Universitat Oberta de Catalunya

uoc.edu

Análisis de la supervivencia y de sus factores pronóstico en pacientes tratados de compresión medular metastásica en el Instituto Catalán de Oncología

Olivia Jordi Ollero
Máster en Bioinformática y Bioestadística – Trabajo final de máster
Área 2 – Subárea11 Análisis de datos y técnicas de clustering

Daniel Fernández Martínez
Marc Maceira Duch

5 de enero del 2020

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis de la supervivencia y de sus factores pronósticos en pacientes tratados de compresión medular metastásica en el Instituto Catalan de Oncología*
Nombre del autor:	*Olivia Jordi Ollero*
Nombre del consultor/a:	*Daniel Fernández Martínez*
Nombre del PRA:	*Marc Maceira Duch*
Fecha de entrega (mm/aaaa):	01/2020
Titulación:	*Máster de Bioinformática y Bioestadística - TFM*
Área del Trabajo Final:	*Área3 – Subárea11 Análisis de datos y técnicas de clustering*
Idioma del trabajo:	*castellano*
Palabras clave	*Biomarcador tumoral* *Escalas RADES, SINS y Blisky* *Enfoque multidisciplinar - Comité*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El objetivo de este trabajo es doble. Por un lado, se quiere comparar la supervivencia de los pacientes afectos de compresión medular (CM) tratados en el Instituto Catalán de Oncología reclutados anteriormente al 2018 con la supervivencia de los que fueron reclutados a partir de ese año, año en que se introdujo un comité multidisciplinar y un nuevo algoritmo de decisión terapéutica. Por otro lado, se quiere estudiar qué factores podrían ser pronósticos de supervivencia para dichos pacientes. Mediante el estimador de Kaplan-Meier se han hallado las curvas de supervivencia de la totalidad de la muestra y de los grupos comentados anteriormente. Se han realizado regresiones de Cox univariantes y multivariantes, así como modelados machine learning mediante árboles de decisión, Random Forest y GBM que han sido validados sobre un conjunto de los datos. No se han podido establecer diferencias en la supervivencia de los dos grupos de pacientes. Resultan variables pronóstico de supervivencia, el tratamiento oncológico tras la CM, el índice Rades, el estado ambulatorio a los 7 días y no haber realizado previamente radioterapia descompresiva. El modelo con mejor rendimiento es la regresión de Cox.

Abstract (in English, 250 words or less):

The goal of this paper is twofold. On the one hand, we want to compare the survival of patients affected by spinal cord compression (CM) treated at the Catalan Institute of Oncology recruited prior to 2018 with the survival of those who were recruited from that year on. On the other hand, we want to study which factors could be prognostic of survival. Using the Kaplan-Meier estimator, the survival curves of the entire sample and the groups discussed above have been found. Univariate and multivariate Cox regressions have been carried out, as well as machine learning modeling using decision trees, Random Forest and GBM that have been validated on a set of data. It was not possible to establish differences in the survival of the two groups of patients. Resulting variables for survival prognosis were oncologic treatment after spine cord compression treatment was finished, Rades index, ambulatory status at 7 days and not having previously performed decompressive radiotherapy. The best performing model was Cox regression.

Índice

1.Introducción

1.1. Contexto y justificación del Trabajo

La compresión medular se define (1) como la aparición de clínica dolorosa o clínica neurológica, con correlación radiológica dónde se observa una lesión en contacto con la dura del cordón medular. Las compresiones subclínicas se caracterizan por la ausencia de clínica cuando aparece una masa que invade el canal medular y contacta con la dura.

Los pacientes oncológicos con lesiones metastásicas en esqueleto axial (vértebras) y con compromiso de la médula espinal tienen tasas de supervivencia bajas. Han sido tratados históricamente con cirugías invasivas o bien con dosis bajas paliativas de radioterapia convencional ("External Beam Radiotherapy", EBRT). En este tipo de cirugías se realiza una resección en bloque de toda la lesión de manera que los márgenes de la pieza extraída están "limpios", no afectos por el tumor. Es por ello que la pieza extraída suele ser grande y por tanto resultan en alta morbilidad del paciente y poco control local a largo plazo (2,3,4). La radioterapia convencional de baja dosis sólo mejora el dolor en el 60% de los pacientes y con una duración media de menos de 4 meses (5,6).

Actualmente, la expectativa de vida de estos pacientes ha aumentado. De hecho, una de las escalas más usadas actualmente en este tipo de pacientes, escala RADES (7), estima el porcentaje de supervivencia a los 6 meses. Esta expectativa de vida superior hace necesaria una solución paliativa a largo plazo con buen control local.

Hoy en día, los pacientes afectos de compresión medular tienen entre las opciones de tratamiento de mayor radicalidad a menor radicalidad: cirugía más radioterapia adyuvante, radioterapia exclusiva o quimioterapia. Entre las opciones quirúrgicas básicas se contemplan desde una operación únicamente de estabilización de la columna hasta una vertebrectomia (cirugía con resección radical), pasando por una laminectomia (cirugía descompresiva simple). Sólo cirujanos entrenados en un programa multidisciplinar de gestión de la columna vertebral, saben realizar la técnica quirúrgica acuñada por Lilyana Angelov y Edward Benzel en el hospital de Clevelant (EUA) llamada "cirugía de separación". Esta cirugía se basa en la creación de un vacío entre la médula espinal y el tumor. La finalidad es permitir la irradiación del tumor con altas dosis mediante técnicas de alta precisión (Stereotactic Body Radiotherapy, SBRT) sin dañar la médula. Sin embargo, la última actualización de la guía clínica de la Asociación Americana de Oncología Radioterápica (ASTRO) (8) para el tratamiento radioterápico de pacientes con metástasis en la columna sólo aconseja por el momento el uso de la SBRT en ensayos clínicos. De hecho, en

la actualidad solo existe un ensayo clínico, el *NCT02320825*, en que se haya investigado el papel de la SBRT postoperatoria en estos pacientes sin que se hayan publicado hasta el momento los resultados. Existen varios esquemas posibles de radioterapia, entre los más comunes 8Gy en una sesión, 5Gy en 4 sesiones, 4Gy en 5 sesiones o 3Gy en 10 sesiones. Varias publicaciones recogen la comparativa de los distintos fraccionamientos como paliativo del dolor (9,10,11). Pese a su equivalencia para gestionar el dolor del paciente, para tratamientos de una sola fracción existen tasas de reincidencia tumoral mayores que en fraccionamientos múltiples (12,13). Es una práctica habitual clínica adoptar esquemas rápidos para pacientes con supervivencias cortas que no vivirán lo suficiente para poder reincidir, y esquemas más largos para pacientes con supervivencias prolongada. En aquellos pacientes con mal estado general o en situación de últimos días sólo se contempla el tratamiento de la paliación medicamentosa de síntomas.

Las últimas guías publicadas (en 2017(14) y 2018(15)) son confluyentes en tres ideas:

i) La herramienta básica de diagnóstico de estos pacientes es la Resonancia Magnética Nuclear

ii) La decisión terapéutica de estos pacientes exige un tratamiento multidisciplinar que englobe oncólogos, radioterapeutas, anestesistas, cirujanos y especialistas en radiodiagnóstico

iii) La decisión terapéutica de estos pacientes deberá tomarse teniendo en cuenta la estabilidad de la columna, el grado de radioresistencia del tumor, el grado de afectación neurológica y el estado de la enfermedad sistémica. La estabilidad de la columna se valora mediante el Spinal Inestability Neoplastic Score (SINS), término acuñado por Fisher et al (16) en el 2007 y que según su puntuación permite clasificar la columna como estable, potencialmente inestable o inestable, con una sensibilidad del 95.7% y una especifidad del 79.5%. Dicha puntuación toma en cuenta 6 características de la columna vertebral (Tabla1, Anexo). Para cuantificar el grado de afectación neurológica, se usa la escalera radiológica introducida por Bilsky et al en el 2010 (17). Dicha escalera establece 6 niveles que valoran el grado de compresión de la médula de menos a más. Grado 1a: Lesión epidural. Grado1b: Contacto con saco dural, pero sin deformación del mismo, Grado 1c deformación del saco dural y contacto con el cordón medular sin deformación del mismo, Grado 2: Deformación del cordón medular, pero con liquido cefalorraquídeo visible alrededor del cordón y Grado3 deformación del cordón sin liquido cefalorraquídeo visible. (Tabla2, Anexo). En cuanto a la valoración de la enfermedad sistémica, no existe consenso. La guía publicada en 2017 usa tan sólo el estado de Karnofsky, que cuantifica el estado de salud actual del paciente, como factor pronóstico (18,19). Mientras que la publicación del 2018 ya introduce índices más avanzados específicamente acuñados como predictores o pronósticos de la supervivencia del paciente, como los índices de Bauer o Bauer-modificado (20,21). De hecho, ha sido estudiada varias veces la dificultad de establero claramente los factores pronósticos (22,23,24). Ello es

preocupante, especialmente cuando la estimación de la supervivencia en los pacientes afectos de compresión medular guarda una estrecha relación con la radicalidad y el tratamiento de elección.

Estas guías establecen relaciones generales indicativas del tratamiento a escoger. Así pues, técnicas quirúrgicas están más indicadas a mayor inestabilidad de la columna, mayor grado de radioresistencia del tumor, mayor compresión de la médula espinal y mejor pronóstico del paciente. Sin embargo, no existe un algoritmo claro de decisión del tratamiento más eficaz para estos pacientes. Las guías recogen la existencia de un par de publicaciones presentando algoritmos de decisión tales como el NOMS (Neurologic – Oncologic - Mechanical Stability - Systemic Desease) (25,26) o el LMNOP (Lieu - Instabilité Méchanique – Néurologie – Oncologie - État du Patient) (27). Estos algoritmos son cualitativos ya que no establecen indicaciones claras basadas en índices para valorar el estado del paciente o el índice de compresión de la columna. Son marcos muy generales que carecen por el momento de un respaldo científico robusto o de uso extendido en la práctica clínica.

Además de todo ello, el único ensayo clínico en faseIII que demostraba una mejor supervivencia en pacientes con compresión medular sometidos a cirugía más radioterapia que aquellos sometidos exclusivamente a radioterapia (9), ha sido fuertemente criticado a nivel metodológico. Primeramente, la muestra de pacientes fue tomada durante 10 años, con los cambios tecnológicos que ello acarrea. En segundo lugar, se excluyeron los tumores más radiosensibles. Por último, se usaron técnicas de cirugía más avanzadas mínimamente invasivas pero técnicas estándar más groseras de radioterapia convencional. De hecho, Rades et al. en un estudio posterior retrospectivo obtuvo resultados opuestos, planteando la necesidad de otro ensayo de fase III conclusivo sobre la materia (28,29).

Un algoritmo de decisiones detallado y validado para pacientes afectos de compresión medular es inexistente. Éste es especialmente necesario en un momento en que se ha generalizado el acceso a la resonancia magnética (imagen de elección en el diagnóstico de CM) y la incidencia de compresiones medulares va en aumento. Por un lado, la población general está más envejecida y por el otro, los pacientes con enfermedades avanzadas cada vez sobreviven más. Los resultados de la inmunoterapia en el cáncer de pulmón avanzado o el melanoma ejemplifican muy bien este cambio, donde las supervivencias han mejorado de forma evidente (30). Hace 20 años, entre el 2,5% y el 5% de los pacientes que morían de cáncer desarrollaban compresión medular (31,32). Imaginemos el número de pacientes afectos de esta patología actualmente.

Hace menos de dos años, en el Instituto Catalán de Oncología, en varios casos, el paciente había muerto entre el transcurso de tiempo entre el diagnóstico y el inicio del tratamiento con radioterapia. Ello plantea la idea de una mala gestión de estos pacientes a los que no se les proporciona el mejor cuidado, invirtiendo recursos de manera ineficaz.

Coincidiendo con la publicación de estas nuevas guías, en el 2018 en el Instituto Catalán de Oncología se adoptó de forma corporativa un enfoque multidisciplinar en el tratamiento de pacientes con compresión medular. La figura1 del apéndice muestra el algoritmo de decisión que se ha adoptado. De acuerdo con el marco teórico que hemos visto, en el algoritmo se considera la estabilidad de la columna mediante el índice SINS, el estado pronóstico del paciente mediante el índice RADES, la radiosensibilidad del tumor y el déficit neurológico mediante la escala ASIA (American Spinal Injury Association). Además, se tiene en consideración el número de vertebras afectas contiguas.

La escala ASIA divide las lesiones medulares en 5 categorías determinadas por la absencia o preservación de la función motora y sensitiva que indican la severidad de la lesión y su posible pronóstico. Estas categorías son: A) Completa. Absencia de función motora y sensitiva que se extiende hasta los segmentos del sacro S4-S5 B) Incompleta. Preservación de la función sensitiva por debajo del nivel neurológico de la lesión que se extiende hasta los segmentos del sacro S4-S5 C) Incompleta. Preservación de la función sensitiva por debajo del nivel neurológico y más de la mitad de los músculos clave por debajo del nivel neurológico tiene un balance muscular menor a 3 D) Incompleta. Preservación de la función motora por debajo del nivel neurológico y más de la mitad de los músculos clave por debajo del nivel neurológico tienen un balance muscular de 3 o más E) Normal. Las funciones sensitiva y motora son normales. Aunque no se ha tenido en cuenta el índice Bilsky para la decisión terapéutica, sí que se ha recogido en la base de datos para cada paciente.

En lo referente al pronóstico del paciente, el índice Rades fue acuñado a raíz de una publicación en la que un análisis multivariante de la tasa de supervivencia de varios pacientes identificaba distintas variables potencialmente pronósticas y que combinaba hábilmente en un solo índice (7). El índice Rades tiene en cuenta el tipo de tumor, el intervalo de tiempo entre el diagnóstico y la compresión medular, la existencia de otras metástasis al inicio de la radioterapia, el estatus ambulatorio y la duración de los déficits motores (Tabla3, Anexo). El índice Rades funciona de forma que atribuye a cada paciente una puntuación que es aproximadamente asociada con una tasa de supervivencia. Es por ello que el índice Rades interviene al decidir el tratamiento de elección.

Si un análisis de los pacientes con compresiones medulares en nuestro centro revelara un aumento de la supervivencia a partir del 2018, ratificaría el algoritmo de decisión como un algoritmo detallado, funcional y efectivo. En la base de datos de estos pacientes se han recogido además muchas otras variables que podrían ser influyentes en la supervivencia. Es el caso, como hemos comentado, del índice Bilsky o por ejemplo de la existencia de biomarcadores. En la guía del 2018 (17) se comenta la posible incidencia de estos sobre la supervivencia. Esta idea surge en analogía al algoritmo de tratamiento de metástasis cerebrales en el que los biomarcadores de mama y pulmón se incluyen en la clasificación pronóstica de pacientes (18,33,34). La hipótesis es que la incorporación de biomarcadores de mama, próstata y pulmón (tumores primarios en porcentaje mayoritario en el caso de compresiones

tumorales) a la escala pronóstica, así como la combinación de la escala de estabilidad, escala RADES y grado de compresión, permitirá predecir con mayor exactitud la supervivencia global.

Los objetivos por tanto de este trabajo son por tanto por un lado comparar de manera estadísticamente significativa la supervivencia de los pacientes antes y después del 2018 y, por otro lado, establecer posibles variables significativas.

1.2 Objetivos del Trabajo

1.2.1 Objetivos generales

a. Comparar la supervivencia antes del 2018 y después del 2018 en pacientes tratados de compresión medular metastásica en el Instituto Catalán de Oncología
b. Estudiar la influencia de otras variables en la supervivencia de estos pacientes

1.2.2 Objetivos específicos

Relativos a preparar y conocer la base de datos para cualquier análisis (se podrían achacar tanto al objetivo general a. como al b.):

1. Preparar la base de datos para el análisis
2. Realizar un análisis descriptivo de la muestra de pacientes (de la variable respuesta y de toda otra variable) y análisis de missing y de outliers

Relativos al primer objetivo general:

3a Comprobar que existe una diferencia significativa en la supervivencia para los distintos grupos Rades
4a Valorar para toda variable si existe una diferencia significativa antes del 2018 y después del 2018
5a Estimar la supervivencia antes del 2018 y después del 2018 y, si existe diferencia, estudiar su significancia

Relativos al segundo objetivo general:

6b Realizar un estudio univariante para toda variable
7b Realizar un estudio multivariante
8b Modelar varios sistemas que predigan la tasa de supervivencia del paciente y valorar su eficiencia
9b Valoración de los resultados obtenidos

1.3 Enfoque y método seguido

La variable respuesta es el tiempo de supervivencia o de censura, medido como el intervalo entre el fin del tratamiento (sea por radioterapia o por cirugía) y el exitus del sujeto o fecha del último control.

Las variables de entrada serían:

- variables demográficas
  i. edad
  ii. sexo
  iii. hospital al que acuden (Hospital Duran i Reynals, Hospital Germans Tries I Pujol)
  iv. estatus (vivo, muerto o pérdida de seguimiento)
- variables de tiempo relativas al momento del diagnóstico por resonancia magnética:
  i. tiempo desde la existencia de una sospecha clínica al diagnóstico por resonancia magnética
  ii. tiempo des del diagnóstico de resonancia magnética al aviso de compresión
  iii. tiempo des del diagnóstico por resonancia magnética al inicio de radioterapia
- estado previo al tratamiento
  i. existencia de radioterapia descompresiva previa
  ii. evaluación del dolor
  iii. debut de la clínica (asintomático, dolor neuropático o clínica sensitiva, clínica motora, clínica somática y sus combinaciones)
  iv. tiempo des de la aparición de clínica sensitiva hasta el diagnóstico
  v. tiempo des de la aparición de clínica motora hasta el diagnóstico
- datos relativos a la compresión
  i. Localización (cervical, dorsal, lumbar, sacro y múltiple)
  ii. Tipo de compresión (medular, radicular, intramedular)
  iii. Número de vértebras afectas
  iv. Número de vértebras comprimidas
- índices:
  i. RADES y las variables de cuya combinación se obtiene (tumor primario, existencia de metástasis óseas y viscerales, estado ambulatorio, tiempo des del desarrollo de déficits motores, tiempo des del descubrimiento del primario a la compresión)
  ii. SINS y las variables de cuya combinación se obtiene (vértebras afectas, dolor, lesión en el hueso, alineación de la columna, colapso vertebral, afectación posterolateral)
  iii. Vicky (6 grados posibles)
- Datos relativos al tratamiento
  i. Tipo de test diagnóstico
  ii. Opción terapéutica
  iii. Dosis de radioterapia

- Estado del paciente tras la compresión:
    i. Evaluación del dolor a los 7 días
    ii. Estado a los 7 días
    iii. Empeoramiento neurológico
    iv. Tratamiento clínico posterior

La estrategia a seguir durante el trabajo es la de lograr completar las guías clínicas actuales sobre el manejo de este tipo de pacientes. Para ello, realizaremos un análisis estadístico de la base de datos mediante R. Pensando en nuestros objetivos:

→ Para el estudio estadístico de nuestra base de datos:

- variables numéricas. A nivel analítico se describirán las medidas de dispersión y desviación estándar que procedan según la distribución de la variable, se estudiará la normalización y simetría, así como la correlación entre las distintas variables. A nivel gráfico se presentarán cajas de dispersión, gráficos de puntos bivariados con histogramas de las distribución y gráficos de correlación de las variables. Se realizará la transformación según la escalera de Tucker de las variables con distribuciones alejadas a la normal. Se realizará un análisis de los outliers y valores extremos para variables transformadas y no transformadas mediante el uso de cajas de dispersión y mediante el uso del test de Rosner si la variable presenta distribuciones normales.

-variables categóricas. Se presentará para cada variable un gráfico de barras con el porcentaje de las distintas categorías. Se realizarán cajas de dispersión de la supervivencia (variable independiente) según las distintas categorías de cada variable. Se repetirá dicha operación para la variable transformada de la supervivencia si procede.

-se estudiarán los valores *missing.* Se eliminarán las variables con un porcentaje de NA superior al 35%. Se describirán los porcentajes de NA para cada variable. Se usará una aproximación a los vecinos “más parecidos” para reemplazar los valores NA, si no son excesivos.

- valorar si existe o no una mejora en la supervivencia de los dos grupos

→ En el caso de la supervivencia según Rades o según grupos (antes y después del 2018) se trata de estudios de supervivencia en relación a variables categóricas.

- Se realizará un test chi cuadrado para cada variable comparando su distribución antes y después del 2018

- Estudiaremos la supervivencia mediante el estimador de Kaplan-Meier. Presentaremos las tablas de vida y la curva de supervivencia. Primero en función de los distintos grupos Rades para “testear” la bondad de nuestra base de datos, después de toda la base de datos y finalmente para los grupos1 y 2 (antes/después del 2018). Se calculará la esperanza de vida y la tasa de supervivencia a 6 meses.

- Valoraremos la significancia estadística obtenida mediante un log-Rank test.

- encontrar posibles variables de influencia

→ se usará el método semiparamétrico de la regresión de Cox de riesgos proporcionales que supone un clásico para temáticas de supervivencia con datos censurados. Las regresiones de Cox admiten variables predictoras tanto categóricas como cuantitativas.

- Estudio de la posible dependencia de las variables con el tiempo. Si se consideran independientes, se procederá al modelado univariante por regresión de Cox simple para cada una de las variables.

- Comprobación del cumplimiento de la hipótesis de riesgos proporcionales para toda regresión univariante de forma analítica mediante los residuos de Schoenfeld y del test de Kolgonorov. Si la hipótesis no se cumple para alguna variable, se supondrá un coeficiente dependiente del tiempo de forma escalonada para esa variable y se volverá a comprobar el cumplimiento de la hipótesis.

- Consideración de problemas de multicolinealidad en una regresión de Cox multivariante. Eliminación de variables categóricas y numéricas por correlación con otras. Si hay demasiadas variables, estudio mediante PCA.

- Regresión de Cox simple multivariante. Se irá afinando la concordancia de la regresión mediante la eliminación de variables que no sean de interés.

- Comprobación del cumplimiento de la hipótesis de riesgos proporcionales para la regresión escogida de forma analítica mediante los residuos de Schoenfeld y del test de Kolgonorov. Si la hipótesis no se cumple para alguna variable, se supondrá coeficientes dependientes del tiempo de forma escalonada para esas variables y se volverá a comprobar el cumplimiento de la hipótesis según muestra.

- Diagnósticos para el último modelado multivariante mediante la regresión de Cox. Comprobación gráfica del cumplimiento de la hipótesis mediante los residuos de Shoenfeld. Estudio de outliers y puntos extremos mediante la representación de los residuos desviados. Estudio de la linealidad de las covariables cuantitativas mediante la representación gráfica de los residuos de Martingale.

- En referencia a la publicación acuñando el término de índice de RADES (7) se realizará la misma regresión multivariante

eliminando esta variable e introduciendo todas aquellas que se han utilizado para su cálculo.

- crear un modelo que nos permita predecir la tasa de supervivencia del paciente

→ Modelaremos por un lado con el algoritmo clásico en temáticas de supervivencia y por otro lado con los algoritmos más novedosos de machine learning de mayor uso en temática de supervivencia. En cuanto a los algoritmos de machine learning se intentará combinar algoritmos para la clasificación, así como algoritmos para la regresión, así como algoritmos de uso combinado.

- Creación de dos juegos de data sets de entrenamiento y aprendizaje. Los conjuntos de datos de entrenamiento (o aprendizaje) entre ellos serán exactamente iguales exceptuando la creación en uno de ellos de una variable que almacenará en un tiempo arbitrario fijado el estado vivo/muerto del paciente. Esta paciente se usará para aquellos algoritmos que trabajan como "clasificadores".

- Modelización de toda la base de datos mediante una regresión de Cox multivariante, considerando exclusivamente las covariables demostradas anteriormente de interés, con aproximación de Bridge para los coeficientes dependientes del tiempo. La predicción para modelados de regresión de Cox multivariante con función escalonada para los coeficientes de las variables que no cumplen la hipótesis de proporcionalidad de riesgos con un coeficiente constante, no está aún implementada en R. Comprobación de la concordancia del modelo y significación de las variables en referencia al anterior modelado. Si es relativamente parecido, se procederá al modelado exclusivo de la base de datos de entrenamiento. Se observará su significancia y concordancia. Se usará dicho modelo para realizar una predicción del conjunto de datos de validación o test para un tiempo fijo t_0.

- Modelización del grupo de entrenamiento mediante el árbol de decisión de clasificación C5.0. Predicción mediante el modelo sobre el grupo de validación para un tiempo fijo t_0.

- Modelización del grupo de entrenamiento mediante el árbol de decisión de regresión rpart. Predicción mediante el modelo sobre el grupo de validación para un tiempo fijo t_0. Se mostrará mediante soporte gráfico las variables de mayor relevancia en el árbol de decisión.

- Modelización del grupo de entrenamiento mediante dos algoritmos Random Forest, del paquete ranger y del paquete SRS, algoritmos

de uso dual (clasificación y regresión). Predicción mediante el modelo sobre el grupo de validación para un tiempo fijo t_0. Se mostrará mediante soporte gráfico cuáles son las variables de mayor utilidad para el modelo.

- Modelización del grupo de entrenamiento mediante el algoritmo GBM ("Gradient Boosting Model"), algoritmo avanzado de clasificación. Predicción mediante el modelo sobre el grupo de validación para un tiempo fijo t_0.

- Cuantificación del rendimiento de cada algoritmo para la predicción en un tiempo t_0 mediante matrices de confusión.

- Cuantificación del rendimiento de cada algoritmo para la predicción en un tiempo t_0 mediante la eficiencia, sensibilidad, sensitividad, precisión y estadístico kappa.

- Cuantificación del rendimiento de cada algoritmo para la predicción en un tiempo t_0 mediante el área bajo la curva. Representación gráfica de curvas ROC.

- Representación del rendimiento de los algoritmos de regresión mediante curvas de predicción sobre el conjunto de datos de validación para todo instante de tiempo

1.4 Planificación del Trabajo

1.4.1 Tareas

La planificación temporal se lleva a cabo mediante un diagrama de Gantt que se incrusta en este apartado. En lo relativo a las tareas usaremos la misma numeración usada anteriormente para la descripción de los objetivos específicos, creando subapartados cuando haya varias tareas que se refieran a un mismo objetivo específico. Recordamos que aquellos objetivos con "*a"* hacen referencia al primer objetivo general, así como los de "*b"* al segundo.

1. Objetivo: Preparar la base de datos para el análisis

1.1 Detectar datos no introducidos y errores. Informar a los médicos para que sean rellenados/corregidos.

1.2 Definir las variables de interés (existencia de la variable, definición del tipo de variable, definición de los valores que puede tomar esa variable).

2. Objetivo: Realizar un análisis descriptivo de la muestra de pacientes y análisis de missing y de outliers
   2.1 Describir visualmente la base de datos. Variables categóricas: gráfico de barras. Variables numéricas: histogramas, boxplot, scatter plot para análisis bivariado, gráficos de correlación. Boxplot de la supervivencia en función de las variables categóricas.
   2.2 Localizar y describir con técnicas estadísticas de exploración la existencia de valores missing
   2.3 Realizar para las variables numéricas cálculo de medidas de tendencia central, de dispersión, de normalidad y de simetría. Describir visualmente mediante qqplots la normalidad de las distribuciones de variables.
   2.4 Localizar, describir y tratar con técnicas estadísticas de exploración la existencia de valores outliers y extremales.
   2.5 Realizar transformaciones para la normalización de las distribuciones de las variables numéricas.
   2.6 Describir visualmente y analíticamente la base de datos como en los puntos anteriores (2.1 y 2.3), tras la transformación de variables y tratamiento de valores missing y outliers.

3. Objetivo: Comprobar que existe una diferencia significativa en la supervivencia para los distintos grupos RADES
   3.1 Visualizar las curvas de supervivencia para los tres grupos
   3.2 Comprobar que existe diferencia entre ellas y que es estadísticamente significativa
   3.3 Estimar la esperanza de vida para cada grupo y la tasa de supervivencia a los 6 meses

4. Objetivo: Valorar para toda variable si existe una diferencia significativa en su distribución antes del 2018 y después del 2018
   4.1 Todas las variables del estudio pueden ser tratadas como categóricas (con pocos grupos) por lo que podemos realizar un test chi cuadrado para valorar la homogeneidad de su distribución antes del 2018 y después del 2018

5. Objetivo: Estimar la supervivencia antes del 2018 y después del 2018 y, si existe diferencia, estudiar su significancia
   5.1 Encontrar las tablas de vida para la muestra general y para ambos grupos. Trazar e interpretar las curvas de Kaplan-Meier
   5.2 Estimar la supervivencia a 6 meses y la esperanza de vida de la muestra en general y de ambos grupos

   5.3 Estudiar la significancia de la diferencia en la supervivencia antes del 2018 y después del 2018 mediante un log-Rank test
   5.4 Valorar los datos censurados en ambos grupos y discutir su influencia en los resultados obtenidos

6b Objetivo: Realizar un estudio univariante para toda variable

6.1b Analizar si existe alguna variable dependiente del tiempo
6.2b Realizar una regresión de Cox univariante simple para toda variable.
6.3b Comprobar la hipótesis de proporcionalidad de riesgos analíticamente (test Kolmogorov – residuos de Schoenfeld) y gráficamente
6.4b Realizar una regresión de Cox con coeficiente definido por una función escalón para aquellas variables que no cumplan la hipótesis de proporcionalidad de riesgos. Reanalizar analíticamente el cumplimiento de la hipótesis de proporcionalidad de riesgos
6. 5b Discutir los resultados. Concordancia de los modelos, tests probabilísticos y riesgos proporcionales de las variables.

7b Objetivo: Realizar un estudio multivariante
7.1b Reducir el número de variables predictoras para evitar problemas de multicolinealidad. Análisis PCA en caso de muchas variables.
7.2b Realizar una regresión de Cox multivariante simple para toda variable. Ir eliminando variables hasta alcanzar el mejor modelo posible.
7.3b Comprobar analíticamente la hipótesis de proporcionalidad de riesgos
7.4b Realizar una regresión de Cox multivariante estratificada para incluir aquellas variables que no cumplan la hipótesis de proporcionalidad de riesgos. Recomprobar analíticamente la hipótesis de proporcionalidad de riesgos.
7.4b Realizar una regresión de Cox multivariante con los coeficientes de aquellas variables que no cumplan la hipótesis de proporcionalidad de riesgos, definidos en el tiempo mediante la función escalón. Recomprobar analíticamente la hipótesis de proporcionalidad de riesgos.
7.5b Discutir los resultados y elección de un modelo
7.6b Diagnósticos para el modelo de Cox resultante. Comprobación visual de la hipótesis de proporcionalidad de riesgos mediante los residuos de Schoenfeld. Comprobación de puntos extremos y outliers mediante los residuos desviados. Comprobación de la linealidad de las variables cuantitativas mediante los residuos de Martingale.
7.7b Repetir la regresión de Cox escogida eliminando la variable Rades e introduciendo todas aquellas variables a partir de la cual se calcula.

8b Objetivo: Modelar un sistema que prediga la tasa de supervivencia del paciente y valorar su eficiencia
8.1b Dividir la base de datos en un conjunto de entrenamiento y otro de validación. Clonarlos y añadir una variable que tenga en cuenta el estado del paciente en un tiempo fijo t_0. De manera, que finalmente tengamos 2 conjuntos de entrenamiento y validación.
8.2b Realizar una regresión de Cox multivariante con extensión de Bridge para toda la base de datos. Comparar su modelado con la regresión de Cox de elección del punto anterior. Si es comparable, realizar una regresión de Cox multivariante con extensión de Bridge para la base de datos de entrenamiento. Predecir sobre los datos de validación.

8.3b Modelizar los datos mediante *tree analysis* y predecir sobre los datos de validación. Usar los algoritmos C5.0 de clasificación y el rpart de regresión.

8.4b Modelizar los datos mediante un *random forest analysis* y predecir sobre los datos de validación. Usar los algoritmos del paquete ranger y el paquete RandomForestSRS.

8.5b Modelizar los datos mediante *GBM* y predecir sobre los datos de validación.

9b Objetivo: Valoración de los resultados obtenidos

9.1b Visualización de las matrices de confusión para todos los métodos usados

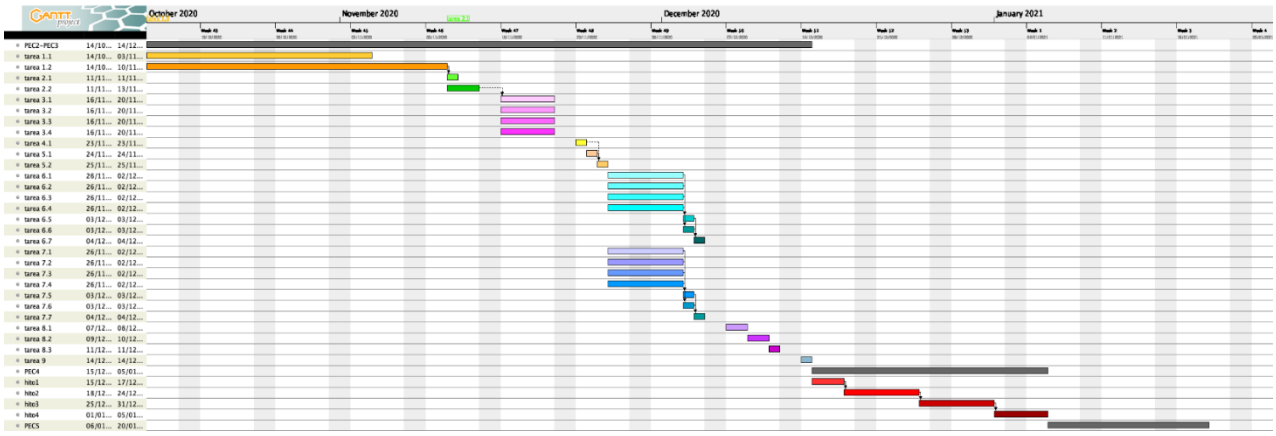
9.2b Comparación de la regresión y el los algoritmos de machine learning mediante la eficiencia, sensibilidad, sensitividad, precisión, recall y el estadístico kappa

9.3b Comparación de la regresión y los algoritmos de machine learning mediante el área bajo la curva ROC (Receiver Operating Caracterization). Comparación visual y analítica.

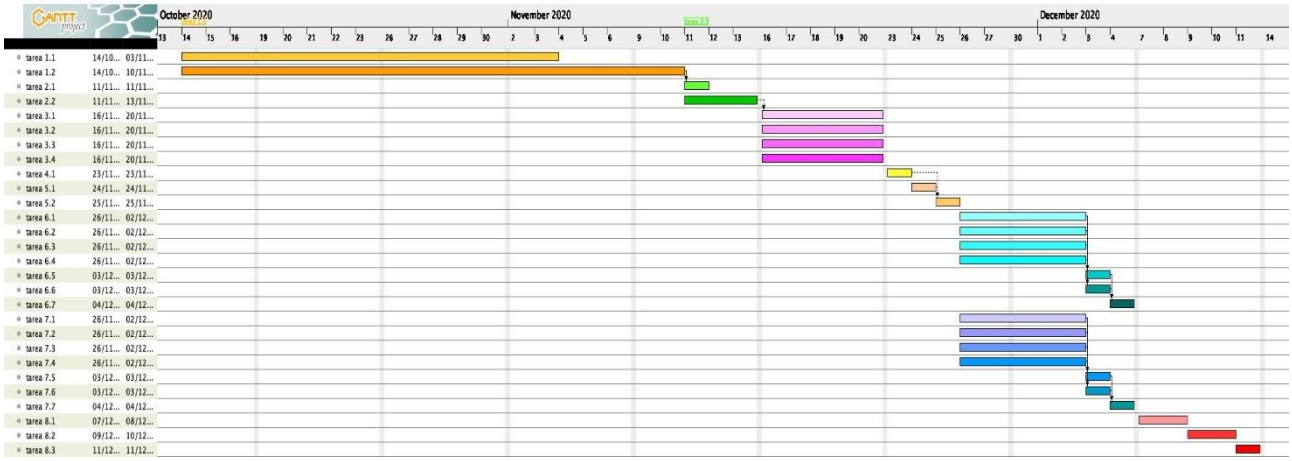
9.4c Comparación de la regresión y el modelado machine learning por regresión mediante curvas de predicción. Uso del paquete de R *Pec* (36-38)

1.4.2 Calendario

Para una estimación completa del desarrollo de todo el TFM podemos visualizar la siguiente planificación que contempla las fechas de todas las PEC además de las tareas e hitos. Las PEC figuran en gris, los hitos en rojo y las tareas en distintos colores agrupadas en una misma tonalidad si pertenecen al mismo objetivo específico.



Si nos centramos en la PEC2 y PEC3 que comprenden el desarrollo del trabajo y visualizamos la planificación:



1.4.3 Hitos

Definimos los hitos en base a las tareas planificadas.

Hito1 Sección de exploración descriptiva de la base de datos ejecutada y escrita en el TFM (grupo de tareas 2)

Hito2 Sección referente a la supervivencia comparada de los sujetos previamente y después del 2018 ejecutada y escrita en el TFM (grupos de tareas 3, 4 y 5)

Hito3 Sección referente al estudio de las variables influyentes en la supervivencia ejecutada y escrita en el TFM (grupos de tareas 6 y 7)

Hito4 Sección referente a la modelización de la base de datos para predecir la tasa de supervivencia de los sujetos, ejecutada y escrita en el TFM (grupo de tareas 8)

1.4.4 Análisis de riesgos

Los factores que pueden influir negativamente en el seguimiento de la planificación realizada son a parte del tiempo, muy especialmente la obtención de la base de datos completada con los datos que no fueron rellenados cuando se introdujo el paciente en la base de datos.

1.5 Breve sumario de productos obtenidos

Al final del proyecto se esperan obtener los siguientes entregables:

5.1 Plan de trabajo
5.2 Memoria del TFM
5.3 Presentación virtual
5.4 Autoevaluación del proyecto

1.6. Breve descripción de los otros capítulos de la memoria

La memoria contiene 6 capítulos más, el capítulo final de conclusiones y cinco capítulos más con título respectivamente: “Conociendo la base de datos”, “Hablando sobre la supervivencia”, “Variables influyentes en la supervivencia”, “Prediciendo la tasa de supervivencia” “Valorando la bondad del ajuste”. En el primero estudiaremos visualmente y analíticamente nuestra base de datos, gestionaremos outliers y missings, y transformaremos las variables si es necesario. En el segundo estimaremos mediante Kapplan-Meier las curvas de supervivencia en distintos escenarios: globalmente, por grupos y según índice Rades. Se calculará también la significancia de las diferencias, así como la tasa de supervivencia a 6 meses y la esperanza de vida. En el tercero, realizaremos regresiones de Cox univariante y multivariante para descubrir qué variables resultan pronósticas de la supervivencia. Se procederá a distintos modelados para aquellas variables que no cumplan la hipótesis de proporcionalidad de riesgos. En el cuarto capítulo, ejecutaremos varios modelados mediante los algoritmos de machine learning más usados en temática de supervivencia: árboles de decisión, Random Forest y Gradient Boosting Modelling. En el cuarto, valoraremos la bondad del ajuste de las predicciones tanto de los modelados por machine learning como mediante la regresión de Cox. Usaremos la curva ROC, matrices de confusión y curvas de predicción. El apartado de conclusiones contendrá obviamente las conclusiones que se han hallado, una reflexión crítica sobre el grado de compleción de los objetivos, un análisis crítico del seguimiento de la planificación y una discusión sobre la repercusión futura del trabajo y las posibles líneas de investigación que podrían seguirse.

2. Conociendo la base de datos

2.1 De Access a R

La base de datos original, en formato Access, contiene 320 episodios de CM y 121 variables recogidas (Tabla1). La obtención de una base de datos sobre la cual trabajar ha sido un proceso laborioso y complejo. La base de datos ha sido pensada por médicos y no por estadísticos, que al no analizar la base de datos después no han tenido en cuenta algunos detalles. Dejando de lado que la mayoría de variables de interés se han tenido que crear a partir de las recogidas, los problemas han sido numerosos: problemas de acceso informático al software que contenía algunos datos, retrasos en la entregas sucesivas de la base de datos que no llegaba a estar completa, intervención de diferentes sujetos que entendieron las tabulaciones de distintas variables de forma distinta dando lugar a una heterogeneidad que ha tenido que ser luego perseguida y "limpiada", múltiples errores en la introducción de las variables (cambios de mes por día en fechas por ejemplo), sujetos introducidos varias veces sin tener varios episodios de CM y problemas en el diseño de la base de datos al almacenar los datos. Por ejemplo, ante el evento de compresión múltiple registrada como un sólo episodio múltiple, la puntuación SINS de ambas CM se sumaba dando lugar a clasificaciones incoherentes que no se detectaron hasta realizar el estudio de outliers. La limpieza de la base de datos ha ocupado gran parte del tiempo dedicado al TFM y se ha vuelto a ella de manera recursiva ante la obtención de resultados raros.

Durante ese proceso de limpieza se detectan varias singularidades. La mayoría de ellas implican la eliminación de ese episodio de CM. Existen id distintos asociados a un mismo SAP (número clínico identificativo del paciente). Son los id 7/8, 10/11, 15/16, 26/27, 28/29, 30/31, 34/35, 38/39, 112/155, 131/132, 159/160/161, 166/167/229, 192/193, 135/271/272 y 293/295/296. Se busca en los registros clínicos para discernir si se trata de distintos episodios de CM o por el contrario se trata de un error. A resultas de ello, se eliminan todos aquellos id que corresponden a episodios de CM ya introducidos. Son los id: 27, 35, 155, 166, 167, 193, 295. Además, se detectan 3 pacientes que no llegan a iniciar el tratamiento. El paciente asociado al id 109 fallece el mismo día que debía empezar el tratamiento (en su caso, radioterapia). En el caso de los id 278 y 287, los pacientes asociados deciden no someterse al tratamiento (en el primer caso tratamiento radioterápico y en el segundo tratamiento farmacológico de soporte). Estos 3 casos, aunque diagnosticados no llegaron a iniciar su tratamiento por lo que eliminamos el evento del estudio, son los id: 109, 278,287. Además, encontramos otro caso que, aunque sí que empieza al tratamiento no lo termina. Es el paciente de id 112. Al operar el paciente se detecta que es una tuberculosis vertebral que no requiere radioterapia al no tener origen tumoral. Este paciente no se elimina del estudio ya que sí inicia tratamiento, supone una pérdida informada. Por último, el paciente asociado al id 106 presenta unas fechas de

diagnóstico, éxitus y tratamiento de radioterapia incoherentes, por lo que se decide eliminarlo también del estudio. Todo ello supone una pérdida de 10 id en el estudio.

Además, se realizan múltiples correcciones en las fechas en las que se constata claramente errores al introducir los datos. Tal y como funciona el circuito clínico del paciente afecto de CM, existe una sospecha clínica previa o coincidente con el diagnóstico, que desencadena un aviso de compresión medular para desembocar en un tratamiento de radioterapia. Por tanto: la fecha de sospecha clínica debe ser menor o igual a la de diagnóstico, la de diagnóstico debe ser menor o igual que la de aviso de compresión medular y la de aviso de compresión medular debe ser menor o igual a la del inicio de tratamiento que a su vez será igual o menor que la del final de tratamiento. También se corrige alguna fecha relacionada con la edad, al mostrar edades incoherentes. Se realizan correcciones en los siguientes id: 14, 15, 16, 18, 24, 25, 52, 81, 94, 109, 112, 131, 156, 157, 158, 159, 172, 191, 195, 225, 240, 259, 282, 283, 306 y 314.
Todo ello se deja debidamente documentado en la base de datos en la columna de "Comentarios".

Se crean 22 variables nuevas (Tabla2) y se retabulan las categorías de la variable *histología* de forma que cada categoría contenga al menos 4 episodios de CM. Las categorías con frecuencia menor a 4, pasan a engrosar la categoría de "otros". Así se pasa para dicha variable de 28 categorías a 12. Se eliminan 85 variables que caen en desuso tras la creación de las nuevas. Las nuevas variables (*puntuacion_RADES, puntuacion_SINS)*, se calculan a partir de la suma de sus componentes. En la tabla podemos observar los puntos que cada categoría de cada variable aporta a esa puntuación.

Nuestra base de datos tras todos los cambios introducidos consta de 290 episodios de CM y 49 variables, como queda descrito en la Tabla3. Se redefinen las variables como factor y numérica según proceda.

Variable	Definición	Categorías	
		Código	Significado
Relativas a datos identificativos del paciente			
ID	nº identificativo del episodio de compresión medular (CM)		
SAP	nº identificativo del paciente que sufre el episodio de CM		
fecha_nacimiento	fecha de nacimiento		
genero	genero	0	hombre
		1	mujer
centro	centro donde se realiza el tratamiento aunque no la hospitalización	0	Hospital Germans Tries i Pujol
		1	Hospital Duran i Reynals
estatus_UC	estado del paciente en el último control previo al cierre del estudio	0	vivo
		1	muerto
Relativas a fechas temporales que describen el curso clínico del paciente			
fecha_UC	fecha del último control		
fecha_exitus	fecha del fallecimiento del sujeto		
fecha_sospechaCM	fecha de la sospecha clínica de CM		
fecha_diagnostico	fecha de establecimiento de diagnóstico de CM		
fecha_avisoCM	fecha de aviso de CM		
fecha_Inicio_RT	fecha en que se inicia el tratamiento de Radioterapia (RT) si existe		
fecha_FIN_RT	fecha en que finaliza el tratamiento de RT si existe		
fecha_cirugía	fecha en que se realizala cirugía si existe		
Relativas a la CM			
loc_cervical	4 variables que recogen como variables booleanas de verdadero/falso si la CM se encuentra en ese tramo de la columna	0	no
loc_dorsal			
loc_lumbar		1	si
loc_sacra			
vert_afect_C1/C2/.../D3	25 variables booleanas que recogen para cada vértebra si está afecta	0 / 1	no / si
vert_comp_C1/C2/.../D3	25 variables booleanas que recogen para cada vértebra si está comprimida	0 / 1	no / si

histología	histología del tumor	0	Sin AP
		1	Adenocarcinoma
		2	Carcinoma escamoso
		3	Carcinoma de célula pequeña
		4	Células claras
		5	Mieloma múltiple
		6	Carcinoma no célula pequeña
		7	Carcinoma ductal infiltrante
		8	Carcinoma lobulillar infiltrante
		9	Carcinoma urotelial
		10	Hepatocarcinoma
		11	Linfoma
		12	Otros
CM_PrimarioConocido	existencia de un tumor primario conocido	0 / 1	no / si
Biomarcadores	existencia de biomarcadores tumorales	0	no existen / no disponible
		1	existen
EGFR/ALK/RP/RE	4 variables que recogen, si lo hay, el valor del biomarcador tumoral de esa CM (los dos primeros son relativos a cáncer pulmón y los dos últimos a cáncer de mama)	0	negativo
		1	positivo
PDL1/Her2 Neu	2 variables que recogen, si lo hay, el valor del biomarcador tumoral de esa CM (el primero es relativo a cáncer pulmón y el último a cáncer de mama)	0	negativo
		1	low
		2	high
PSA	valor numérico del biomarcador específico de próstata		
Relativas a índices			
RADES_HC	clasificación del paciente afecto de CM en la escalera RADESque se tomó clínicamente previamente a ser tratado. La escalera de RADES valora la supervivencia del paciente.	0	Grupo1 (20-30ptos)
		1	Grupo2 (31-35 ptos)
		2	Grupo3 (36-45ptos)
tumPrimario_RADES	variable recogida a efectos del cálculo de la puntuación RADES. Información sobre el tipo de tumor primario que ha originado la CM	0	Mama (8ptos)
		1	Próstata (7ptos)
		2	Mieloma/Linfoma (9ptos)
		3	Pulmón (3ptos)
		4	Otros/orígen desc/sin AP(4ptos)
metOseas_RADES	variable a efectos del cálculo de la puntuación RADES. Existencia de metástasis óseas	0 / 1	no(7ptos) / si(5ptos)
metVisc_RADES	variable a efectos del cálculo de la puntuación RADES. Existencia de metástasis viscerales	0 / 1	no(8ptos) / si(2ptos)
estAmbulatorio_RADES	variable a efectos del cálculo de la puntuación RADES. Estado ambulatorio del paciente	0 / 1	no(3ptos) / si(7ptos)
intervalo_diagCM_RADES	variable a efectos del cálculo de la puntuación RADES. Intérvalo de tiempo entre el diagnóstico primario y la aparición de CM	0	=< 15meses (4ptos)
		1	> 15meses (7ptos)
tiempoDesarrollo_DefMotor_ RADES	variable a efectos del cálculo de la puntuación RADES. Tiempo de desarrollo de déficits motores	0	De uno a siete (3ptos)
		1	De ocho a catorce (6ptos)
		2	Más de catorce (8ptos)
		3	Sin déficit motor/Subclínica (8ptos)
SINS_HC	clasificación del paciente afecto de CM en la escalera SINS que se tomó clínicamente previamente a ser tratado. La escalera SINS valora la estabilidad de la columna.	0	Estable (0-6ptos)
		1	Potencialmente inestable (7-12ptos)
		2	Inestable (13-18ptos)
Nivel_M1_C1C2/C3C6/.../S2S5 _SINS	8 variables booleanas que recogen la afectación por la CM en ese tramo de la columna. A efectos del cálculo de la puntuación SINS.	0	no
		1	si
dolor_SINS	A efectos del cálculo de la puntuación SINS. Variable que recoge el dolor que experimenta el paciente.	0	ocasional (1pto)
		1	sí (3ptos)
		2	sin dolor (0ptos)
caract_lesion_SINS	A efectos del cálculo de la puntuación SINS. Características de la lesión en la columna.	0	lítica (2ptos)
		1	mixta (1pto)
		2	blástica (0ptos)
alineacColumna_SINS	A efectos del cálculo de la puntuación SINS. Alineación de la columna.	0	subluxación o translación (4ptos)
		1	deformidad nueva (2ptos)
		2	normal (0ptos)
colapsoVertebral_SINS	A efectos del cálculo de la puntuación SINS. Colapso vertebral de la columna.	0	> 50% (3ptos)
		1	< 50% (2ptos)
		2	sin colpaso,>50%cuerpo afecto(1p)
		3	ninguna de las anteriores(0ptos)
afectPosterolat_SINS	A efectos del cálculo de la puntuación SINS. Afectación posterolateral de la columna.	0	bilateral (3ptos)
		1	unilateral (1pto)
		2	ninguna de las anteriores (0ptos)
Blisky	Clasificación en la escalera de Blisky, que valora el grado de afectación del canal medular en la CM	0	0
		1	A1
		2	A2
		3	A3
		4	B
		5	C
Relativas al estado del paciente previo al tratamiento			
clinDebt_somatico/ neuropatico/ motora/	4 variables booleanas que recogen el début clínico del paciente	0	no
		1	si
dolor_eval	escala de dolor numérica (del 0-10) que siente el paciente previamente al tratamiento		
tiempo_clinSensitiva_diagn	intérvalo de tiempo en horas desde la aparición de clínica sensitiva al diagnóstico de CM		
tiempo_clinMotora_diagn	intérvalo de tiempo en horas desde la aparición de clínica motora al diagnóstico de CM		
RT_descompPrevia	existencia de radioterapia previa descompresiva para ese paciente	0 / 1	no / si
Relativas al estado del paciente tras el tratamiento			
estatus_lesion_UC	Estado de la lesión de CM tras el tratamiento en el último control realizado antes del cierre del estudio	0	Estable
		1	Progresión en PTV
		2	Progresión fuera de PTV
		3	RC Radiológica
		4	RC Metabólica
		5	No evaluada
empeoramiento_neurologico	empeoramiento neurológico a 7 días tras la finalización del tratamiento	0 / 1	no / si
dolor_eval7	escala de dolor numérica (del 0-10) que siente el paciente 7 días tras finalizar el tratamient	0 / 1	no / si
estatusAmbulatorio7	estatus ambulatorio del paciente 7 días después de la finalización del tratamiento	0 / 1	no / si

tto_trasCM	tratamiento oncológico realizado tras finalizar el tratamiento	0	No inicia tto
		1	No continúa tto
		2	Continúa tto
		3	Inicia tto
Relativas al tratamiento			
fijacionPrevia_segIrradiar	Episodio de CM que requiere la estabilización previa del segmento antes del tratamiento	0 / 1	no / si
dosis_RT	Fraccionamiento de la dosis de radioterapia: número de sesiones y dosis por sesión	0	8Gy x 1sesión
		1	5Gy x 4sesiones
		2	4Gy x 5sesiones
		3	3Gy x 10sesiones
		4	otros
test_diagnostico	Test realizado para determinar el diagnóstico de CM	0	Resonancia Magnética Nuclear (RM)
		1	Tomografía computarizada (TC)
		2	otros
opción terapéutica	opción terapéutica realizada ante el evento de CM	0	Cirugía (CR)
		1	CR + RT
		2	RT + CR
		3	RT
		4	tto farmacológico de soporte
comentarios	variable que recoge comentarios sobre el episodio de CM		

Tabla1 Variables y tabulación de la base de datos original en Access

Variables de nueva creación	Definición	Categorías	
		Código	Significado
Relativas a datos identificativos del paciente			
edad	intérvalo de tiempo en años desde la fecha de nacimiento a la fecha de diagnóstico de CM		
grupo	divide los pacientes según la fecha del diagnóstico	0	Grupo1 (antes del 2018)
		1	Grupo2 (durante o después del 2018)
Relativas a fechas temporales que describen el curso clínico del paciente			
fecha_fin*	A efectos de cálculo.. fecha del último control o fecha de fallecimiento del sujeto.		
fecha_FIN_tto*	A efectos de cálculo de la variable survival. Fecha que recoge la fecha de finalización del tto sea con cirugía (*fecha_cirugia*) o con RT (*fecha_FIN_RT*).		
survival	variable output, tiempo de supervivencia o censurado. Intérvalo de tiempo entre la finalización de tto (sea con CR o RT) y el último control o éxitus del paciente (*fecha_fin).* Para pacientes en soporte recoge el tiempo desde el diagnóstico a su muerte.		
t_ac	intérvalo de tiempo (días) desde el diagnóstico al aviso de CM		
t_sc	intérvalo de tiempo (días) desde la sospecha clínica de CM al diagnóstico		
t_itto	intérvalo de tiempo (días) desde el diagnóstico de CM al inicio de tratamiento		
Relativas a la CM			
localizacion	variable que recoge la zona de la columna que está afecta por CM. Substituye 4 antiguas variables booleanas *(loc_cervical/dorsal/lumbar/sacra*)	0	cervical
		1	dorsal
		2	lumbar
		3	sacra
		4	múltiple
n_vert_afect	nº de vértebras afectas (substituye 25variables booleanas (*vert_afect_C1/C2/.../D3)*		
n_vert_comp	nº de vértebras comprimidas (substituye 25variables booleanas (*vert_comp_C1/C2/.../D3)*		
vert_afect	variable categórica que recoge por intérvalos en número de vértebras afectas	0	1
		1	1-5
		2	>5
vert_comp	variable categórica que recoge por intérvalos en número de vértebras comprimidas	0	0
		1	1
		2	>1
valor_biomarcador	Variable que distingue la situación de que el biomarcador tumoral específico para esa CM sea positivo de cualquier otra situación	0	No disponible / No existe / Negativo
		1	Positivo
histologia**	histología del tumor. Se recodifican todas las categorías de manera que se observan las frecuencias y se eliminan todas aquellas que tienen un número menor de episodios de CM de 4. Esas categorías pasan a "alimentar" la categoría de otros.	0	Sin AP
		1	Adenocarcinoma
		2	Carcinoma escamoso
		3	Carcinoma de célula pequeña
		4	Células claras
		5	Mieloma múltiple
		6	Carcinoma no célula pequeña
		7	Carcinoma ductal infiltrante
		8	Carcinoma lobulillar infiltrante
		9	Carcinoma urotelial
		10	Hepatocarcinoma
		11	Linfoma
		12	Otros
Relativas a índices			
puntuacion_RADES	puntuación calculada a partir de las variables componentes de su definición		
RADES_Calc*	A efectos de cálculo. clasificación del paciente afecto de CM en la escalera RADES según la puntuación RADES calculada a partir de las variables de su definición	0	Grupo1 (20-30ptos)
		1	Grupo2 (31-35 ptos)
		2	Grupo3 (36-45ptos)
RADES	clasificación del paciente afecto de CM en la escalera RADES. Recoge el valor de *RADES_HC* y en su absencia de *RADES_Calc*	0	Grupo1 (20-30ptos)
		1	Grupo2 (31-35 ptos)
		2	Grupo3 (36-45ptos)
puntuacion_SINS	puntuación calculada a partir de las variables componentes de su definición		
SINS_Calc*	A efectos de cálculo. clasificación del paciente afecto de CM en la escalera SINS según la puntuación SINS calculada a partir de las variables de su definición	0	Estable (0-6ptos)
		1	Potencialmente inestable (7-12ptos)
		2	Inestable (13-18ptos)
SINS	clasificación del paciente afecto de CM en la escalera SINS. Recoge el valor de *SINS_HC* y en su absencia de *SINS_Calc*	0	Estable (0-6ptos)
		1	Potencialmente inestable (7-12ptos)
		2	Inestable (13-18ptos)
location_SINS*	A efectos de cálculo de la puntuación SINS. Recoge el tramo de la columna en el que aparece la CM. Substituye 8 antiguas variables booleanas (Nivel_M1_C1C2/C3C6/.../S2S5_SINS)	0	C7-D2 / D11-L1/L5-S1/occip-C2 (3ptos)
		1	C3-C6 / L2-L4 (2ptos)
		2	D3- D10 (1ptos)
		3	S2-S5 (0ptos)
Relativas al estado del paciente previo al tratamiento			
clinica_debut	Variable que recoge en distintas categorías el début clínico del paciente afecto de CM. Recoge las diversas combinaciones de 4 variables booleanas originales (*clinDebt_somatico/ neuropatico/ motora/ asintomatica*)	0	Asintomático
		1	Clínica sensitiva/ dolor neuropático
		2	Somático
		3	Clínica motora
		4	sens / neurop + somatico
		5	sens / neurop + motor
		6	somático + motor
		7	sens / neurop + somat + motor

Tabla 2 Variables y tabulación de nueva creación
**variables a efectos de cálculo*
***variables originales, sólo se han reticulado de sus categorías*

Variable	Definición	Categorías	
		Código	Significado
Relativas a datos identificativos del paciente			
ID	nº identificativo del episodio de compresión medular (CM)		
SAP	nº identificativo del paciente que sufre el episodio de CM		
edad	intérvalo de tiempo en años desde la fecha de nacimiento a la fecha de diagnóstico de CM		
genero	genero	0	hombre
		1	mujer
centro	centro donde se realiza el tratamiento aunque no la hospitalización	0	Hospital Germans Tries i Pujol
		1	Hospital Duran i Reynals
grupo	divide los pacientes según la fecha del diagnóstico de CM	0	Grupo1 (antes del 2018)
		1	Grupo2 (durante o después del 2018)
estatus_UC	estado del paciente en el último control previo al cierre del estudio	0	vivo o pérdida de seguimiento*
		1	muerto
Relativas a fechas temporales que describen el curso clínico del paciente			
survival	variable output, tiempo de supervivencia o censurado. Intérvalo de tiempo entre la finalización de tto (sea con CR o RT) y el último control o éxitus del paciente (*fecha_fin).* Para pacientes en soporte recoge el tiempo desde el diagnóstico a su muerte.		
t_ac	intérvalo de tiempo (días) desde el diagnóstico al aviso de CM		
t_sc	intérvalo de tiempo (días) desde la sospecha clínica de CM al diagnóstico		
t_itto	intérvalo de tiempo (días) desde el diagnóstico de CM al inicio de tratamiento		
Relativas a la CM			
localizacion	variable que recoge la zona de la columna que está afecta por CM. Substituye 4 antiguas variables booleanas *(loc_cervical/dorsal/lumbar/sacra*)	0	cervical
		1	dorsal
		2	lumbar
		3	sacra
		4	múltiple
n_vert_afect	nº de vértebras afectas (substituye 25variables booleanas (*vert_afect_C1/C2/.../D3)*		
n_vert_comp	nº de vértebras comprimidas (substituye 25variables booleanas (*vert_comp_C1/C2/.../D3)*		
vert_afect	variable categórica que recoge por intérvalos en número de vértebras afectas	0	1
		1	1-5
		2	>5
vert_comp	variable categórica que recoge por intérvalos en número de vértebras comprimidas	0	0
		1	1
		2	>1
histologia	histología del tumor. Se recodifican todas las categorías de manera que se observan las frecuencias y se eliminan todas aquellas que tienen un número menor de episodios de CM de 4. Esas categorías pasan a "alimentar" la categoría de otros.	0	Sin AP
		1	Adenocarcinoma
		2	Carcinoma escamoso
		3	Carcinoma de célula pequeña
		4	Células claras
		5	Mieloma múltiple
		6	Carcinoma no célula pequeña
		7	Carcinoma ductal infiltrante
		8	Carcinoma lobulillar infiltrante
		9	Carcinoma urotelial
		10	Hepatocarcinoma
		11	Linfoma
		12	Otros
CM_PrimarioConocido	existencia de un tumor primario conocido	0 / 1	no / si
valor_biomarcador	Variable que distingue la situación de que el biomarcador tumoral específico para esa CM sea positivo de cualquier otra situación	0	No disponible / No existe / Negativo
		1	Positivo
Relativas a índices			
tumPrimario_RADES	variable recogida a efectos del cálculo de la puntuación RADES. Información sobre el tipo de tumor primario que ha originado la CM	0	Mama (8ptos)
		1	Próstata (7ptos)
		2	Mieloma/Linfoma (9ptos)
		3	Pulmón (3ptos)
		4	Otros/orígen desc/sin AP(4ptos)
metOseas_RADES	variable a efectos del cálculo de la puntuación RADES. Existencia de metástasis óseas	0 / 1	no(7ptos) / si(5ptos)
metVisc_RADES	variable a efectos del cálculo de la puntuación RADES. Existencia de metástasis viscerales	0 / 1	no(8ptos) / si(2ptos)
estAmbulatorio_RADES	variable a efectos del cálculo de la puntuación RADES. Estado ambulatorio del paciente	0 / 1	no(3ptos) / si(7ptos)
intervalo_diagCM_RADES	variable a efectos del cálculo de la puntuación RADES. Intérvalo de tiempo entre el diagnóstico primario y la aparición de CM	0	=< 15meses (4ptos)
		1	> 15meses (7ptos)
tiempoDesarrollo_DefMotor_RADES	variable a efectos del cálculo de la puntuación RADES. Tiempo de desarrollo de déficits motores	0	De uno a siete (3ptos)
		1	De ocho a catorce (6ptos)
		2	Más de catorce (8ptos)
		3	Sin déficit motor / Subclínica (8ptos)
puntuacion_RADES	puntuación calculada a partir de las variables componentes de su definición		
RADES	clasificación del paciente afecto de CM en la escalera RADES. Recoge el valor de *RADES_HC* y en su absencia de *RADES_Calc*	0	Grupo1 (20-30ptos)
		1	Grupo2 (31-35 ptos)
		2	Grupo3 (36-45ptos)
dolor_SINS	A efectos del cálculo de la puntuación SINS. Variable que recoge el dolor que experimenta el paciente.	0	ocasional (1pto)
		1	sí (3ptos)
		2	sin dolor (0ptos)

caract_lesion_SINS	A efectos del cálculo de la puntuación SINS. Características de la lesión en la columna.	0	lítica (2ptos)
		1	mixta (1pto)
		2	blástica (0ptos)
alineacColumna_SINS	A efectos del cálculo de la puntuación SINS. Alineación de la columna.	0	subluxación o translación (4ptos)
		1	deformidad nueva (2ptos)
		2	normal (0ptos)
colapsoVertebral_SINS	A efectos del cálculo de la puntuación SINS. Colapso vertebral de la columna.	0	> 50% (3ptos)
		1	< 50% (2ptos)
		2	sin colapso,>50% cuerpo afecto(1pto)
		3	ninguna de las anteriores(0ptos)
afectPosterolat_SINS	A efectos del cálculo de la puntuación SINS. Afectación posterolateral de la columna.	0	bilateral (3ptos)
		1	unilateral (1pto)
		2	ninguna de las anteriores (0ptos)
puntuacion_SINS	puntuación calculada a partir de las variables componentes de su definición		
SINS	clasificación del paciente afecto de CM en la escalera SINS. Recoge el valor de *SINS_HC* y en su absencia de *SINS_Calc*	0	Estable (0-6ptos)
		1	Potencialmente inestable (7-12ptos)
		2	Inestable (13-18ptos)
Blisky	Clasificación en la escalera de Blisky, que valora el grado de afectación del canal medular en la CM	0	0
		1	A1
		2	A2
		3	A3
		4	B
		5	C
Relativas al estado del paciente previo al tratamiento			
clinica_debut	Variable que recoge en distintas categorías el début clínico del paciente afecto de CM. Recoge las diversas combinaciones de 4 variables booleanas originales (*clinDebt_somatico/ neuropatico/ motora/ asintomatica*)	0	Asintomático
		1	Clínica sensitiva/ dolor neuropático
		2	Somático
		3	Clínica motora
		4	sens / neurop + somatico
		5	sens / neurop + motor
		6	somático + motor
		7	sens / neurop + somat + motor
dolor_eval	escala de dolor numérica (del 0-10) que siente el paciente previamente al tratamiento		
tiempo_clinSensitiva_dia	intérvalo de tiempo en horas desde la aparición de clínica sensitiva al diagnóstico de CM		
tiempo_clinMotora_diag	intérvalo de tiempo en horas desde la aparición de clínica motora al diagnóstico de CM		
RT_descompPrevia	existencia de radioterapia previa descompresiva para ese paciente	0 / 1	no / si
Relativas al estado del paciente tras el tratamiento			
estatus_lesion_UC	Estado de la lesión de CM tras el tratamiento en el último control realizado antes del cierre del estudio	0	Estable
		1	Progresión en PTV
		2	Progresión fuera de PTV
		3	RC Radiológica
		4	RC Metabólica
		5	No evaluada
empeoramiento_neurologico	empeoramiento neurológico a 7 días tras la finalización del tratamiento	0 / 1	no / si
dolor_eval7	escala de dolor numérica (del 0-10) que siente el paciente 7 días tras finalizar el tratamien	0 / 1	no / si
estatusAmbulatorio7	estatus ambulatorio del paciente 7 días después de la finalización del tratamiento	0 / 1	no / si
tto_trasCM	tratamiento oncológico realizado tras finalizar el tratamiento	0	No inicia tto
		1	No continúa tto
		2	Continúa tto
		3	Inicia tto
Relativas al tratamiento			
fijacionPrevia_segIrradiar	Episodio de CM que requiere la estabilización previa del segmento antes del tratamiento	0 / 1	no / si
dosis_RT	Fraccionamiento de la dosis de radioterapia: número de sesiones y dosis por sesión	0	8Gy x 1sesión
		1	5Gy x 4sesiones
		2	4Gy x 5sesiones
		3	3Gy x 10sesiones
		4	otros
test_diagnostico	Test realizado para determinar el diagnóstico de CM	0	Resonancia Magnética Nuclear (RM)
		1	Tomografía computarizada (TC)
		2	otros
opción terapéutica	opción terapéutica realizada ante el evento de CM	0	Cirugía (CR)
		1	CR + RT
		2	RT + CR
		3	RT
		4	tratamiento farmacológico de

Tabla 3 Variables y tabulación de la base de datos final. En azul las variables de nueva creación.

3.2 Descripción visual de la base de datos

Nuestra base de datos refleja un estudio clínico longitudinal prospectivo, con fecha de inicio el 19 de mayo del 2009 y fecha de cierre el 6 de noviembre del 2020, que aglutina 290 episodios de CM reclutados entre el 19 de mayo del 2009 y el 12 de octubre del 2018. 276 pacientes tienen episodios únicos mientras que 6 de ellos tienen episodios múltiples. Las CM fueron diagnosticadas la gran mayoría mediante Resonancia Magnética Nuclear y tratadas en ICO Badalona o ICO Hospitalet. Un 79% (228) de las CM fueron tratadas anteriormente al 2018 (grupo1) y un 21% (62) durante ese año o posterior (grupo2). Entre las opciones terapéuticas el tratamiento más frecuente de manera aplastante (267 casos sobre un total de 290) es el tratamiento radioterápico exclusivo. La siguiente figura muestran gráficamente la distribución en grupos comentada y la distribución por centros. Así como la frecuencia de tratamientos entre las distintas opciones terapéuticas: cirugía (CR) 1%, radioterapia (RT) 92%, RT-CR 5%, CR-RT 2% y SOPORTE 0.3%, y el tipo de fraccionamiento en RT.

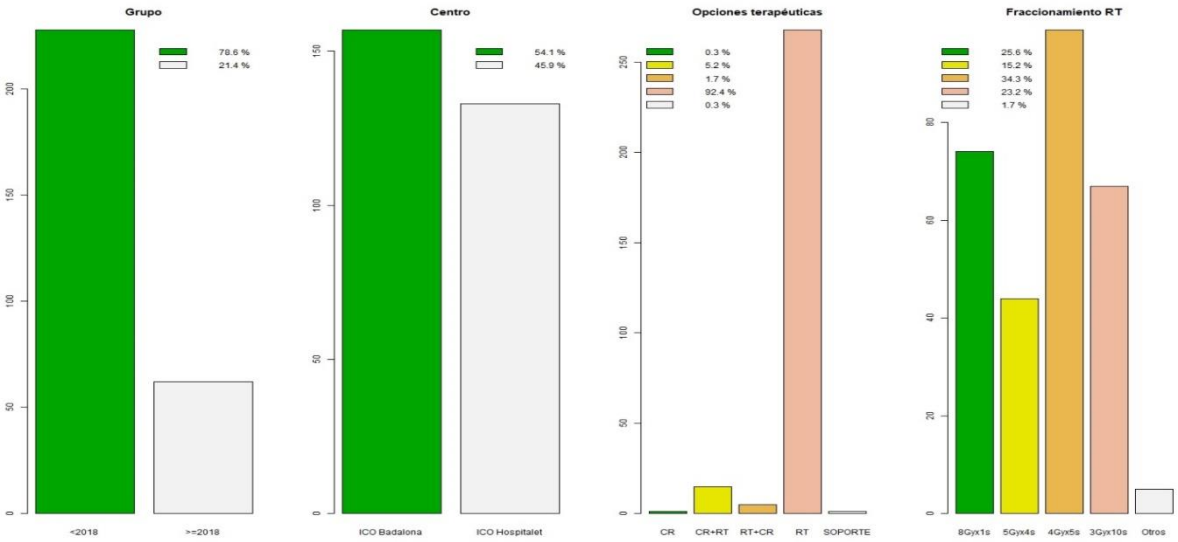


Figura1 Distribución de los episodios de CM por grupos (anterior o posterior al 2018), centro, opción terapéutica y fraccionamiento de RT de elección

Tenemos proporcionalmente muchos más pacientes hombres (210 hombres por 80 mujeres). La mediana es parecida a la media (alrededor de 66 años), con 10 años más o menos para el tercer o primer cuartil respectivamente, lo que indica una distribución aproximadamente normal en referencia a la edad. A fecha de cierre del estudio sólo 21 pacientes siguen vivos lo que constituye pérdidas aceptables des del punto de vista del cálculo de la supervivencia. En referencia a la supervivencia la media (293 días) es muy superior a la mediana (74 días) e incluso al tercer cuartil (227 días), lo que significa que hay una mayoría de pacientes que viven menos de 3 meses y una minoría que viven más de un año.

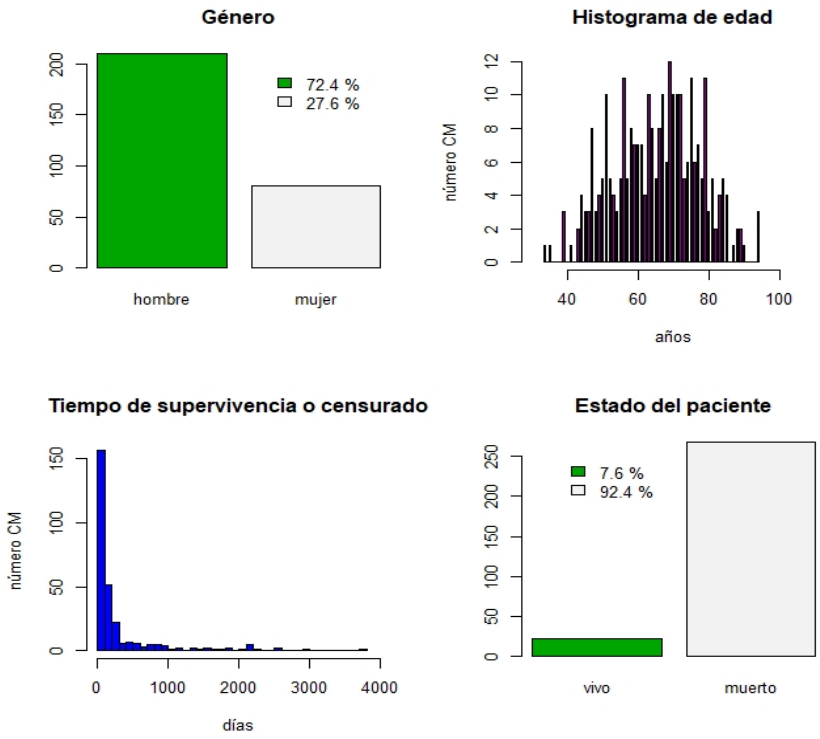


Figura2 Distribución de los episodios de CM por género y edad. Supervivencia y estatus del paciente a cierre del estudio.

En lo referente a la CM, la localización más frecuente es de manera evidente la dorsal (188 casos), seguida a distancia de la lumbar (50) y la cervical (39). De la

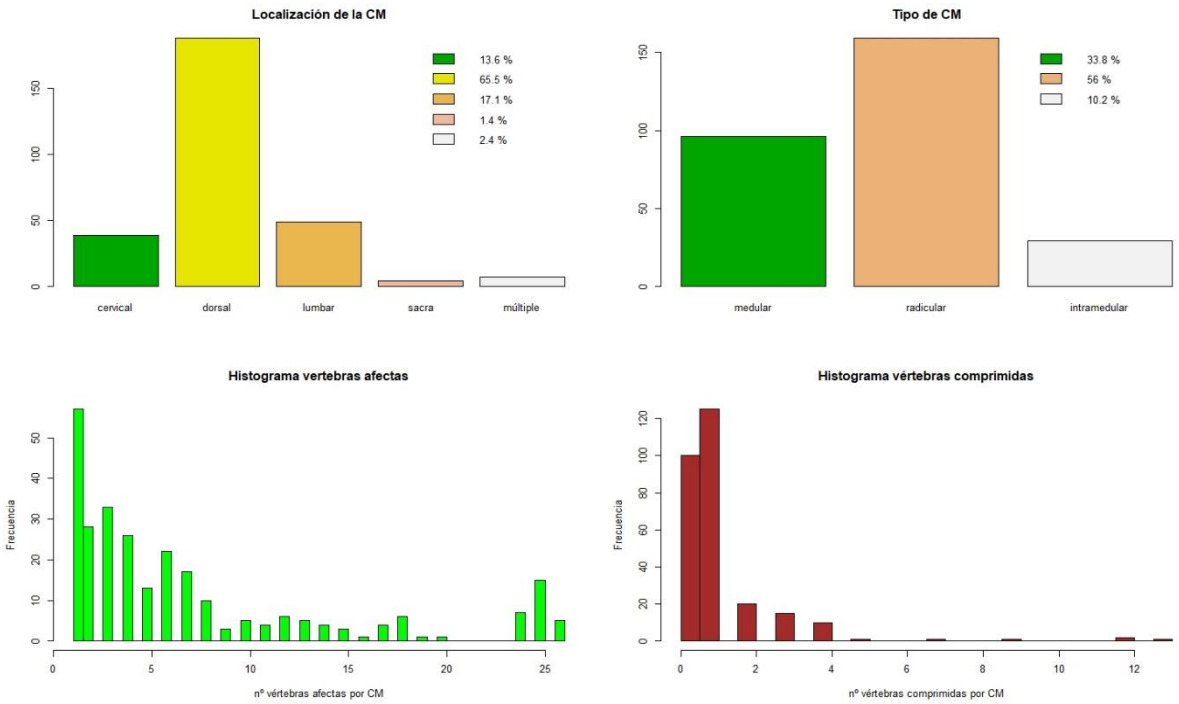


Figura3 Datos relativos a la CM

misma forma, las CM más frecuentes son las radiculares (159), seguidas de las medulares (96) e intramedulares. Las dos situaciones más frecuentes en las CM son: en lo referente a vértebras afectas, entre 1-5(100) o más de 5 (119), en lo referente a vértebras comprimidas, ninguna (100) o una vértebra (125).

En referencia a la histología del tumor, el tumor más frecuente (43% de los casos) es el adenocarcinoma, seguido a distancia por el carcinoma de célula pequeña (10%). En la mayoría de los casos el valor del biomarcador (EGFR, ALK, PDL1, RE, RP, Her2/neu, PSA) o bien es negativo o no se halla disponible (235) frente a los de valor positivo (50). Y también en la mayoría de los casos (200 frente 87), el tumor primario de la CM es conocido. Sólo en 19 de los casos el paciente había recibido radioterapia con intención descompresiva previamente.

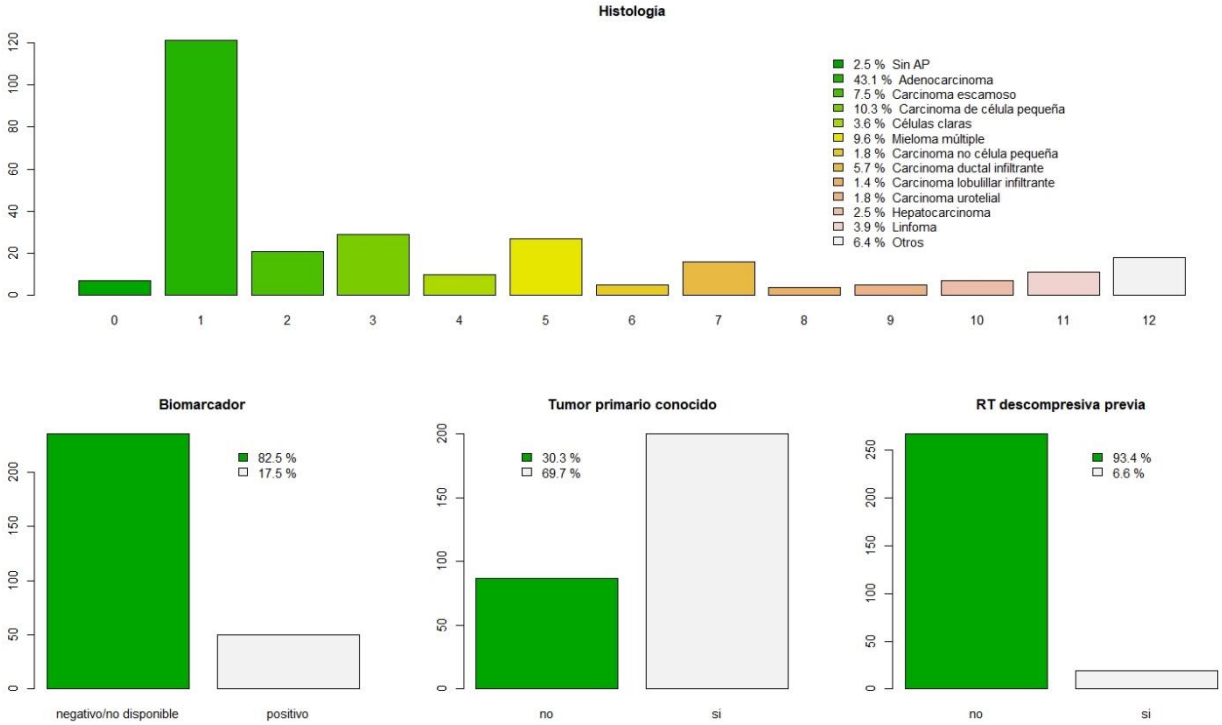


Figura4 Distribución de la histología del tumor, el valor del biomarcador, existencia de un tumor primario conocido y existencia de radioterapia previa compresiva

En cuanto al estado del paciente previo a la CM, se produce un debut en la mayoría de los casos somático (108 casos). De no ser así, se distribuye de manera cuasi equitativa entre las demás categorías. En cuanto a las opciones de tratamiento oncológico tras la CM se distribuyen por igual entre todas las opciones. En referencia al estado post tratamiento, en la gran mayoría de casos no hay estatus ambulatorio (60%) ni empeoramiento neurológico (256 vs 22).

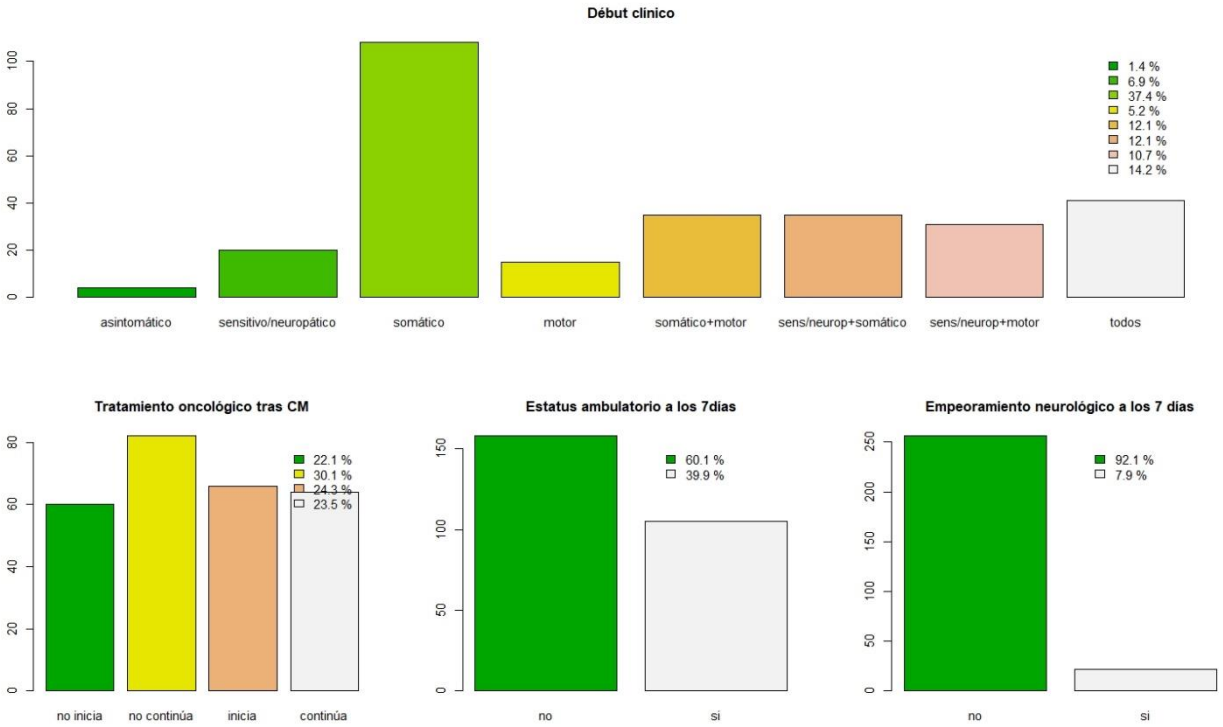


Figura 5 Estado del paciente previo y post tratamiento

En referencia a los índices tenemos una muestra dónde casi 2/3 están clasificados como RADES 1 (entre 20 y 30 puntos), más de la mitad muestra un SINS catalogado como potencialmente inestable (con un 13% de NA) y en referencia a la escalera Blisky una mayoría de B y C (existe un 14% en este caso de NA).

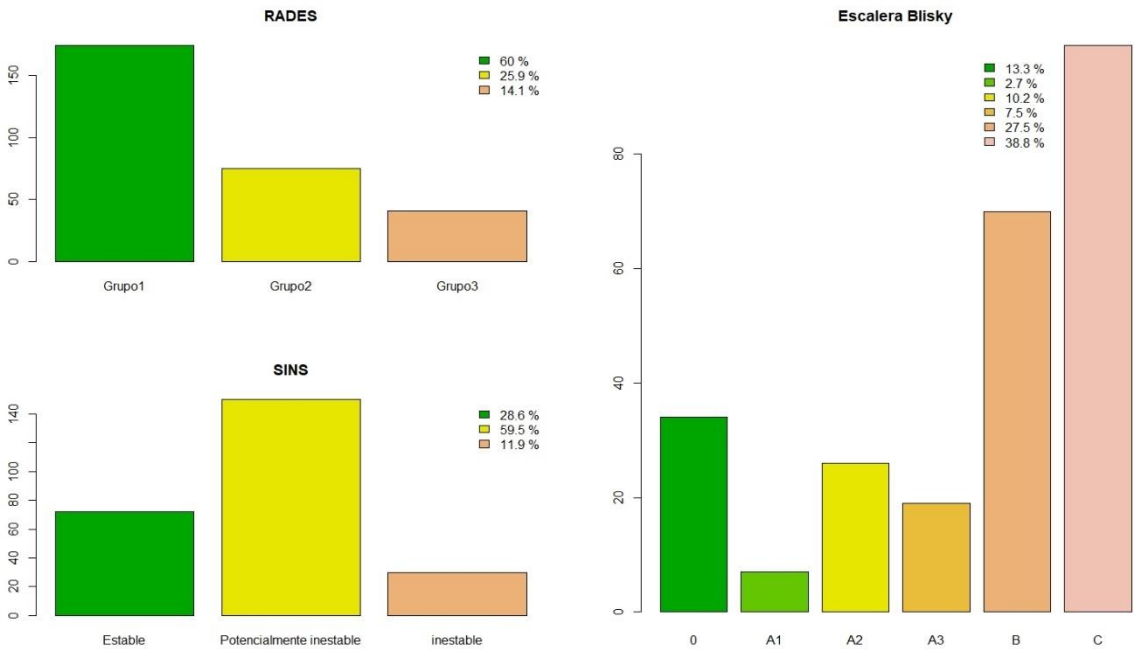


Figura 6 Índices de RADES, SINS y Blisky

Finalmente estamos hablando de una base de datos de 49 variables, la mayoría de las cuales son variables factor. Son variables numéricas la edad, la supervivencia, el número de vértebras afectas o comprimidas, el tiempo del diagnóstico al aviso de compresión, el tiempo des de la sospecha clínica al diagnóstico, el tiempo des del diagnóstico al inicio de tratamiento y la puntuación RADES o SINS. A la derecha podemos observar los scatterplots bivariado y el histograma de la distribución en la diagonal

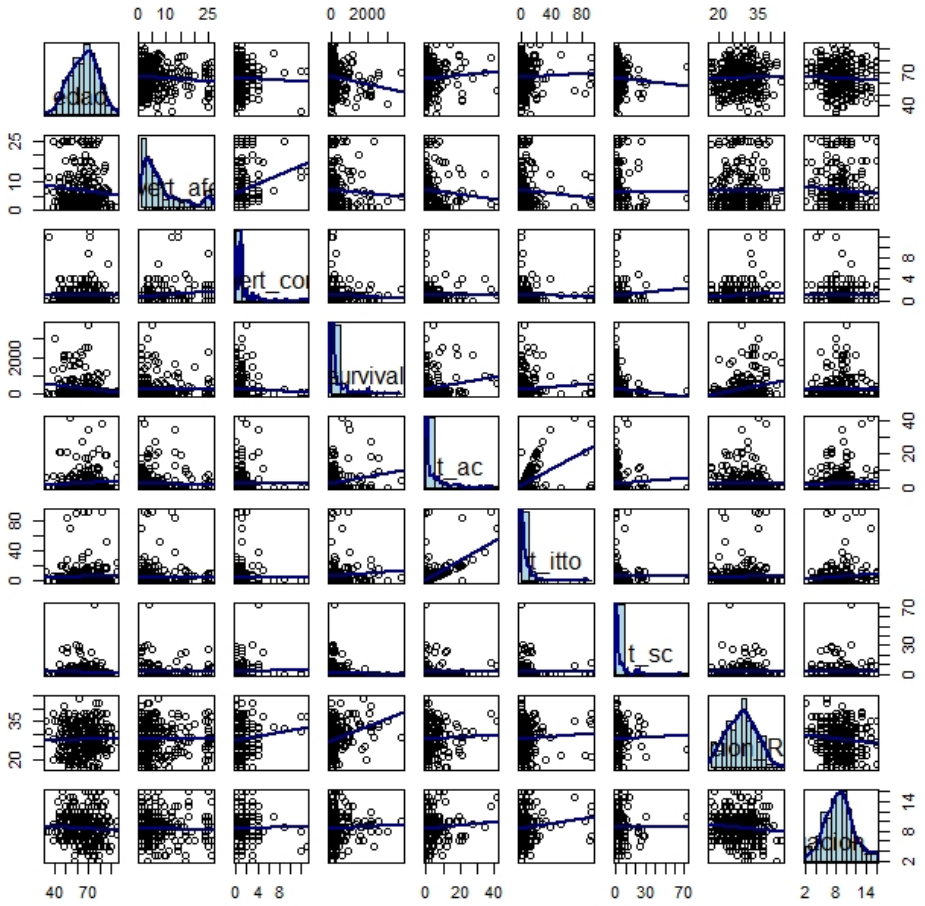


Figura 7 Scatterplot bivariado con histograma en la diagonal

Exceptuando la edad y las puntuaciones RADES y SINS, todas las demás muestran distribuciones alejadas de la normal que deberían ser transformadas previamente a un análisis. No parece que se pueda apreciar ninguna correlación lineal sólida entre la variable independiente supervivencia y las demás variables. Tampoco se encuentra que las variables numéricas entre ellas se hallen relacionadas. En las dos figuras más abajo podemos observar una correlación pobre de sentido positiva ente el tiempo de aviso de compresión y el inicio de tratamiento (0.6).

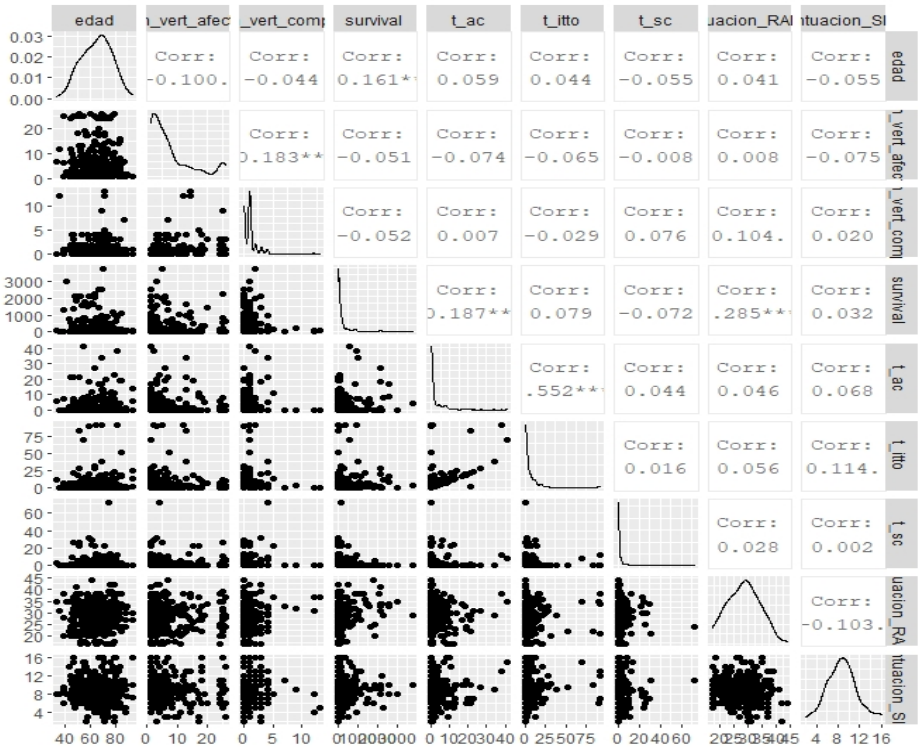


Figura 8 Scatterplot bivariado con histograma en la diagonal y coeficiente de correlación a su derecha

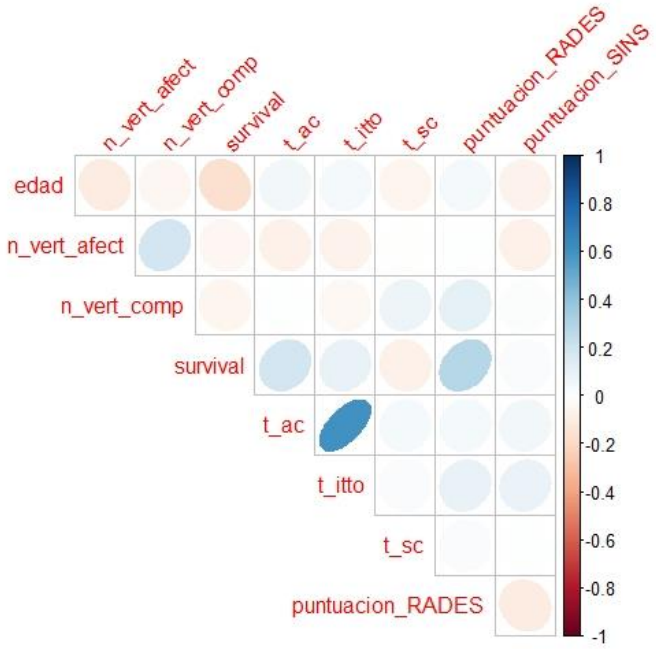


Figura 9 Imagen visual de la correlación entre las distintas variables

Cuando graficamos las cajas de dispersión de nuestras variables numéricas, podemos ver las variables edad, puntuación RADES y SINS presentan una distribución equidistante de la mediana tanto para los cuartiles como para el mínimo y el máximo, y son las únicas con pocos/ningún outlier. Básicamente, podemos decir que presentan distribuciones simétricas no sesgadas. Una buena

medida de tendencia central seria la media (que coincidirá con la mediana y con la moda) y una buena medida de la dispersión la desviación estándar.

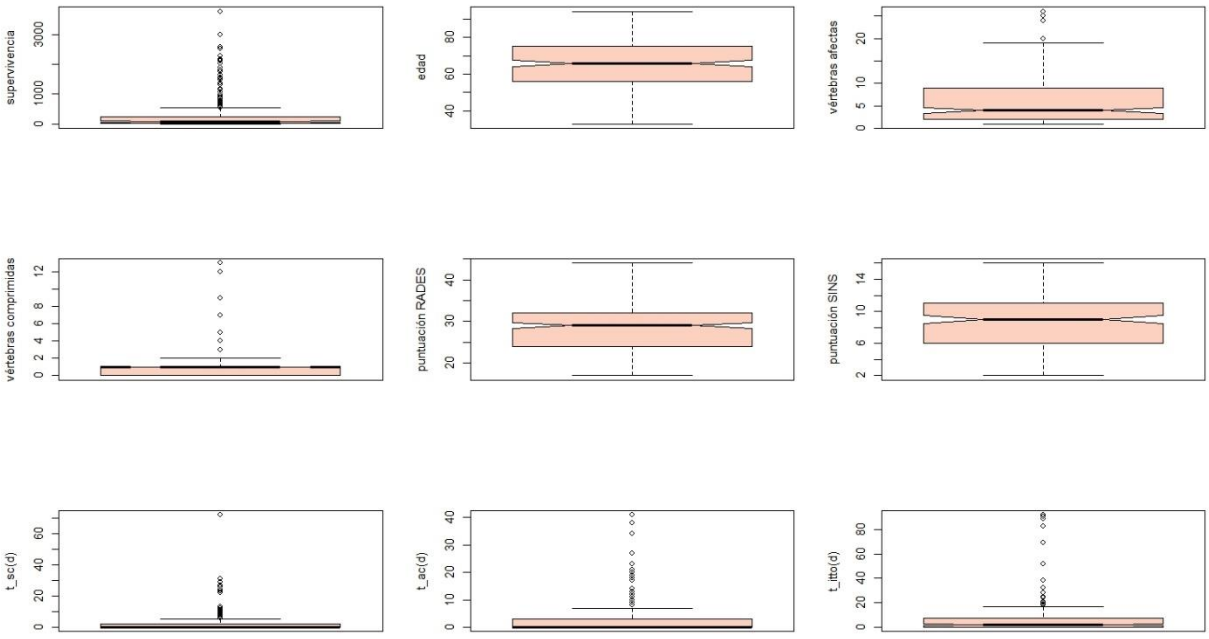


Figura 10 Cajas de dispersión para las distintas variables numéricas

Por el contrario, las variables de supervivencia, vértebras comprimidas, vértebras afectas, tiempo de sospecha clínica al diagnóstico, tiempo de aviso de la compresión al diagnóstico y tiempo del diagnóstico al inicio del tratamiento presentan histogramas sesgados a la derecha con lo que las medias suelen coincidir con el valor del tercer cuartil. Además, estas distribuciones no sólo tienen muchísimos outliers, sino que también tienen valores extremales. Existe una concentración muy fuerte de los valores en el rango inferior de la variable y una dispersión muy alta en los valores superiores del rango. Una buena medida de la tendencia central para estas distribuciones sería la mediana mientras que una buena medida de dispersión con tantos extremales sería el rango entre cuartiles.

3.3 Tratando con los valores missing

Existen 4 variables con un porcentaje de NA alrededor del 50%. Son las variables de evaluación del dolor (*eval_dolor)* tanto previo al tratamiento como posterior a este *eval_dolor7)* y las variables que recogen el tiempo en horas desde la aparición de clínica sensitiva y motora (respectivamente) hasta el diagnóstico (*tiempo_clinSensitiva_diag, tiempo_clinMotora_diag)*. Este porcentaje de NA es

muy superior al 35% estimado, a partir del cual es coherente para la fiabilidad del análisis estadístico eliminar estas variables. En referencia a la variable *Estatus_lesión_UC*, la categoría más frecuente con un casi 50% de los datos es la de “No evaluada”. En la práctica estadística, la categoría “No evaluada” es la forma clínica de definir un *missing.* Eliminamos las 5 variables explicadas. Nuestra base de datos pasa a contener ahora 44 variables.

En la base de datos existen 561 datos *missing* que representan un 4% de los datos, de los cuales 78 datos (que representa el 2,7%) pertenecen a variables numéricas. Vamos ahora a estudiar el porcentaje por variables:

```
ID SAP edad genero centro estatus_UC grupo survival t_ac t_sc t_itto
 0   0  1.4      0      0          0     0        0  0.3    0      0
localizacion tipo_CM vert_afect vert_comp n_vert_afect n_vert_comp
           1     2.1        4.8       4.8          4.8         4.8
histologia CM_PrimarioConocido valor_biomarcador RADES RADES_Calc
       3.1                   1               1.7     0          0
puntuacion_RADES tumPrimario_RADES metOseas_RADES metVisc_RADES
               0                0              0             0
intervalo_diagCM_RADES estAmbulatorio_RADES
                     0                    0
tiempoDesarrollo_DefMotor_RADES SINS SINS_Calc puntuacion_SINS
                            0.7 13.1      15.5            15.5
location_SINS dolor_SINS caract_lesion_SINS alineacColumna_SINS
          0.3       15.2               15.9                15.9
colapsoVertebral_SINS afectPosterolat_SINS Blisky clinica_debut
                 15.5                 15.5   12.1          0.3
RT_descomprPrevia estatus_lesion_UC empeoramiento_neurologico
              1.4               3.8                       4.1
estatusAmbulatorio7 tto_trasCM fijacionPrevia_SegmIrradiar dosis_RT
                9.3        6.2                          2.8      0.3
test_diagnostico opcion_terapeutica
               0                  0
```

Figura11 Porcentaje de missings para cada variable

Como podemos ver hay muchas variables que el porcentaje es prácticamente 0 y en todas se mantiene por debajo del 5% exceptuando en el valor de SINS (alrededor del 13% de missings), de la escalera Blisky (12%) y del tratamiento realizado tras la CM (6%). Aunque valores inferiores al 5% podrían ser simplemente omitidos, como existen un par de variables con valores de missings superiores al 10% consideramos que lo mejor es substituir esos valores missings por una aproximación a primeros vecinos. El algoritmo busca los k casos más parecidos y realiza un promedio ponderado de todos ellos. Efectivamente, comprobamos que después de dicha sustitución no hay ningún missing.

3.4 Descripción analítica y transformación de las variables cualitativas

En el caso de variables numéricas, si el histograma es próximo a una gaussiana normal es recomendable como medida de tendencia central la media (que coincide con la mediana) y como medida de dispersión la desviación estándar o varianza. Si el histograma es sesgado a la derecha o izquierda se recomienda como medida de tendencia central la mediana y como medida de dispersión el

rango. En el caso de que haya muchos outliers o valores extremos se usará el rango intercuartil en vez del rango. Ya hemos visto que las variables de edad, puntuación Rades y puntuación SINS tienen histogramas cercanos a la normal mientras que las demás variables presentan histogramas sesgados a la izquierda.

	variable	na	mean	sd	IQR	skewness	kurtosis	p05	p50	p75	p99
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	edad	4	65.2	12.2	18.8	-0.104	-0.561	45.2	66	74.8	9.06e1
2	n_vert_a~	14	7.16	7.30	7	1.44	0.946	1	4	9	2.60e1
3	n_vert_c~	14	1.12	1.67	1	4.10	22.8	0	1	1	9.75e0
4	survival	0	293.	553.	202.	3.13	10.8	6	74.5	227.	2.54e3
5	t_ac	1	2.77	5.90	3	3.43	14.4	0	0	3	2.78e1
6	t_itto	0	5.89	12.5	7	4.85	27.7	0	2	7	8.37e1
7	t_sc	0	2.33	6.44	2	6.07	51.6	0	0	2	2.72e1
8	puntuaci~	0	28.2	5.73	8	0.0420	-0.557	18	29	32	4.11e1
9	puntuaci~	45	8.70	2.99	5	0.228	-0.231	4	9	11	1.60e1

Tabla4 Medidas de tendencia central, dispersión, normalidad y simetría de las distribuciones de variables numéricas

	vars	statistic	p_value	sample
	<chr>	<dbl>	<dbl>	<dbl>
1	edad	0.991	9.25e- 2	290
2	n_vert_afect	0.784	2.67e-19	290
3	n_vert_comp	0.567	3.04e-26	290
4	survival	0.550	1.18e-26	290
5	t_ac	0.533	4.55e-27	290
6	t_itto	0.461	1.17e-28	290
7	t_sc	0.387	3.87e-30	290
8	puntuacion_RADES	0.986	6.34e- 3	290
9	puntuacion_SINS	0.980	4.86e- 4	290

Tabla5 Test de Shapiro-Wilk para la normalidad de las variables. Valores del p-valor inferiores a 0.05 nos llevan a rechazar la hipótesis nula de normalidad de las distribuciones

La edad presenta una media de 65años con 12 años de desviación estándar, la puntuación de RADES una media de 28 puntos (Rades1) con 6 puntos de desviación estándar y la puntuación SINS de 8 puntos (potencialmente inestable) con una desviación de 3 puntos. Las tres distribuciones presentan una simetría y una curtosis cercana a 0. Sin embargo, sólo la edad pasa el test de Shapiro-Wilk en lo referente a la normalidad.

La variable independiente *supervivencia* tiene una mediana de 74 días con una distancia intercuartil muy elevada de 202 días. Su valor para la simetría es de 3 con una curtosis alrededor de 10. El test de Shapiro-Wilk es altamente significativo.
El número de vértebras afectas / comprimidas tiene una mediana respectivamente de 4 y 1 vértebra con distancia intercuartil de 7 y 1. Mientras que los valores de simetría y curtosis son moderados para el número de vértebras afectas, toman valores mucho más elevados (4 para la simetría y 23 para la curtosis) en el caso de vértebras comprimidas. Ninguno de los dos “pasa” el test de Shapiro-Wilk para la normalidad.

En referencia a los tiempos todos muestran valores de simetría y curtosis elevados de acuerdo a los histogramas que hemos visto anteriormente, siendo

la variable del tiempo de la sospecha clínica la más perjudicada con curtosis de 50 y simetría de 6. Las tres son altamente significativas para el test de normalidad. Muestras medianas de 0 para la sospecha clínica y el aviso de CM, y de 2 para el inicio de tratamiento. Las distancias intercuartiles son respectivamente de 2, 3 y 7 días.

Vamos a visualizar la “normalidad” de las variables mediante un gráfico Q-Q. Un gráfico Q-Q compara dos distribuciones de probabilidad representando el valor de cada uno de sus cuantiles en cada uno de los ejes. Es decir, un punto (x,y) cualesquiera del gráfico toman el valor en el eje x de uno de los cuantiles de la primera distribución de probabilidad mientras que en el eje y toma el valor del mismo cuantil pero esta vez de la segunda distribución de probabilidad. Así, cuando se comparan distribuciones parecidas el gráfico Q-Q debería asemejarse a una recta de pendiente 1. Un gráfico Q-Q puede ser una buena manera de estudiar la normalidad o no de una distribución, si esa distribución se compara contra la distribución teórica normal. Puntos externos a la envolvente de la línea que atraviesa la nube de puntos pueden ser considerados como alejados de la normal.

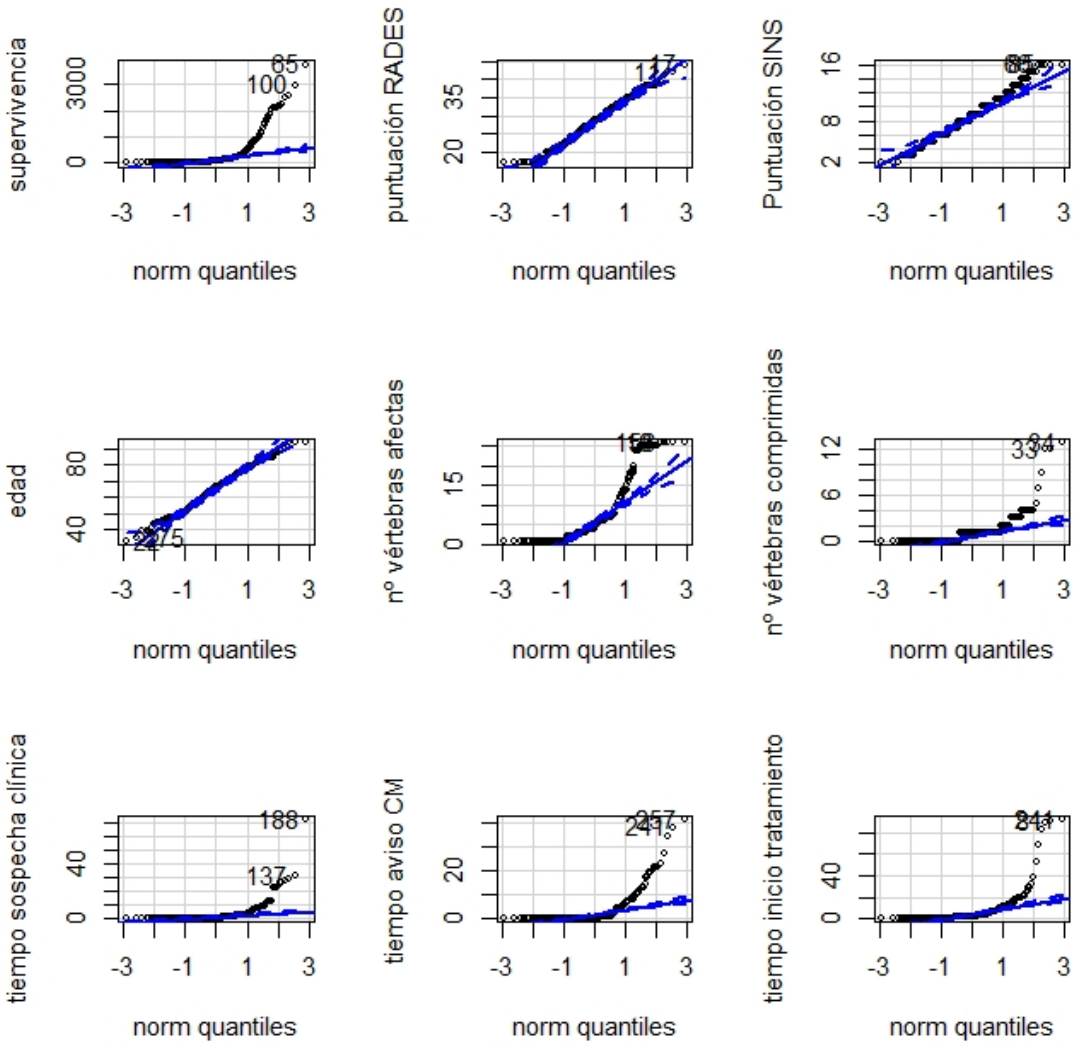


Figura 12 Gráficos Q-Q para las variables numéricas

Como podemos ver, sólo para la edad y quizás la puntuación de RADES el comportamiento se acerca a una distribución normal. En el caso, de la puntuación SINS el comportamiento se aleja de la normal para cuantiles elevados. Y en todos los demás casos (supervivencia, vértebras afectas, vértebras comprimidas, t_ac, t_sc, t_itto) tanto en cuantiles inferiores, pero sobre todo en los superiores los puntos se alejan muchísimo de la normalidad.

Vamos a probar de "normalizar" nuestras variables de forma conveniente. Para ello usaremos la escalera de potencias de Tukey que usa el test de Shapiro-Wilk de manera iterativa para encontrar el valor de lambda que aproxima más la distribución a una distribución normal. En la siguiente tabla encontramos la transformación asociada a cada valor de lambda.

λ	-2	-1	-1/2	0	1/2	1	2
y	$\frac{-1}{x^2}$	$\frac{-1}{x}$	$\frac{-1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	x	x^2

Tukey's Ladder of Powers *lamda values and corresponding power transforms. Lambda*

Tabla 6 Escalera de poderes de Tukey

Asemejando lambda a los valores más parecidos en la tabla obtenemos que para normalizar las variables cabría realizar:
* una transformación logarítmica: supervivencia, n_vert_afect
* una transformación sqrt : n_vert_comp, t_sc, t_ac, t_itto
* una transformación lineal (ninguna): RADES, SINS

Observemos que en el caso de los tiempos el valor de lamba (0.3) está prácticamente equidistante de 0 y 0.5, aunque más cercano a 0.5. Una transformación logarítmica (lambda igual a 0) sería imposible ya que los tiempos toman en infinidad de ocasiones el valor de 0 (0días) cuyo logaritmo o transformada no existiría. Para el caso de la distribución puntuación SINS se halla equidistante entre una trasformación línea o una de raíz cuadrada. Como la distribución no tiene outliers y es bastante simétrica decidimos no transformarla.

3.5 Outliers y puntos extremos

Consideramos un outlier como aquel punto que está a más de 1.5 la distancia intercuartil del cuartil más cercano (primero o tercero) y un extremal a más de 3 veces esa distancia. Hemos visto anteriormente en las cajas de dispersión de nuestras variables numéricas, que no sólo existen muchísimos outliers para casi todas las variables, sino que también existen extremals. Efectivamente en las cajas de dispersión de la siguiente figura vemos que con las nuevas variables transformadas existen menos outliers, ninguno para las variables de supervivencia, puntuación RADES, edad y número de vértebras afectas, distribuciones tras la transformación asimilables a una normal. La supervivencia y las vértebras afectas que tenían respectivamente 44 y 29 potenciales outliers, pasan a ser 0 tras ser las variables transformadas. Las variables de tiempo (t_sc,

t_ac, t_itto) y el número de vertebras afectas también mejoran con la transformación. Aunque algunos potenciales outliers se mantienen, la gran mayoría desaparecen pasando de 35 a 12, 37 a 11, 24 a 7 y 12 a 5 outliers respectivamente. Contamos 13 outliers para la puntuación SINS, que no ha sido transformada.

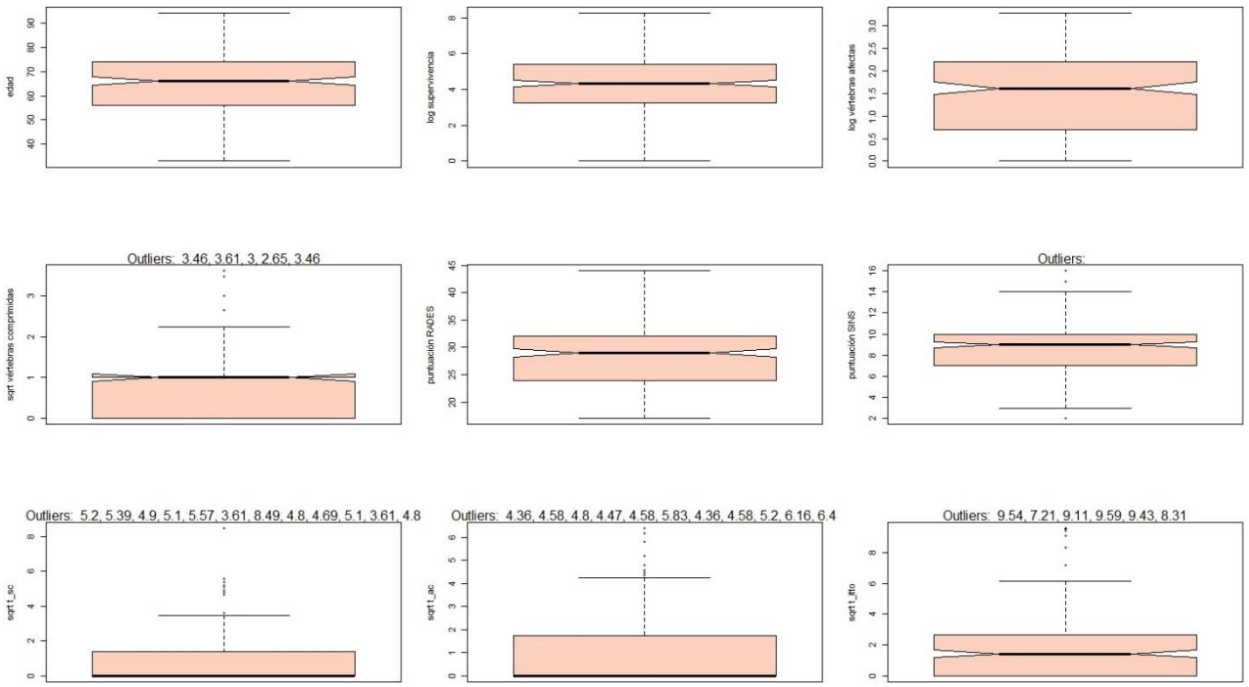


Figura 13 Cajas de dispersión de las variables numéricas transformadas

Para definir exactamente los outliers y valores extremos queremos usar el test de Rosner del paquete *EnvStats* que nos permite detectar de una vez múltiples outliers. Lamentablemente presupone la normalidad de las distribuciones en absencia de outliers. Tras un test de Shapiro Wilk en las distribuciones excluyendo los outliers, todas las pruebas son significativas por las que tenemos que rechazar la hipótesis de normalidad y no puede usarse el test de Rosner. Consideramos todos los “potenciales” outliers como outliers verdaderos. Intentamos ajustar las distribuciones mediante una distribución de Poisson, típica de distribuciones temporales, para poder “reconocer” nuestros outliers, pero nuestras distribuciones no se ajustan a dicha distribución teórica. Lo más sensato sería analizar nuestros outliers uno por uno, sopesar si podría ser un error y decidir si deben ser eliminados o remplazados (por la media/moda/mediana, imputación a k-vecinos o capados al percentil 5%-95% si el outlier es inferior/superior respectivamente). Estudiamos nuestros datos y vemos que los outliers de la puntuación SINS son todos aquellos con puntuaciones 2, 15 y 16. Considerando que una categoría es de 0 a 6 y la última de 13 a 18, parecen datos que "caen" en medio de esas categorías, parecen razonables. De la misma forma los outliers de vertebras comprimidas pertenecen a pacientes que tienen toda la columna afecta y comprimida. En referencia a los tiempos, tras revisar cada caso en concreto, en muchos casos los altos valores para *t_sc* son subproducto de limitaciones en la gestión hospitalaria, debidos a “atascos” en el acceso a la RM para diagnosticar la CM. Los tiempos largos de

t_ac responden a pacientes que tras el diagnóstico son derivados automáticamente y no se realiza el aviso de manera que el paciente se queda "colgado" en el sistema. Es más difícil saber qué está sucediendo con los tiempos largos de inicio de tratamiento, aunque en muchos casos suelen estar relacionados con los dos anteriores. Aunque el valor de los outliers podría parecer real, como la preparación de la base de datos ha sido tan problemática y tantos usuarios han intervenido en su creación, creemos que en espera de poder revisar cada caso con un médico lo más sensato es eliminar cualquier duda y eliminar todos los outliers de nuestra muestra. Se eliminan 34 episodios de CM con ID:

```
> cleanCM_T$ID[c(outt_acT, outt_ittoT, outt_scT, outSINST, outVertAfT, outVertCT)]
 [1]  66  82  99 121 162 164 224 225 230 269 288  86 189 225 269 283 288   1  69
[20]  98 135 145 197 200 237 243 271 319 320  38  39  80 279 307
```

Tabla 7 ID de episodios de CM que son outliers en alguna de las distribuciones de las variables

3.6 La nueva base de datos

La nueva base de datos contiene las variables numéricas transformadas. Ha sido necesario eliminar 2 ID cuya supervivencia era de 0 días, dado que su logaritmo es asintótico. Tras la eliminación de esos Id y de los outliers nuestra base de datos contiene 256 observaciones y 45 variables.

	variable	na	mean	sd	IQR	skewness	kurtosis	p05	p50	p75	p99
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	log.survi~	0	4.43	1.64	2.18	0.0888	-0.359	1.79	4.32	5.43	7.84
2	edad	0	65.0	12.0	18	-0.134	-0.562	45.4	66	74	89.1
3	puntuacio~	0	28.2	5.75	8	0.0457	-0.566	18	29	32	41.1
4	puntuacio~	0	8.70	2.74	3	0.192	0.155	4	9	10	16
5	log.n_ver~	0	1.50	1.01	1.51	0.0155	-0.975	0	1.61	2.2	3.26
6	sqrt.n_ve~	0	0.792	0.694	1	0.696	1.28	0	1	1	3.06
7	sqrt.t_sc	0	0.827	1.29	1.41	2.17	6.09	0	0	1.41	5.22
8	sqrt.t_ac	0	0.954	1.37	1.73	1.51	1.83	0	0	1.73	5.28
9	sqrt.t_it~	0	1.70	1.74	2.65	1.74	4.65	0	1.41	2.65	9.15

	vars	statistic	p_value	sample
	<chr>	<dbl>	<dbl>	<dbl>
1	log.survival	0.991	8.83e- 2	288
2	edad	0.990	5.73e- 2	288
3	puntuacion_RADES	0.986	5.91e- 3	288
4	puntuacion_SINS	0.981	8.22e- 4	288
5	log.n_vert_afect	0.938	1.15e- 9	288
6	sqrt.n_vert_comp	0.816	8.76e-18	288
7	sqrt.t_sc	0.690	1.06e-22	288
8	sqrt.t_ac	0.737	4.45e-21	288
9	sqrt.t_itto	0.832	5.21e-17	288

Tablas 8 Descripción de las variables numéricas transformadas y de su normalidad mediante test de Shapiro-Wilks

Como podemos observar la simetría y curtosis de todas las distribuciones mejora drásticamente. Como era de suponer los valores de la distancia intercuartil en referencia al de la mediana, se achican también. Sin embargo, sólo el logaritmo de la supervivencia a parte de la edad cumple la hipótesis nula de normalidad de su distribución. Las demás variables transformadas siguen mostrando p-valores altamente significativos. Cuando realizamos un gráfico Q-Q nos damos cuentas que valores de cuartiles elevados se alejan especialmente de la normal para la puntuación SINS y las variables transformadas de número de vértebras comprimidas y las relativas a los tiempos.

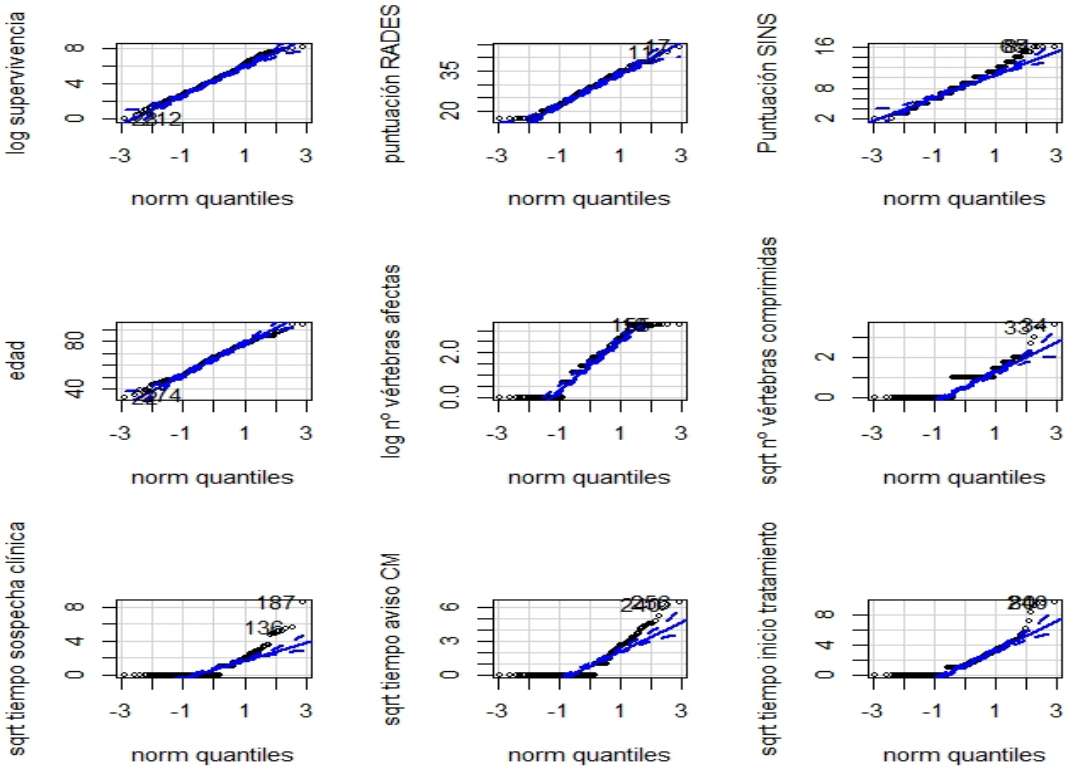


Figura 14 Gráficos Q-Q para las variables transformadas

De forma más intuitiva podemos observar su distribución mediante cajas de dispersión. Tras la transformación la supervivencia muestra un comportamiento normal, el *t_itto* presenta un comportamiento “cuasi-normal”, los histogramas de vértebras comprimidas y afectas presentan distribuciones sesgadas respectivamente a la derecha y a la izquierda

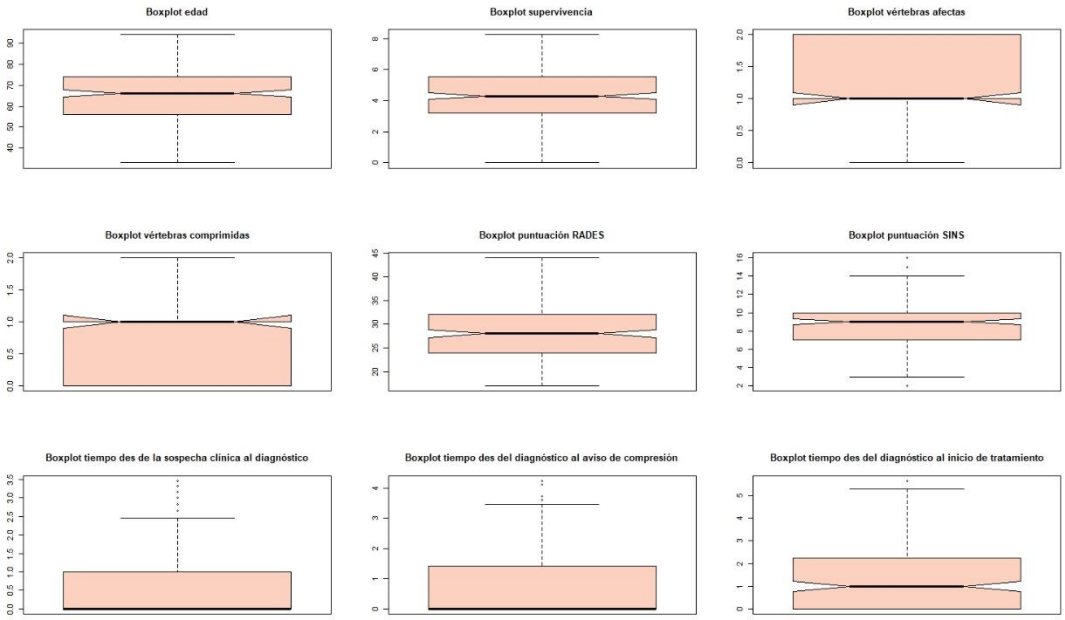


Figura 15 Cajas de dispersión de las variables transformadas

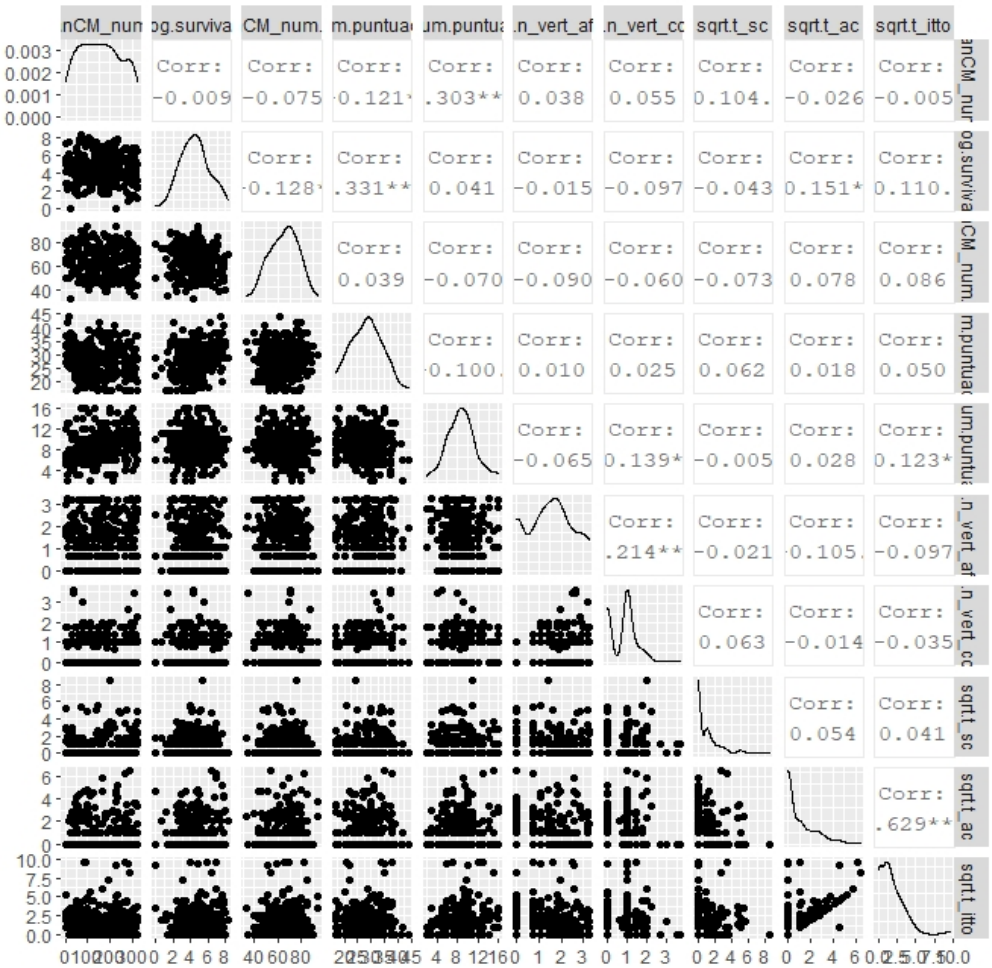


Figura 16 Scatter bivariado con histogramas en la diagonal y coeficientes de correlación entre las distintas variables

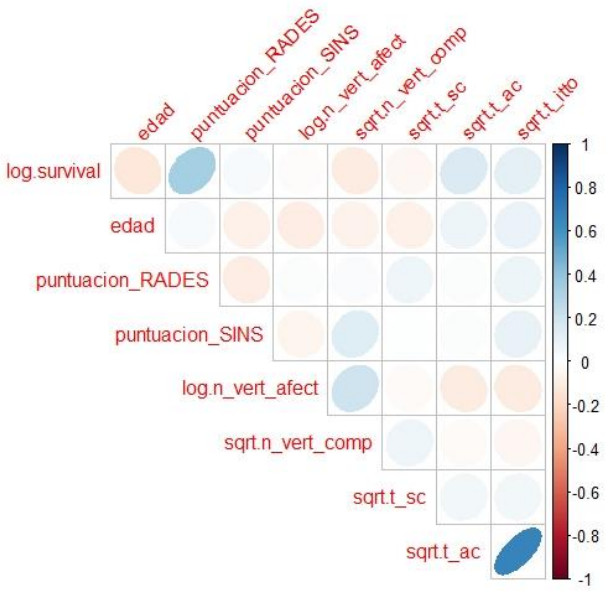


Figura 17 Representación visual de la correlación entre las variables

En el scatterplot bivariado podemos observar el comportamiento más cercano a la normalidad de las variables. La correlación entre el aviso de compresión y el inicio de tratamiento se hace un poco más fuerte superando el 0.6. En la representación visual mediante colores observamos que aparece una cierta correlación entre la puntuación RADES y el logaritmo de la supervivencia.

Cuando observamos el comportamiento de la supervivencia en función a las distintas categorías de las variables categóricas, vemos que en la mayoría de las situaciones los “notch” se solapan y no podemos establecer ninguna diferencia

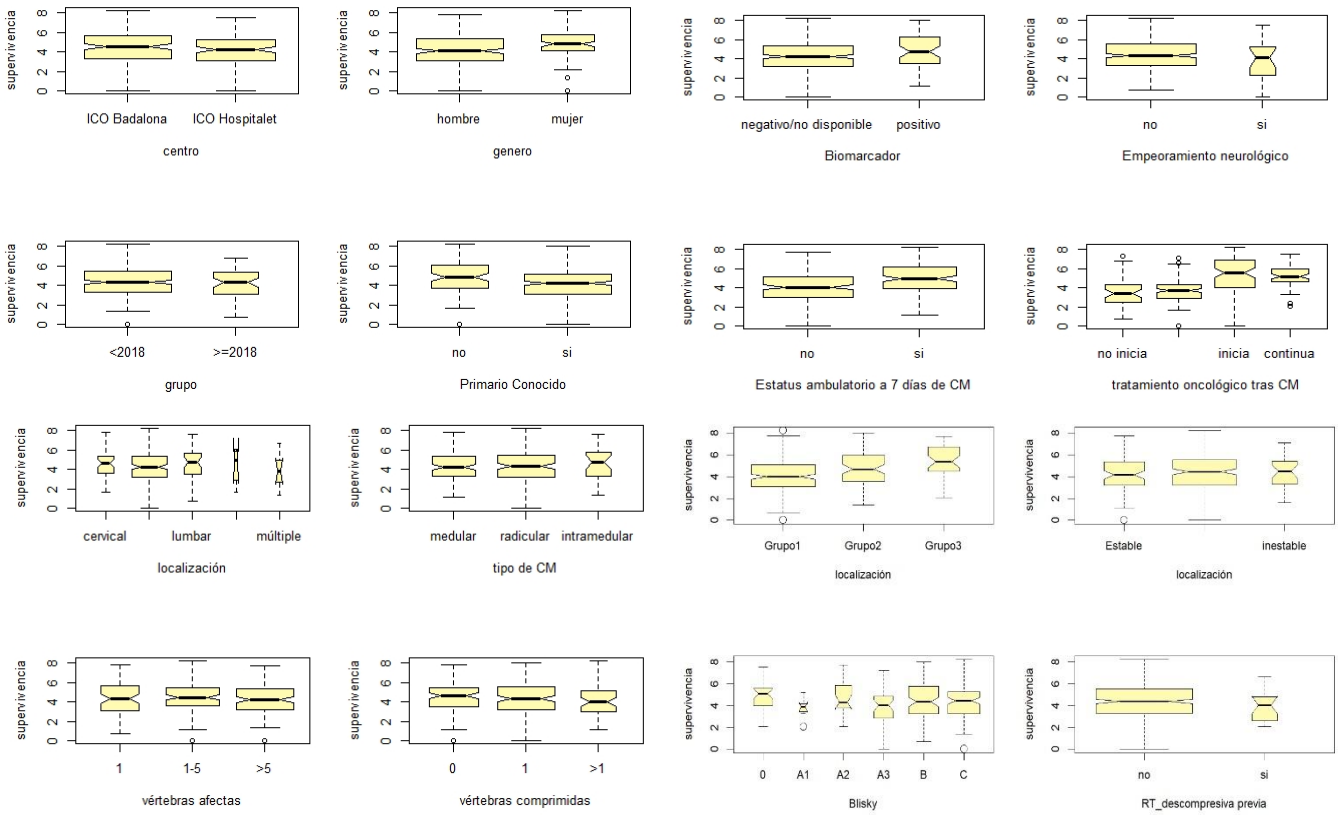


Figura 18 Cajas de dispersión de la supervivencia en función de las categorías de las variables categóricas

significativa entre ellas. Observamos diferencias que parecen significativas entre la supervivencia y el grupo RADES, parece que hay una ascensión clara con cada grupo, la supervivencia y la radioterapia descompresiva siendo superior si no la han recibido, la supervivencia y el tratamiento oncológico tras la CM viéndose muy favorecido los que inician y continúan frente a los que no lo hacen, y la supervivencia y el estado ambulatorio a los 7 días. Relaciones poco claras que requerirían mayor atención son las establecidas entre la supervivencia, el género y el tumor primario conocido.

3. Hablando de la supervivencia

3.1 Testeando la base de datos: supervivencia y RADES

El concepto de índice de Rades fue introducido (7) con la intención de estimar previamente al tratamiento la supervivencia del paciente. A través de un estudio multivariante demostró que distintas clasificaciones de grupos Rades correspondían a distintas supervivencias. Vamos a comprobar que ello se cumple para los pacientes de nuestra base de datos. Calculamos las curvas de supervivencia para cada grupo mediante el estimador de Kaplan-Meier.

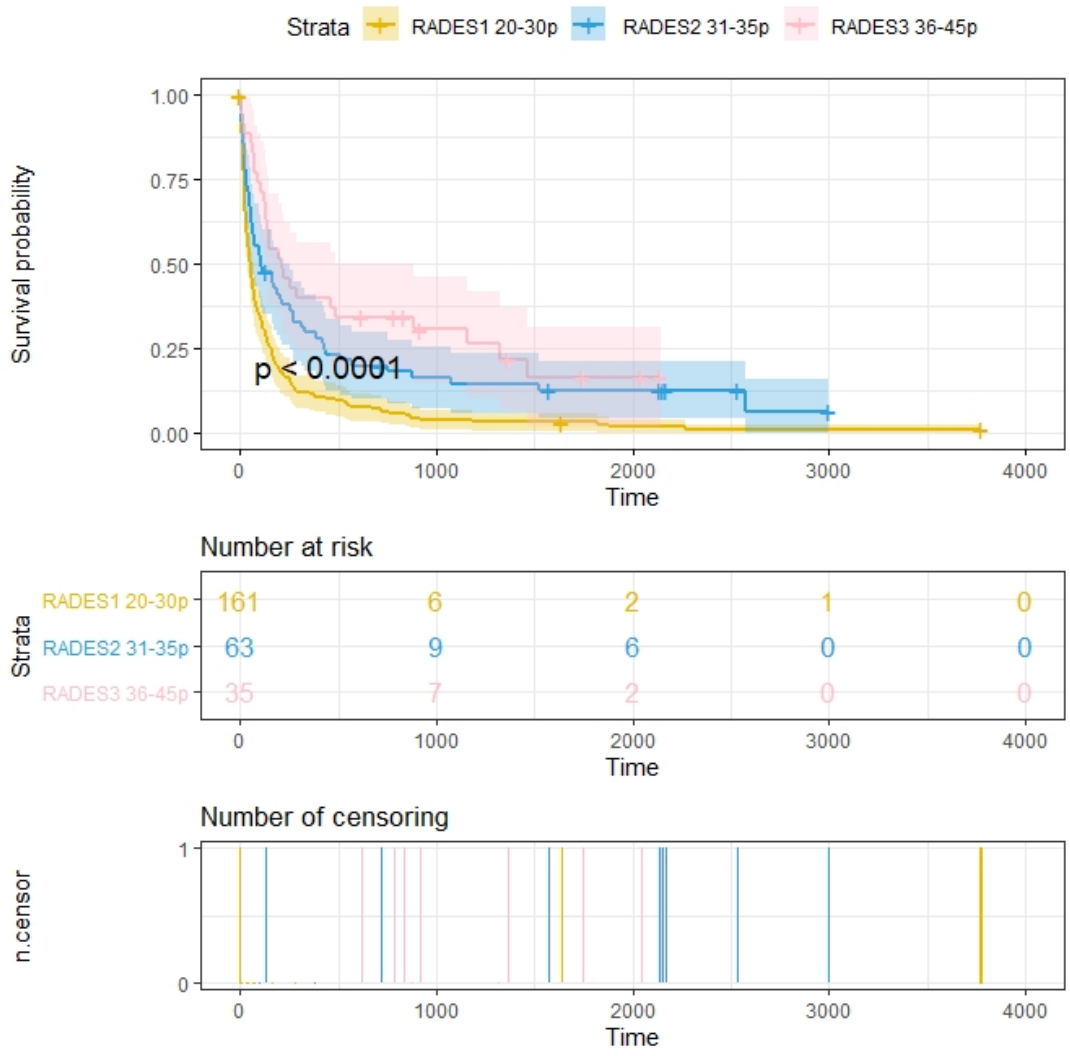


Figura 19 Supervivencia de la muestra en días para los distintos grupos Rades, número de pacientes en riesgo para cada intervalo de tiempo y número de pacientes censurados

Podemos observar una diferencia entre las distintas curvas de supervivencia, aunque los grados 2 y 3 superponen sus intervalos de confianza. Como ya vimos anteriormente, pacientes con un grado de Rades mayor son menos frecuentes. Ello explica que cuando el grado Rades aumenta observemos por un lado una extensión de los intervalos de confianza alrededor de la curva de supervivencia, y por otro lado observemos una desaparición de la cola de la curva de supervivencia. Dado que pacientes afectos de CM con longevidades muy largas son poco frecuentes, en muestras pequeñas es más raro que se den. Existen varios datos censurados para Rades3 a los 3años y varios para Rades2 un poco antes de los 6 años.

Las tasas de supervivencia de estos grupos son respectivamente 20±3 % , 43±6 % y 54±8 %. Si estimamos la esperanza de vida de estos grupos de pacientes (la mediana de la distribución) obtenemos:

	n	events	median	0.95LCL	0.95UCL
cleanCM_KM$RADES=0	161	158	53.5	40	69
cleanCM_KM$RADES=1	63	55	109.0	57	215
cleanCM_KM$RADES=2	35	27	214.0	129	881

Figura 19 Esperanza de vida para los distintos grupos Rades

Ello supone que el 50% de los pacientes estará muerto para Rades1 en menos de 2meses, para Rades2 en poco más de 3meses y para Rades3 pasados los 7 meses.

Realizamos un log-Rank test para estudiar la significancia de la diferencia en la supervivencia por grupos Rades. El interés del log-Rank test es que iguala la densidad de observaciones en todo el tiempo de seguimiento del paciente, aunque el comando que usamos en R, *survdiff*, permite ponderar más las primeras o las últimas observaciones, si se quiere. Obtenemos un resultado significativo con un p-valor de 0.000001 lo que nos lleva a rechazar la hipótesis nula de igualdad de las distribuciones. La diferencia de las curvas de supervivencia, pese a un cierto solapamiento de los intervalos de confianza, resulta ser significativa estadísticamente.

3.2 Análisis de supervivencia de toda la muestra

Visualizamos los primeros 24 días de la tabla de vida de toda nuestra muestra.
La primera columna es el instante de tiempo en que estamos valorando la situación, la segunda el número de pacientes vivos al inicio de ese intervalo, la tercera columna el número de pacientes que fallecen durante ese intervalo, la cuarta la supervivencia o estimador de Kaplan-Meier, la quinta la desviación estándar de esa probabilidad y la sexta y la séptima los límites del intervalo de confianza del 95%.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0	259	1	0.9961	0.00385	0.9886	1.0000
1	257	1	0.9923	0.00545	0.9816	1.0000
2	256	1	0.9884	0.00667	0.9753	1.0000
3	255	2	0.9806	0.00858	0.9638	0.9974
4	253	1	0.9768	0.00938	0.9584	0.9951
5	252	4	0.9613	0.01201	0.9377	0.9848
6	248	4	0.9458	0.01410	0.9181	0.9734
7	244	1	0.9419	0.01457	0.9133	0.9704
8	243	4	0.9264	0.01626	0.8945	0.9582
9	239	1	0.9225	0.01665	0.8899	0.9551
10	238	5	0.9031	0.01841	0.8670	0.9392
11	233	8	0.8721	0.02079	0.8314	0.9129
12	225	2	0.8644	0.02132	0.8226	0.9061
13	223	2	0.8566	0.02182	0.8138	0.8994
14	221	2	0.8488	0.02230	0.8051	0.8926
15	219	4	0.8333	0.02320	0.7879	0.8788
16	215	2	0.8256	0.02362	0.7793	0.8719
17	213	1	0.8217	0.02383	0.7750	0.8684
18	212	2	0.8140	0.02423	0.7665	0.8614
20	210	2	0.8062	0.02461	0.7580	0.8544
21	208	4	0.7907	0.02533	0.7411	0.8403
22	204	4	0.7752	0.02599	0.7243	0.8261
23	200	5	0.7558	0.02675	0.7034	0.8082
24	195	2	0.7481	0.02703	0.6951	0.8010

Tabla 9 Tabla de vida de nuestra muestra de datos

La esperanza de vida de toda la muestra es de 74 días (unos dos meses y medio) con unos intervalos de confianza al 95% de 57 días y 103 días. La tasa de supervivencia a los 6 meses de 30 ± 3 %.

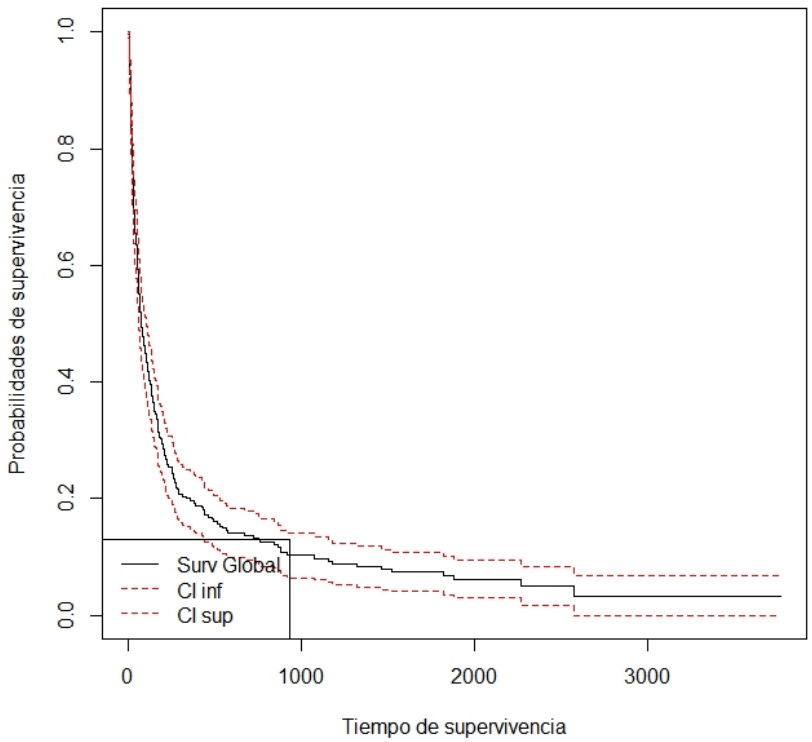


Figura 20 Curva de supervivencia con sus intervalos de confianza para toda la muestra

3.3 Antes y después del 2018

3.3.1 Homogeneidad en la distribución de variables predictoras

Para que el análisis de supervivencia por grupos según si la fecha de diagnóstico es inferior o posterior al 2018, sea robusto, necesitamos conocer si las distribuciones de las demás variables antes y después de ese año son iguales. Si no partimos de igualdad en las distribuciones de las variables, diferencias de la supervivencia entre ambos grupos podrían deberse al estudio de "muestras distintas" y no a una mejor/peor gestión clínica del paciente afecto de CM. Prácticamente todas las variables de interés son categóricas, tienen su homóloga categórica, como por ejemplo las variables *puntuacion_SINS* y *SINS,* o pueden ser tratadas como tales (las variables de tiempo se registran por intervalos diarios y tienen rangos intercuartiles como hemos visto antes muy cortos). Por lo que realizaremos un test chi cuadrado. El test chi cuadrado suele estar pensado para testear diferentes características de un mismo dataset. En nuestro caso queremos testear la misma característica de dos data sets distintos. Para poder operar con dos data sets de dimensiones distintas, tenemos que recurrir en R a alguna "artimaña" (apéndice pxxx).Vamos a comprobar la homogeneidad para los dos data sets de las distribuciones de las variables categóricas más representativas. Diremos que rechazamos la hipótesis de independencia entre ambos cuando p-valor sea menor a 0.05. Sólo la variable histología presenta un p-valor menor, de 0,03. Llama especialmente la atención la desaparición de casos de mieloma múltiple y se aprecia un ligero aumento en los casos de carcinoma de célula pequeña.

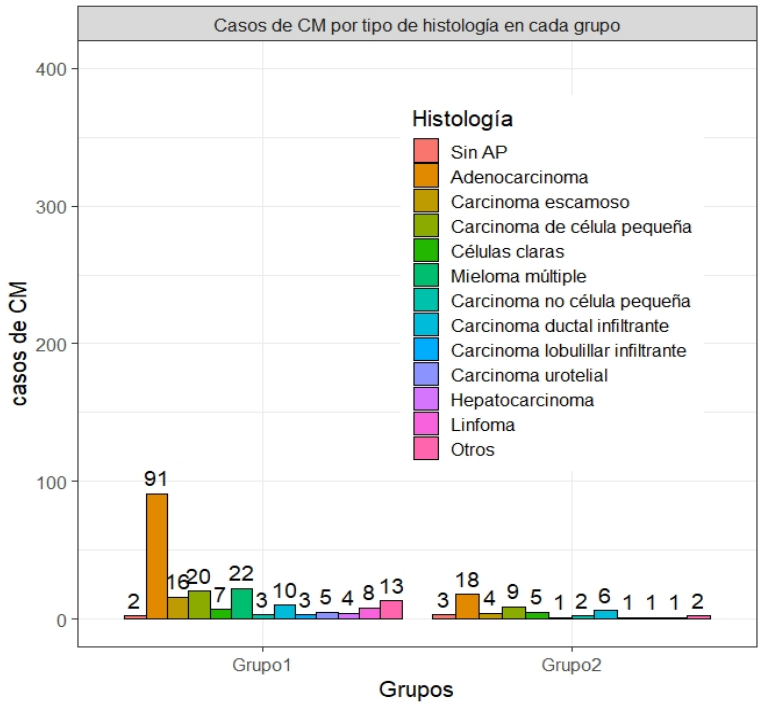


Figura 21 Distribución de las frecuencias de las distintas categorías de la variable histología para los individuos diagnosticados antes del 2018 o durante el 2018

3.3.2 Análisis de supervivencia

Las curvas de supervivencia de los pacientes tratados antes del 2018 y durante ese año y posterior, se solapan. Sus intervalos de confianza se funden completamente. Los pacientes más longevos que fueron diagnosticados a partir del 2018, no han tenido “tiempo de fallecer” a cierre del estudio clínico por lo que representan pérdidas censuradas y no podemos observar “la cola” de la curva de supervivencia. De hecho, podemos observar para el grupo2, en azul, numerosas pérdidas alrededor de los dos años (800 días) que coincide con el intervalo de tiempo desde que se reclutaron (2018) hasta el cierre del estudio (2020). Si estudiamos la cantidad de datos censurados observamos de hecho que la mayoría de datos caen en el grupo1, 7% del total, frente un 5% del grupo2. El problema estriba en que la gran mayoría de los datos censurados para el grupo2 caen en la "cola" de la supervivencia.

Observamos una vez más como la menor frecuencia de casos del grupo2 repercute en un ensanchamiento generoso del intervalo de confianza alrededor de la curva.

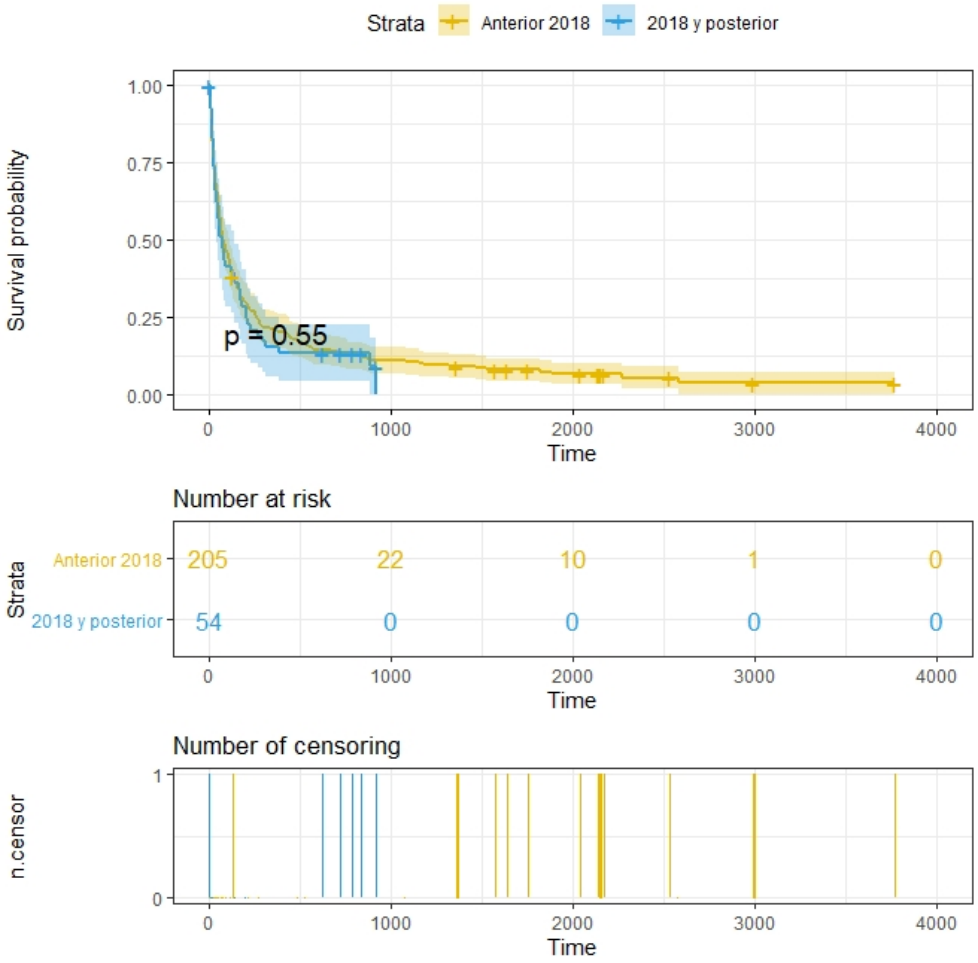


Figura 22 Curvas de supervivencia y datos censurados para los grupos 1 y 2

Al realizar un log-Rank test para ambos grupos obtenemos resultados no significativos (p-valor de 0.5, de forma acorde con la visualización gráfica. Para testear la diferencia podrían realizarse otros test probabilísticos como Wilcoxone, Tarone-Ware, Peto, Flemington-Harrington disponibles en el paquete *survMisc*, sin embargo, los resultados son tan concluyentes que debemos aceptar la igualdad de las distribuciones para ambos grupos.

Se obtienen esperanzas de vida muy similares de 75 días y 71 días con intervalos de confianza respectivamente al 97% de 59-109 y 41-140 respectivamente. Una vez más no podemos “distinguir” entre las esperanzas de vida de ambos grupos. Ni tampoco entre las tasas de supervivencia a 6 meses que se sitúan en 30±3% y 28±6% con desviaciones estándar que superponen sus medianas.

4. Variables influyentes en la supervivencia

Para el estudio de las variables de influencia se procederá a la regresión de Cox de riesgos proporcionales. La regresión de Cox nos da una expresión del riesgo de ocurrencia del evento (en este caso el fallecimiento) en un tiempo t proporcionado por unas variables predictoras x. Es un algoritmo clásico semiparamétrico para ajustar modelos tanto univariantes como multivariantes que tienen la supervivencia como variable de salida. Para el ajuste de los datos no se usarán métodos paramétricos ya que exigen conocer el tipo de distribución que siguen nuestros datos. Aunque en la siguiente sección se usaran algoritmos de machine learning para el modelado y predicción y se observaran también las variables de mayor uso, el "establecimiento" de qué variables son pronósticas para la supervivencia se realizará mediante la regresión de Cox. Los modelos de machine learning son muy útiles para predecir resultados, pero no establecen relaciones causales. En medicina, tenemos interés en actuar sobre la situación. Es decir, si el tiempo de sospecha clínica al diagnóstico está relacionado con la supervivencia, querremos reducirlo con el objetivo de mejorar la supervivencia. En ese sentido los modelos de machine learning podrían relacionar el número de zapatos vendidos en un año con el número de muertes de dicho año. Y, aunque resulta muy útil para predecir el número de muertes del próximo año, ello no significa que si el estado restringiera la venta de zapatos la gente dejara de morir.

Preparamos la base de datos para el estudio de regresión. A fin de evitar problemas de multicolinealidad en el modelado multivariante, eliminamos las variables que se hallan relacionadas entre ellas. En referencia a las variables numéricas, ya hemos visto en el capítulo de "Conociendo nuestra base de datos" que no existen correlaciones fuertes entre ellas, todos los coeficientes se hallan por debajo de 0.6, por lo que introduciremos todas. En referencia a aquellas variables que tengan representación numérica y a la vez categórica, escogeremos la categórica a menos que la variable numérica haya sido transformada y muestre un buen comportamiento. Eliminamos las variables redundantes sobre información del índice Rades, SINS y número de vértebras: *RADES_calc, vert_afect, vert_comp, puntuación RADES, puntuación SINS, SINS_calc, dolor_SINS, caract_lesion_SINS, alineacColum_SINS, location_SINS, afectPosterolat_SINS.* Las variables que permiten el cálculo de la puntuación RADES (*tumPrimario_RADES, metOseas_RADES, metViscRADES, intervalo_diagCM_RADES, estAmbulatorio_RADES, tiempoDesarrollo_DefMotor_RADES)* se dejaran para usarlas exclusivamente con el objetivo de comparar la regresión multivariante con la variable Rades y la regresión multivariante con las variables usadas para la determinación del índice Rades (y eliminando éste). Dichas variables no intervendrán en el resto del estudio. Además, dado que a mayor índice Rades mayor supervivencia del paciente, es de suponer que el oncólogo radioterápico prescriba fraccionamientos de dosis nominales superiores si cree que vivirá más (para así evitar posibles recidivas en un futuro). Realizamos un test chi cuadrado entre las variables *Rades* y *dosis_RT* que resulta ser significativo por lo que rechazamos la

hipótesis nula de independencia de ambas y decidimos eliminar la variable *dosis_RT* de nuestra base de datos.

```
        Pearson's Chi-squared test

data:  table(cleanCM$RADES, cleanCM$dosis_RT)
X-squared = 44.154, df = 8, p-value = 0.000000532
```

Tabla 10 Test chi cuadrado entre la variable RADES y dosis_RT

Tenemos una base de datos con 257 observaciones y 35 variables, 5 de las cuales no se usarán como hemos explicitado para el modelado general. Las variables categóricas con sólo dos categorías o incluso tres categorías pero que tengan un orden, se consideraran numéricas a efectos de realizar las regresiones por simplicidad “estética” y orden. Sólo aquellas variables categóricas con multiplicidad de categorías sin un orden aparente entre ellas, se definirán como factor a efectos de la realización de regresiones.

4.1 Regresiones univariantes de Cox

En la tala visualizamos los resultados para la regresión univariante de Cox en relación a la histología. La regresión, al ser un factor, se desdobla para cada una de sus categorías en relación a la primera de ellas (T.0 “sin anatomía patológica”).

En la tabla visualizamos el Hazard Ratio (HR). Como sabemos el HR corresponde a la exponencial del coeficiente de regresión y describe la ratio entre las tasas instantáneas de ocurrencia del evento (muerte) de los dos grupos considerados. HR inferiores a 1 se consideran "riesgos reducidos" y factores de buen pronóstico, mientras que HR mayores de 1 se consideran riesgos altos de muerte y factores de mal pronóstico. Así por ejemplo vemos que para cada 0.27 casos de mieloma múltiple (T.5) en riesgo hay uno de sin anatomía patológica (AP) (T.0), que por cada 0.24 casos linfoma en riesgo hay uno sin AP o que por cada 0.39 casos de carcinoma urotelial (T.11) en riesgo hay 1 sin AP (T.0). Estas son las tres histologías con un p-valor menor a 0.05 y que resultan ser significativas.
Además de la tabla podemos visualizar para la bondad del ajuste, la concordancia del modelo y los test probabilísticos de likelihood-ratio, de Wald y el logrank test. La concordancia del modelo sería el homólogo en la regresión de Cox al coeficiente de correlación en una regresión lineal. La concordancia del modelo pese a no ser buenísima (0.63) tampoco es mala y los 3 test

Characteristic	HR[1]	95% CI[1]	p-value
histologia			
histologia[T.1]	1.09	0.44, 2.67	0.9
histologia[T.2]	1.64	0.61, 4.37	0.3
histologia[T.3]	1.32	0.51, 3.41	0.6
histologia[T.4]	0.77	0.27, 2.19	0.6
histologia[T.5]	0.27	0.10, 0.76	0.013
histologia[T.6]	1.54	0.45, 5.35	0.5
histologia[T.7]	0.39	0.13, 1.13	0.082
histologia[T.8]	0.46	0.12, 1.73	0.3
histologia[T.9]	2.68	0.77, 9.29	0.12
histologia[T.10]	1.41	0.38, 5.28	0.6
histologia[T.11]	0.24	0.07, 0.81	0.022
histologia[T.12]	0.94	0.34, 2.62	>0.9

[1] HR = Hazard Ratio, CI = Confidence Interval

Tabla11 HR, ICC y p-valor para la regresión univariante con la histología

mencionados son significativos de largo. Al realizar las regresiones univariantes para cada variable obtenemos:

	beta	HR (95% CI for HR)	wald.test	p.value
edad	0.018	1 (1-1)	9.7	0.0018
genero	-0.36	0.7 (0.52-0.93)	6	0.014
centro	0.26	1.3 (1-1.7)	3.9	0.048
grupo	0.1	1.1 (0.81-1.5)	0.4	0.53
t_ac	-0.13	0.88 (0.79-0.99)	4.8	0.028
t_sc	0.15	1.2 (1-1.3)	4.7	0.03
t_itto	-0.098	0.91 (0.83-1)	4.2	0.04
localizacion	-0.03	0.97 (0.81-1.2)	0.11	0.74
tipo_CM	-0.011	0.99 (0.8-1.2)	0.01	0.91
n_vert_afect	0.034	1 (0.93-1.1)	0.39	0.53
n_vert_comp	0.077	1.1 (0.87-1.3)	0.49	0.48
CM_PrimarioConocido	0.44	1.6 (1.2-2.1)	9.3	0.0023
valor_biomarcador	-0.36	0.7 (0.49-0.99)	4	0.045
RADES	-0.49	0.61 (0.5-0.74)	26	0.00000038
SINS	-0.026	0.97 (0.77-1.2)	0.05	0.83
Blisky	0.01	1 (0.94-1.1)	0.07	0.79
clinica_debut	0.059	1.1 (1-1.1)	3.5	0.062
RT_descomprPrevia	0.51	1.7 (1-2.7)	4.2	0.039
empeoramiento_neurologico	0.14	1.1 (0.73-1.8)	0.35	0.55
estatusAmbulatorio7	-0.61	0.54 (0.42-0.71)	20	0.0000084
fijacionPrevia_SegmIrradiar	-0.88	0.42 (0.2-0.84)	5.9	0.015
test_diagnostico	0.077	1.1 (0.79-1.5)	0.24	0.62
opcion_terapeutica	0.46	1.6 (1.2-2.1)	9.1	0.0026
tto_trasCM	-0.44	0.64 (0.56-0.73)	45	2.3e-11

Figura 22 Resultados de las regresiones univariantes para la regresión de Cox. Se muestra en la primera columna el coeficiente de regresión, en la segunda el HR, en la tercera los intervalos de confianza para éste, en la cuarta el valor del test de Wald y el p-valor asociado a éste

Son significativas con un p-valor inferior a 0.05 las regresiones en relación a la edad, el género, las variables temporales (*t_ac, t_sc, t_itto)*, el tumor primario conocido, el índice Rades, la radioterapia descompresiva previa, el estatus ambulatorio a los 7 días de finalizar el tratamiento, la fijación previa del segmento a irradiar , la opción terapéutica y el tratamiento tras la CM, escasamente también lo son el centro de tratamiento y el valor del biomarcador. Las variables más significativas son por orden el tratamiento oncológico tras la CM, el índice Rades y el estado ambulatorio a 7 días. De ellas, son factores pronóstico ser joven, mujer, tratarse en el Hospital Germans Tries i Pujol, tiempos entre la sospecha clínica y el diagnóstico cortos, tiempos largos entre el diagnóstico y el aviso de compresión / inicio de tratamiento, no conocer el tumor primario, tener un valor de biomarcador positivo, grupos altos de Rades, no haber realizado radioterapia descompresiva previa, estado ambulatorio a los 7 días, la fijación previa del segmento a irradiar, opciones terapéuticas que incluyan cirugía y continuar/iniciar tratamiento oncológico tras el tratamiento por CM.

Para testear la hipótesis de proporcionalidad de riegos usaremos el método analítico de los residuos de Schoenfeld. La función *cox.zph*, usa los residuos de Schoenfeld basándose en el test de Kolmogorov. Sólo las variables de género y de tratamiento tras CM son significativas y no cumplen la hipótesis de riesgos. Como dichas variables no dependen del tiempo, debemos suponer que sus coeficientes en el modelado de Cox son dependientes del tiempo. Una primera aproximación válida

según (39) sería suponer que sus coeficientes siguen una función escalón en el tiempo tomando constantes distintas según el intervalo de tiempo considerado. Así pues, podemos usar la función *survSplit* para ampliar nuestra base de datos de manera que, en función de unos cortes de elección en el tiempo, se creen para cada sujeto tantas entradas como momentos en el tiempo de valoración de su estado (muerto/vivo). Por ejemplo, supongamos que definimos 2 cortes de tiempo y por tanto 3 intervalos de tiempo. Para un paciente x se crearán 3 entradas iguales cada la evaluación del sujeto en cada intervalo. Dichas entradas se diferenciarán en la variable *tgroup* que tomará los valores de 1, 2, 3 según el intervalo de tiempo que se considere. Además, la variable que recoge el estado del paciente se actualizará en cada uno de ellos para recoger si en ese intervalo el paciente está vivo o muerto. Si el paciente muriera en el intervalo 2 de tiempo y no llegara al 3, sólo tendría dos entradas en la nueva base de datos "spliteada".

Tras una rápida búsqueda iterativa manual para "encontrar" el momento de tiempo en que el comportamiento de los coeficientes varía, escogemos dos cortes de tiempo al mes y a los 3 meses. De manera que consideramos los intervalos: 0-1mes, 1-3meses y más de 3 meses. Realizamos de nueva la regresión de Cox sobre la base de datos "spliteada" y obtenemos por ejemplo para el tratamiento tras la CM un modelado con una concordancia aceptable (0.65), un p-valor de los test significativa que no es significativo ante la hipótesis de riesgos proporcionales. Este método tiene el interés que nos permite ahondar sobre el comportamiento de la variable en relación a la regresión. Parece ser que el hecho que inicie o continúe con tratamiento oncológico tras finalizar el tratamiento por CM, es un factor pronóstico para la supervivencia hasta los primeros 3 meses, especialmente el primero, pero no más allá.

```
Call:
coxph(formula = Surv(tstart, survival, estatus_UC) ~ tto_trasCM:strata(tgroup),
    data = CMcoxStep, x = TRUE, y = TRUE)

  n= 554, number of events= 238

                                      coef exp(coef) se(coef)      z      Pr(>|z|)
tto_trasCM:strata(tgroup)tgroup=1 -0.7380    0.4781   0.1229 -6.008 0.00000000188 ***
tto_trasCM:strata(tgroup)tgroup=2 -0.5270    0.5904   0.1250 -4.214 0.00002505546 ***
tto_trasCM:strata(tgroup)tgroup=3 -0.0919    0.9122   0.1216 -0.756          0.45
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                  exp(coef) exp(-coef) lower .95 upper .95
tto_trasCM:strata(tgroup)tgroup=1    0.4781      2.092    0.3758    0.6082
tto_trasCM:strata(tgroup)tgroup=2    0.5904      1.694    0.4621    0.7544
tto_trasCM:strata(tgroup)tgroup=3    0.9122      1.096    0.7187    1.1577

Concordance= 0.652  (se = 0.019 )
Likelihood ratio test= 59.27  on 3 df,   p=8e-13
Wald test            = 54.42  on 3 df,   p=9e-12
Score (logrank) test = 59.59  on 3 df,   p=7e-13

                        chisq df    p
 tto_trasCM:strata(tgroup)  2.44  3 0.49
 GLOBAL                     2.44  3 0.49
```

Figura 23 Modelado de la regresión de Cox univariante para el tratamiento oncológico tras la CM con base de datos "spliteada". Test de los residuos de Shoenfeld.

Para la variable de género el modelado tiene baja concordancia, de 0.5, y sólo es factor de buen pronóstico ser mujer durante el primer mes tras finalizar el tratamiento por CM.

4.2 Regresiones multivariantes de Cox

Para el modelado de la regresión multivariante procedemos a incluir todas las variables y eliminamos de forma iterativa variables buscando aumentar la concordancia del modelo, minimizar el p-valor y maximizar la significancia estadística de las variables. Eliminamos primeramente todos aquellas variables con un p-valor mayor a 0.1. Posteriormente, procedemos a eliminar una a una aquellas variables con una significancia mayor a 0.05 y reevaluamos la concordancia del modelo, el p-valor obtenido y los cambios en la significancia de cada una de las variables. El modelo final contiene las siguientes variables. La presencia de las variables *género* y *CM_PrimarioConocido* mejoraban la concordancia del modelo o el p-valor de otras variables.

```
Call:
coxph(formula = Surv(survival, estatus_UC) ~ genero + t_ac +
    t_sc + CM_PrimarioConocido + RADES + RT_descomprPrevia +
    tto_trasCM + estatusAmbulatorio7 + histologia + opcion_terapeutica,
    data = CMcox, x = TRUE, y = TRUE)

  n= 257, number of events= 239

                        coef exp(coef) se(coef)      z     Pr(>|z|)
genero              -0.28645   0.75092  0.18053 -1.587     0.112581
t_ac                -0.12348   0.88384  0.06386 -1.934     0.053170 .
t_sc                 0.22718   1.25506  0.07889  2.880     0.003981 **
CM_PrimarioConocido  0.25884   1.29543  0.16465  1.572     0.115937
RADES               -0.40386   0.66773  0.11032 -3.661     0.000251 ***
RT_descomprPrevia    0.70510   2.02404  0.28227  2.498     0.012492 *
tto_trasCM          -0.40941   0.66404  0.07639 -5.359 0.0000000836 ***
estatusAmbulatorio7 -0.30067   0.74032  0.14870 -2.022     0.043182 *
histologia[T.1]     -0.25784   0.77272  0.47599 -0.542     0.588022
histologia[T.2]      0.04933   1.05057  0.52296  0.094     0.924843
histologia[T.3]      0.05347   1.05492  0.50543  0.106     0.915749
histologia[T.4]     -0.32868   0.71987  0.54823 -0.600     0.548820
histologia[T.5]     -1.11605   0.32757  0.53819 -2.074     0.038105 *
histologia[T.6]      0.22141   1.24783  0.66398  0.333     0.738787
histologia[T.7]     -0.64257   0.52594  0.56876 -1.130     0.258567
histologia[T.8]     -0.99499   0.36973  0.71599 -1.390     0.164630
histologia[T.9]      0.69814   2.01001  0.64660  1.080     0.280271
histologia[T.10]    -0.47461   0.62213  0.70189 -0.676     0.498920
histologia[T.11]    -1.39667   0.24742  0.62767 -2.225     0.026069 *
histologia[T.12]    -0.30102   0.74007  0.54238 -0.555     0.578900
opcion_terapeutica   0.21783   1.24337  0.17379  1.253     0.210067
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance= 0.741  (se = 0.017 )
Likelihood ratio test= 148.6  on 21 df,   p=<2e-16
Wald test            = 129.6  on 21 df,   p=<2e-16
Score (logrank) test = 143.7  on 21 df,   p=<2e-16
```

Figura 24 Modelado de la regresión de Cox multivariante

Nuestro modelo final contiene como covariables: el género, el tiempo des del diagnóstico al aviso de compresión, el tiempo des de la sospecha clínica al diagnóstico, el tiempo des del diagnóstico al inicio del tratamiento, el tumor primario conocido, RADES, la RT descompresiva previa, el estado ambulatorio a los 7 días y el tratamiento tras la CM. Todas las variables son factores de buen pronóstico,

exceptuando el tiempo des de la sospecha clínica al diagnóstico, el tiempo des del diagnóstico al inicio de tratamiento, el conocimiento del tumor primario y la existencia de RT descompresiva previa que son factores de mal pronóstico. Tenemos varios nivele de significancia. Primeramente, RADES y tratamiento tras la CM, en segundo lugar, el tumor primario conocido, en tercer lugar, el género, el tiempo de la sospecha al diagnóstico y la RT descompresiva previa, finalmente el tiempo des del diagnóstico al aviso de CM. El modelado tiene una concordancia significativa del 0.741 y un p-valor muy pequeño en diferentes test (likelihood ratio, logrank y wald test), indicando una buena significación del modelo en general. Sin embargo, como ya podemos presuponer, por nuestra experiencia en el modelado univariante, el modelado no cumple la hipótesis de riesgos proporcionales al analizarlo analíticamente. Las variables de *tto_trasCM y estatusAmbulatorio7* muestran valores significativos inferiores a 0.05 que nos llevan de manera global a rechazar la hipótesis de riesgos proporcionales.

```
> cox.zph(coxmultiF)
                       chisq df       p
genero                3.7030  1 0.05432
t_ac                  0.0490  1 0.82477
t_sc                  0.0206  1 0.88588
CM_PrimarioConocido   0.6875  1 0.40703
RADES                 0.5997  1 0.43870
RT_descomprPrevia     0.2835  1 0.59440
tto_trasCM           14.9898  1 0.00011
estatusAmbulatorio7   4.3520  1 0.03697
histologia           17.1049 12 0.14570
opcion_terapeutica    0.0533  1 0.81734
GLOBAL               40.0974 21 0.00724
```

Figura 25 Test de los residuos de Shoenfeld para la regresión de Cox multivariante

Como ninguna de las 2 variables muestra dependencia en el tiempo suponemos que debe ser el coeficiente el que tiene una dependencia temporal. Para investigar el comportamiento de nuestras variables y crear un modelado de Cox que cumpla sus supuestos procederemos como hemos explicado en el punto anterior con la suposición de una función escalón en el tiempo para los coeficientes de las variables que no cumplen la hipótesis de riesgos proporcionales y un “spliteado” de la base de datos. Mediante este acercamiento (figura 26) la concordancia (0.743) del modelo no disminuye. Se mantienen como covariables significativas: el índice Rades, la RT descompresiva previa, el tiempo de aviso de compresión, el tiempo de sospecha clínica, el tumor primario conocido y la histología para tumores tipo mieloma/linfoma. El modelo sigue siendo significativo para los test de Wald, logrank test y likelihood ratio con un p-valor inferior. Y sobre todo nos permite saber lo que está sucediendo. En los primeros 3 meses (que corresponde aproximadamente a la esperanza de vida), el hecho de iniciar o continuar un tratamiento oncológico es un factor pronóstico para la supervivencia con significancia muy alta. El estado ambulatorio a los 7 días correlaciona sólo con la supervivencia el primer mes. Sin embargo, a partir de los 3 primeros meses, donde se hallan las supervivencias que superan la esperanza de vida, ninguno de esos factores es significativo. Esta forma de proceder supone una “cierta estratificación” en que se presupone que los pacientes de cada “strata”, en este caso de cada intervalo de tiempo, son más similares entre ellos que a otros pacientes escogidos al azar de otros intervalos de tiempo. Sin embargo, las variables se incluyen en el modelo de Cox y no son tratadas independientemente de forma aditiva a él.

```
Call:
coxph(formula = Surv(tstart, survival, estatus_UC) ~ genero +
    t_ac + t_sc + CM_PrimarioConocido + RADES + RT_descomprPrevia +
    tto_trasCM:strata(tgroup) + estatusAmbulatorio7:strata(tgroup) +
    histologia + opcion_terapeutica, data = CMcoxStep, x = TRUE,
    y = TRUE)

  n= 554, number of events= 238

                                           coef exp(coef) se(coef)      z   Pr(>|z|)
genero                                  -0.26870   0.76437  0.18004 -1.492   0.135589
t_ac                                    -0.13141   0.87686  0.06407 -2.051   0.040264 *
t_sc                                     0.23600   1.26617  0.07939  2.973   0.002954 **
CM_PrimarioConocido                      0.24197   1.27376  0.16421  1.474   0.140608
RADES                                   -0.39867   0.67121  0.11060 -3.605   0.000313 ***
RT_descomprPrevia                        0.78954   2.20238  0.28018  2.818   0.004833 **
histologia[T.1]                         -0.13071   0.87748  0.47567 -0.275   0.783483
histologia[T.2]                          0.22318   1.25005  0.52426  0.426   0.670322
histologia[T.3]                          0.22470   1.25195  0.50650  0.444   0.657299
histologia[T.4]                         -0.16452   0.84830  0.54974 -0.299   0.764736
histologia[T.5]                         -1.06941   0.34321  0.53962 -1.982   0.047504 *
histologia[T.6]                          0.08777   1.09174  0.69937  0.125   0.900128
histologia[T.7]                         -0.55683   0.57302  0.57301 -0.972   0.331170
histologia[T.8]                         -0.89505   0.40859  0.71727 -1.248   0.212085
histologia[T.9]                          0.87196   2.39160  0.64820  1.345   0.178562
histologia[T.10]                        -0.34713   0.70671  0.70173 -0.495   0.620826
histologia[T.11]                        -1.13446   0.32160  0.63049 -1.799   0.071966 .
histologia[T.12]                        -0.20779   0.81238  0.54157 -0.384   0.701220
opcion_terapeutica                       0.22880   1.25709  0.17482  1.309   0.190598
tto_trasCM:strata(tgroup)tgroup=1       -0.68393   0.50463  0.13365 -5.117 0.00000031 ***
tto_trasCM:strata(tgroup)tgroup=2       -0.52673   0.59054  0.13393 -3.933 0.00008393 ***
tto_trasCM:strata(tgroup)tgroup=3       -0.07215   0.93040  0.13137 -0.549   0.582875
strata(tgroup)tgroup=1:estatusAmbulatorio7 -0.77704   0.45977  0.28570 -2.720   0.006533 **
strata(tgroup)tgroup=2:estatusAmbulatorio7 -0.09461   0.90972  0.27128 -0.349   0.727259
strata(tgroup)tgroup=3:estatusAmbulatorio7 -0.17816   0.83681  0.22852 -0.780   0.435626


Concordance= 0.743  (se = 0.017 )
Likelihood ratio test= 165.5  on 25 df,   p=<2e-16
Wald test            = 133.6  on 25 df,   p=<2e-16
Score (logrank) test = 152.7  on 25 df,   p=<2e-16
```

Figura 26 Regresión de Cox multivariante en base de datos "spliteada" en que los coeficientes de las variables tto_trasCM y estatusAmbulatorio7 dependen del tiempo

Imaginemos si trabajáramos con un modelo de Cox estratificado realmente en el que las variables que no cumplen la hipótesis de proporcionalidad de riesgos se excluyen

Existen clásicamente dos formas de manejar variables que no cumplen la hipótesis de proporcionalidad de riesgos. Una de ellas es la introducción de las covariables como interacciones dependientes del tiempo al modelo básico de Cox, veremos un ejemplo así en el próximo apartado. La otra, es pasar a trabajar con un modelo de Cox estratificado en la que esas variables no se introducen en el modelo, de manera que éste no viola la proporcionalidad de riesgos, y son tratadas independientemente como un aditivo al modelo.

```
Call:
coxph(formula = Surv(survival, estatus_UC) ~ genero + t_ac +
    t_sc + CM_PrimarioConocido + RADES + RT_descomprPrevia +
    strata(tto_trasCM) + strata(estatusAmbulatorio7) + histologia +
    opcion_terapeutica, data = CMcox)

  n= 257, number of events= 239

                       coef exp(coef) se(coef)      z Pr(>|z|)
genero              -0.28906   0.74896  0.18538 -1.559  0.11893
t_ac                -0.05123   0.95006  0.06824 -0.751  0.45277
t_sc                 0.23502   1.26493  0.08496  2.766  0.00567 **
CM_PrimarioConocido  0.06043   1.06229  0.20414  0.296  0.76723
RADES               -0.28560   0.75156  0.12012 -2.378  0.01743 *
RT_descomprPrevia    0.62557   1.86931  0.29963  2.088  0.03682 *
histologia[T.1]      0.12819   1.13677  0.49110  0.261  0.79407
histologia[T.2]      0.54771   1.72929  0.54552  1.004  0.31537
histologia[T.3]      0.44245   1.55652  0.52332  0.845  0.39785
histologia[T.4]      0.04367   1.04463  0.57801  0.076  0.93978
histologia[T.5]     -0.84040   0.43154  0.56266 -1.494  0.13528
histologia[T.6]      0.40776   1.50345  0.68179  0.598  0.54979
histologia[T.7]     -0.22464   0.79881  0.58870 -0.382  0.70277
histologia[T.8]     -0.79084   0.45347  0.74630 -1.060  0.28929
histologia[T.9]      1.06801   2.90957  0.67656  1.579  0.11443
histologia[T.10]    -0.32656   0.72140  0.75766 -0.431  0.66646
histologia[T.11]    -0.86957   0.41913  0.67386 -1.290  0.19690
histologia[T.12]     0.11222   1.11876  0.56296  0.199  0.84200
opcion_terapeutica   0.26928   1.30902  0.18159  1.483  0.13809
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance= 0.647  (se = 0.024 )
Likelihood ratio test= 59.79  on 19 df,   p=0.000004
Wald test            = 51.73  on 19 df,   p=0.00007
Score (logrank) test = 55.53  on 19 df,   p=0.00002
```

Figura 28 Modelo de Cox estratificado

Efectivamente el estado ambulatorio a 7 días y el tratamiento tras la CM no son añadidos al modelo. Podemos ver que el modelo pierde significancia en el test estadístico, la concordancia pasa de aceptable (0.742) a mediocre (0.647) y las variables muestran p-valores menos significativos. Además, no podemos acceder a ningún conocimiento sobre cómo dependen del tiempo las variables “estratificadas”.

4.3 Diagnósticos para el modelo final de Cox

Para describir cómo de bien nuestro modelo de Cox describe los datos debemos asegurar no sólo que cumpla la hipótesis de proporcionalidad de riesgos sino comprobar las demás presunciones del modelo. Deberemos estudiar la existencia de puntos influyentes y extremales, así como la linealidad del logaritmo del riesgo y sus covariables.

En referencia a la hipótesis de proporcionalidad de riesgos,
realizamos el test analítico y además representamos gráficamente para cada covariable los residuos de Shoenfeld escalados respecto a la transformada del tiempo. En principio, y siempre que se cumple la hipótesis de proporcionalidad de riesgos, los residuos de Schoenfeld son independientes del tiempo. Por tanto, todo comportamiento con patrones no aleatorios supone una violación de esta hipótesis.

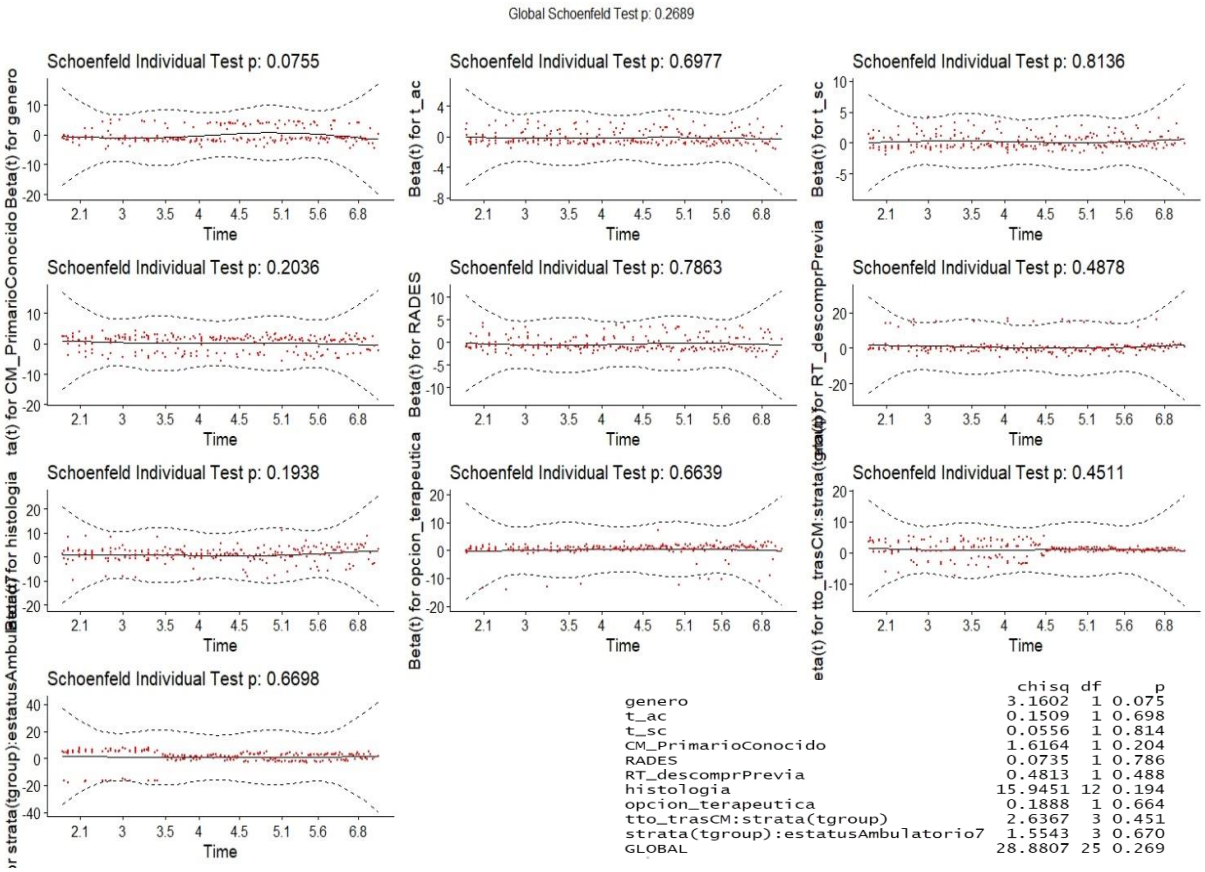


Figura 29 A la derecha, test analítico basado en los residuos de Schoenfeld con resultados no significativos. Arriba, representación gráfica de los residuos de Shoenfeld escalados respecto a la transformada en el tiempo. En el gráfico la línea sólida corresponde a una interpolación por esplines ajustada a la nube de puntos con una banda de más/menos 2 de desviación estándar alrededor de la recta.

Efectivamente los puntos parecen seguir una distribución de patrón aleatorio.

Una forma de testear gráficamente la existencia de puntos influyentes y extremales es mediante la representación de los residuos desviados respecto al tiempo. Los residuos desviados corresponden a la transformada normalizada de los residuos de Martingale. Los residuos desviados deben grosso modo distribuirse alrededor del eje de manera simétrica con una desviación estándar de 1. Valores positivos corresponden a individuos que murieron demasiado pronto, mientras que residuos negativos corresponden a individuos que vivieron “demasiado”. Puntos muy distantes no están bien definidos por el modelo.

.

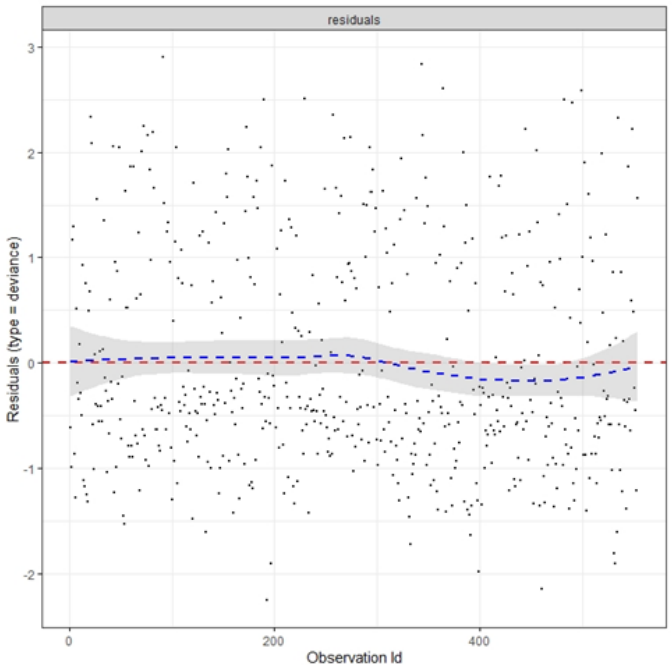


Figura 30 Residuos desviados. En la parte positiva del eje, correspondiente a individuos que "mueren demasiado pronto", los puntos se hallan más espaciados que en la parte negativa del eje. De la misma forma en ese lado observamos más observaciones entre 2 y 3 desviaciones estándar. Ello podría suponer la presencia de outliers de supervivencias cortas

El estudio de la linealidad del logaritmo del riesgo y sus covariables se lleva a cabo mediante los residuos de Martingale. Esta presunción del modelo es frecuentemente omitida. Los residuos de Martingale tienen un rango entre menos infinito y 1. Valores de los residuos cercanos a 1 representan individuos que fallecieron demasiado pronto mientras que valores negativos altos corresponden a individuos que "vivieron demasiado". La representación gráfica de las covariables respecto a los residuos de Martingale debe ser lineal. Este estudio sólo procede para las covariables numéricas y continuas.

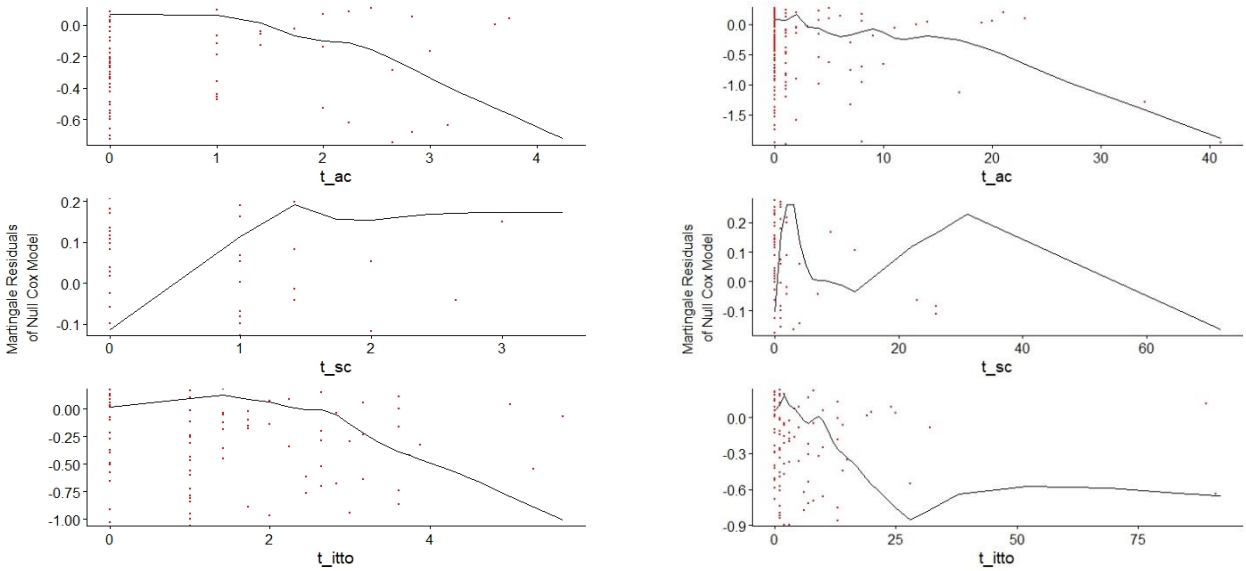


Figura 31 A la izquierda comportamiento de las variables temporales transformadas respecto a los residuos de Martingale. A la derecha comportamiento de las variables temporales originales respecto a los residuos de Martingale.

Aunque la forma de las covariables no acaba de ser del todo lineal, es bastante aceptable. Es muchísimo más lineal que si no hubiéramos realizado una transformación sobre las variables. Además, cabe considerar, como observamos

los puntos en la gráfica, que esas variables no son del todo continuas. Sus datos son recogidos en intervalos de 1día y su rango es relativamente corto.

En general podemos concluir que las presunciones del modelo de Cox se cumplen de manera suficiente para que el modelado tenga una cierta entidad a nivel estadístico.

4.4 Regresión de Cox con las variables para el cálculo del índice Rades

Encontrado un buen ajuste de Cox para nuestro modelado de datos y comprobadas las presunciones del modelo, comprobaremos qué sucede al extraer de nuestro modelo la covariable de Rades e introducir en su lugar las variables a partir de las cuales se calcula este índice.

	coef	exp(coef)	Pr(>|z|)
genero	-0.359	0.698	0.055
t_ac	-0.027	0.973	0.691
t_sc	0.174	1.190	0.037
CM_PrimarioConocido	0.295	1.343	0.109
RT_descomprPrevia	0.802	2.230	0.005
histologia[T.1]	-0.279	0.757	0.564
histologia[T.2]	0.264	1.302	0.627
histologia[T.3]	-0.169	0.845	0.754
histologia[T.4]	-0.227	0.797	0.694
histologia[T.5]	-1.608	0.200	0.015
histologia[T.6]	-0.027	0.973	0.970
histologia[T.7]	-0.865	0.421	0.373
histologia[T.8]	-1.463	0.232	0.209
histologia[T.9]	0.549	1.731	0.408
histologia[T.10]	-0.537	0.584	0.458
histologia[T.11]	-1.691	0.184	0.031
histologia[T.12]	-0.365	0.694	0.510
opcion_terapeutica	0.127	1.135	0.493
tiempoDesarrollo_DefMotor_RADES[T.1]	-0.319	0.727	0.341
tiempoDesarrollo_DefMotor_RADES[T.2]	0.030	1.031	0.880
tiempoDesarrollo_DefMotor_RADES[T.3]	-0.197	0.821	0.276
estAmbulatorio_RADES	-0.422	0.656	0.007
tumPrimario_RADES[T.1]	-0.337	0.714	0.712
tumPrimario_RADES[T.2]	0.299	1.349	0.720
tumPrimario_RADES[T.3]	0.163	1.177	0.857
tumPrimario_RADES[T.4]	0.019	1.019	0.983
metOseas_RADES	0.287	1.333	0.143
metVisc_RADES	0.626	1.871	0.000
intervalo_diagCM_RADES	0.026	1.026	0.885
tto_trasCM:strata(tgroup)tgroup=1	-0.643	0.526	0.000
tto_trasCM:strata(tgroup)tgroup=2	-0.519	0.595	0.000
tto_trasCM:strata(tgroup)tgroup=3	-0.133	0.875	0.315
strata(tgroup)tgroup=1:estatusAmbulatorio7	-0.717	0.488	0.016
strata(tgroup)tgroup=2:estatusAmbulatorio7	-0.069	0.934	0.806
strata(tgroup)tgroup=3:estatusAmbulatorio7	-0.219	0.804	0.351

```
> summary(coxStepPron)$concordance
         C      se(C)
0.75500809 0.01597616
> summary(coxStepPron)$waldtest
       test          df       pvalue
1.532800e+02 3.500000e+01 9.559486e-17
```

Figura 31 A la izquierda modelo de regresión multivariante de Cox considerando las variables propias del cálculo del índice de Rades. A bajo concordancia del modelo significancia del estadística

Como podemos observar la concordancia del modelo aumenta de 0.743 a 0.755, manteniendo la significancia en los test estadísticos realizados. La variable del tiempo de aviso de compresión deja de ser significativa. Extrañamente muchas de las variables introducidas no se muestran significativas. Es el caso de: el tumor primario, las metástasis óseas, el tiempo de desarrollo de déficits motores o el intervalo de tiempo entre la aparición de CM y el diagnóstico primario.

5. Modelizando y prediciendo

5.1 Creación de data sets

Separamos nuestra base de datos en dos conjuntos, uno con el 80% de los datos con finalidades de modelado y otro, con el 20% de los datos, con finalidad de validación del modelo. Como la base de datos está ordenada por orden de reclutamiento de los pacientes y por tanto cronológica en el tiempo, generamos los conjuntos escogiendo de manera aleatoria entre nuestra muestra para evitar sesgos. El dataset de validación contiene pues 26 pacientes y 231 pacientes el dataset de modelización. Para aquellos algoritmos cuya "tarea de aprendizaje" es la clasificación y no la regresión, definimos un punto arbitrario en el tiempo t_0 y replicamos dichos data sets añadiendo una variable adicional (*ReachedEvent*) que contabiliza para cada sujeto si dicho sujeto ha fallecido o no antes de t_0. Estos data sets nos permitirán modelar y validar con algoritmos que no trabajan con probabilidades de supervivencia si no que clasifican los individuos en vivos o muertos para el momento de tiempo considerado. Decidimos usar como t_0 un punto de tiempo cercano a la esperanza de vida de la muestra general.

5.2 Regresión de Cox-Ridge

Aunque hemos realizado un buen modelado de los datos que nos ha permitido establecer cuáles son la variable pronóstica para la supervivencia, no podemos usar en R dicho modelo para la predicción. La función *predictSurvProb* del paquete *Pec,* no admite bases de datos "spliteadas" para la predicción de una regresión de Cox. Como hemos visto, la extensión estratificada del modelo tiene una concordancia mediocre. Nos vemos obligados a usar en este caso el segundo acercamiento clásico cuando se produce una violación de la proporcionalidad de riesgos, introducir dentro del modelo las variables que no cumplen la hipótesis como una interacción dependiente del tiempo. Usaremos en este caso la aproximación de Ridge.

Primero de todo realizaremos la regresión de Cox-Ridge sobre toda la base de datos para comprobar que la significancia y concordancia del modelo son suficientemente buenas. Después modelizaremos mediante Cox-Ridge sobre el dataset de entrenamiento para después predecir sobre el test de validación.

Las variables de *t_ac, t_sc*, tumor primario conocido, radioterapia descompresiva previa, Rades, estado ambulatorio, tratamiento oncológico tras CM e histologías de linfoma / mieloma múltiple siguen siendo significativas para el modelo. Aunque la concordancia es algo más baja 0.731 sigue siendo muy aceptable

```
                                  coef   se(coef)             p
genero                      -0.25777740 0.17941411 0.15078241017
t_ac                        -0.13258163 0.06310723 0.03565021720
t_sc                         0.24942527 0.07833252 0.00145161946
CM_PrimarioConocido          0.18584612 0.16195813 0.25117717562
RADES                       -0.43033906 0.10926761 0.00008202832
RT_descomprPrevia            0.64728177 0.27818138 0.01997412650
ridge(tto_trasCM)           -0.24057035 0.05634079 0.00001955483
ridge(estatusAmbulatorio7 -0.21129729 0.10988466 0.05449252054
histologia[T.1]             -0.23504017 0.47281482 0.61911268697
histologia[T.2]              0.05215084 0.52012519 0.92013319978
histologia[T.3]             -0.01130230 0.50185054 0.98203216867
histologia[T.4]             -0.38472673 0.54450066 0.47983500181
histologia[T.5]             -1.24153411 0.53478269 0.02025623574
histologia[T.6]              0.25884878 0.65770579 0.69390357012
histologia[T.7]             -0.75136124 0.56485571 0.18345811145
histologia[T.8]             -1.15777542 0.71220549 0.10403054168
histologia[T.9]              0.66719673 0.64455455 0.30060894591
histologia[T.10]            -0.24951516 0.69385851 0.71914236566
histologia[T.11]            -1.37053105 0.62739626 0.02892740804
histologia[T.12]            -0.24895971 0.53971076 0.64459516227
opcion_terapeutica           0.24468749 0.17132551 0.15323363337
> summary(coxRidge0)$concordance
         C      se(C)
0.73123688 0.01806265
> summary(coxRidge0)$logtest
        test           df       pvalue
1.425469e+02 2.005700e+01 1.760685e-20
```

Figura 32 Encima, regresión multivariante de Cox-Bridge. Coeficientes, p-valor, Concordancia y test estadístico de Wald

5.3 Árboles de decisión, random forest y GBM

El modelo de Cox es una regresión en su esencia, lo que supone que asume que una línea o curva son suficientes para separar grupos (vivo o muerto) o para estimar probabilidades. A veces modelados del tipo partición como árboles de decisión o sus extensiones de random forest y más avanzado cronológicamente el Gradient Boosting Modelling (GBM), son mejores aproximaciones. Des de la introducción de machine learning en la estadística, este tipo de modelos suponen en estudios de supervivencia la alternativa más común a la regresión de Cox. En este trabajo, hemos aprovechado para alternar algoritmos centrados en la “tarea de aprendizaje” de clasificar, de regresión o ambas.

Figura 33 Algoritmos machine learning y tarea de aprendizaje asociada

Model	Learning Task	Method name	Parameters
k-Nearest Neighbors	Classification	knn	k
Naive Bayes	Classification	nb	fL, usekernel
Decision Trees	Classification	C5.0	model, trials, winnow
OneR Rule Learner	Classification	OneR	None
RIPPER Rule Learner	Classification	JRip	NumOpt
Linear Regression	Regression	lm	None
Regression Trees	Regression	rpart	cp
Model Trees	Regression	M5	pruned, smoothed, rules
Neural Networks	Dual use	nnet	size, decay
Support Vector Machines (Linear Kernel)	Dual use	svmLinear	C
Support Vector Machines (Radial Basis Kernel)	Dual use	svmRadial	C, sigma
Random Forests	Dual use	rf	mtry

En este apartado modelizaremos mediante árboles de decisión usando el algoritmo *C5.0* estándar para clasificación y el *rpart* cuya tarea de aprendizaje es la regresión. Modelizaremos también con dos implementaciones distintas del algoritmo Random Forest (la del paquete *ranger* y la del paquete *randomForestSRC).* Este modelo de supervivencia se basa en la idea de construir varios árboles de decisión asociados cada uno de ellos al procesado de un pedazo de la muestra de datos de manera que al final un promedio de las estimaciones realizadas por cada uno de ellos, conlleve una estimación de la supervivencia, el modelo usa ambas tareas (clasificación y regresión) como aprendizaje. Ambas implementaciones usan una tarea u otra según los datos de entrada. El modelado más reciente por GBM también usa varios árboles de decisión con tarea dual de aprendizaje regresión o clasificación, pero empieza el proceso de combinado de sus estimaciones des del principio y no al final de todo. Se usará como clasificador.

Cuando visualizamos el árbol de decisión usado en el paquete *rpart* aparecen como variables importantes en el algoritmo de decisión en los 5 primeros nodos, de manera análoga a la regresión de Cox-Bridge, el tratamiento tras CM, la histología, el índice RADES y el tiempo des de la sospecha clínica al diagnóstico. Sin embargo, a partir de aquí aparecen otras variables que no habían salido anteriormente como tiempo hasta el inicio del tratamiento, el centro, el número de vértebras afectas, el índice Blisky o el empeoramiento neurológico. Así intervalos de tiempo superiores a 9 días en el inicio de tratamiento combinados con tiempos de sospecha clínica cortos, unas histologías determinadas y realización de tratamiento oncológico posterior, son augurio de supervivencia. De manera yuxtapuesta a la regresión de Cox univariante para el centro, tratamientos en el Hospital Duran i Reynals son preferibles para ciertos pacientes. Así como un número de vértebras afectas inferior a 3. Para pacientes de mal pronóstico con un tipo de histologías determinada, un índice C de Blisky sin empeoramiento neurológico son síntoma de buen pronóstico.

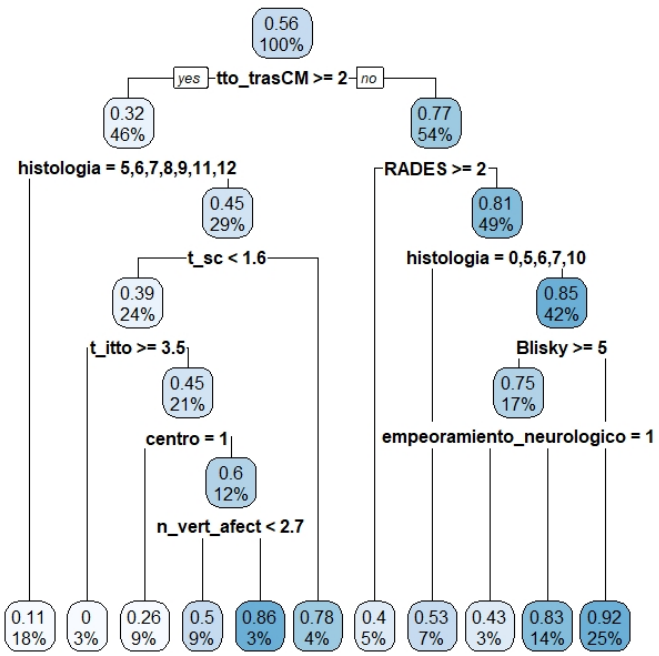


Figura 34 Árbol de decisión rpart

Para el modelado Random Forest obtenemos las siguientes variables de importancia. Como podemos observar son bastante parecidas que las obtenidas en la regresión de Cox. En los primeros puestos aparece también el tratamiento tras CM, la histología del tumor, Rades y el estado ambulatorio a los 7 días. Las siguientes variables son el género, la opción terapéutica y el tumor primario conocido que, aunque no tenían significancia en nuestro modelo sí que fueron introducidas ya que mejoraban la concordancia general del modelo. Llama la atención la poca importancia asociada a la radioterapia descompresiva previa, así como una importancia parecida asociada al índice SINS que la que goza el tumor primario conocido. El índice SINS no aparecía ni en la regresión de Cox ni en el árbol de decisión.

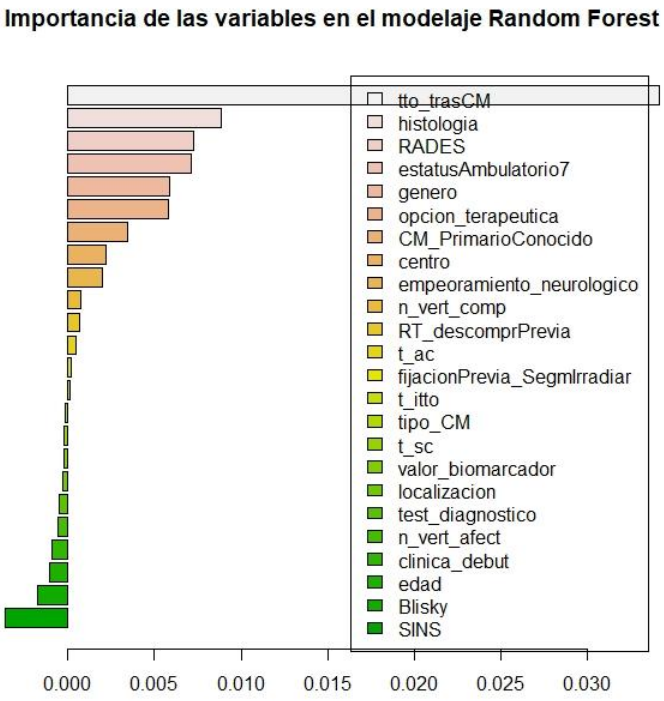
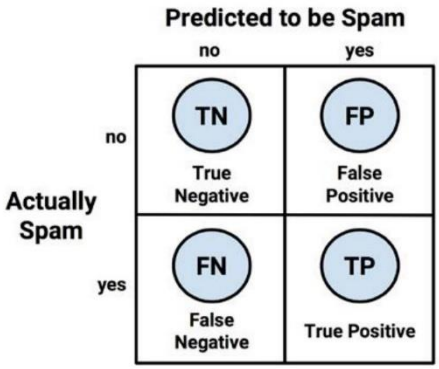


Figura 35 Importancia de las variables en el modelado RandomForest

6. Analizando la bondad del ajuste

Comparamos los resultados obtenidos mediante diferentes métodos, presentaremos las matrices de confusión de cada uno de los modelados usados, junto con sus estimaciones derivadas de la eficiencia, sensibilidad, sensitividad, precisión y estimador kappa. Definimos estos conceptos según (38) como ejemplificamos en la siguiente figura.



$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad \kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

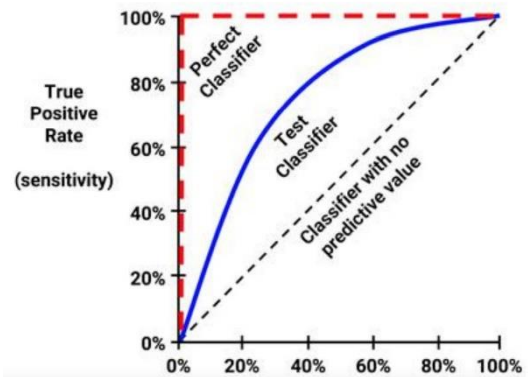
$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Figura 36 Matrices de confusión, eficiencia, sensitividad, sensibilidad, precisión y estimador kappa

El estadístico kappa representa un "ajuste" a la clásica eficiencia teniendo en cuenta la posibilidad de realizar una predicción correcta al azar. Ello pone de manifiesto clasificadores que obtienen una alta eficiencia simplemente escogiendo siempre la clase más frecuente. Consideraremos un muy buen ajuste valores por encima de 0.8, un ajuste bueno valores entre 0.6 y 0.8, un ajuste moderado valores entre 0.4 y 0.6 y un ajuste aceptable entre 0.2 y 0.4. La precisión es también conocida como "positive predictive value" y el recall como "negative predictive value".

En referencia al área bajo la curva consideraremos también la estimación de (38)



- **A**: Outstanding = 0.9 to 1.0
- **B**: Excellent/good = 0.8 to 0.9
- **C**: Acceptable/fair = 0.7 to 0.8
- **D**: Poor = 0.6 to 0.7
- **E**: No discrimination = 0.5 to 0.6

Figura 37 A la izquiera curva ROC, a la derecha significado del valor del área bajo la curva ROC

Para las curvas de predicción usaremos la implementación del paquete *Pec.* Paquetes como *Goffte, Gof, Crskdiag* que implementa tests basados en la función de distribución empírica tipo tests Kolmogorov-Smirnov, Cramer-von Mises o Anderson-Darling, con la intención de valorar la bondad del ajuste en el caso de la regresión de Cox no han podido ser usados ya que la última versión disponible es para la versión de R 1.0.

6.1 Matrices de confusión

Observamos que los modelados por árbol tienen ambos problemas tanto con los falsos positivos como con los falsos negativos. El algoritmo cree que el sujeto ha fallecido cuando está vivo o a la inversa respectivamente, independientemente de la tarea de aprendizaje en la que se base. También presenta problemas similares con los falsos positivos la predicción GBM y el algoritmo Random Forest del paquete ranger. Los algoritmos Random Forest pese estar generados con parámetros parecidos presentan soluciones muy distintas. La predicción Random Forest del paquete SRC presenta los mismos problemas que los algoritmos de árbol con los falsos negativos, aunque no muestre problema con los falsos positivos como le sucede a su homólogo. De todas formas, parece que en general los modelos tienden a tener más dificultades en identificar a un sujeto que está vivo como vivo, que a un sujeto que ha muerto como muerto. Eso significa que todos los modelos tienden a sobreestimar el evento (el fallecimiento).

realidad	predicción Cox-Ridge 0	1	Row Total
0	18	8	26
	0.383	0.170	
1	3	18	21
	0.064	0.383	
Column Total	21	26	47

realidad	predicción GBM 0	1	Row Total
0	15	11	26
	0.319	0.234	
1	3	18	21
	0.064	0.383	
Column Total	18	29	47

realidad	predicción Árbol C5.0 0	1	Row Total
0	13	13	26
	0.277	0.277	
1	5	16	21
	0.106	0.340	
Column Total	18	29	47

realidad	predicción Árbol rpart 0	1	Row Total
0	16	10	26
	0.340	0.213	
1	5	16	21
	0.106	0.340	
Column Total	21	26	47

realidad	predicción random forest RANGER 0	1	Row Total
0	14	12	26
	0.298	0.255	
1	2	19	21
	0.043	0.404	
Column Total	16	31	47

realidad	predicción random forestSRC 0	1	Row Total
0	20	6	26
	0.426	0.128	
1	5	16	21
	0.106	0.340	
Column Total	25	22	47

Figura 37 Matrices de confusión para los diferentes modelos al clasificar los sujetos vivos y muertos existentes en un tiempo t_0

6.2 Eficacia, sensibilidad, sensitividad, precisión y estimador kappa

Vamos a estimar analíticamente nuestras primeras impresiones visuales.

```
               Accuracy : 0.766
                 95% CI : (0.6197, 0.877)
    No Information Rate : 0.5532
    P-Value [Acc > NIR] : 0.002096

                  Kappa : 0.5372

 Mcnemar's Test P-Value : 0.227800

            Sensitivity : 0.8571
            Specificity : 0.6923
         Pos Pred Value : 0.6923
         Neg Pred Value : 0.8571
             Prevalence : 0.4468
         Detection Rate : 0.3830
   Detection Prevalence : 0.5532
      Balanced Accuracy : 0.7747
```

```
               Accuracy : 0.7021
                 95% CI : (0.5511, 0.8266)
    No Information Rate : 0.617
    P-Value [Acc > NIR] : 0.14645

                  Kappa : 0.4187

 Mcnemar's Test P-Value : 0.06137

            Sensitivity : 0.8333
            Specificity : 0.6207
         Pos Pred Value : 0.5769
         Neg Pred Value : 0.8571
             Prevalence : 0.3830
         Detection Rate : 0.3191
   Detection Prevalence : 0.5532
      Balanced Accuracy : 0.7270
```

```
               Accuracy : 0.617
                 95% CI : (0.4638, 0.7549)
    No Information Rate : 0.617
    P-Value [Acc > NIR] : 0.56407

                  Kappa : 0.2527

 Mcnemar's Test P-Value : 0.09896

            Sensitivity : 0.7222
            Specificity : 0.5517
         Pos Pred Value : 0.5000
         Neg Pred Value : 0.7619
             Prevalence : 0.3830
         Detection Rate : 0.2766
   Detection Prevalence : 0.5532
      Balanced Accuracy : 0.6370
```

```
               Accuracy : 0.6809
                 95% CI : (0.5288, 0.8091)
    No Information Rate : 0.5532
    P-Value [Acc > NIR] : 0.05185

                  Kappa : 0.3688

 Mcnemar's Test P-Value : 0.30170

            Sensitivity : 0.7619
            Specificity : 0.6154
         Pos Pred Value : 0.6154
         Neg Pred Value : 0.7619
             Prevalence : 0.4468
         Detection Rate : 0.3404
   Detection Prevalence : 0.5532
      Balanced Accuracy : 0.6886
```

```
               Accuracy : 0.7021
                 95% CI : (0.5511, 0.8266)
    No Information Rate : 0.6596
    P-Value [Acc > NIR] : 0.32710

                  Kappa : 0.4238

 Mcnemar's Test P-Value : 0.01616

            Sensitivity : 0.8750
            Specificity : 0.6129
         Pos Pred Value : 0.5385
         Neg Pred Value : 0.9048
             Prevalence : 0.3404
         Detection Rate : 0.2979
   Detection Prevalence : 0.5532
      Balanced Accuracy : 0.7440
```

```
               Accuracy : 0.766
                 95% CI : (0.6197, 0.877)
    No Information Rate : 0.5319
    P-Value [Acc > NIR] : 0.0008185

                  Kappa : 0.5287

 Mcnemar's Test P-Value : 1.0000000

            Sensitivity : 0.8000
            Specificity : 0.7273
         Pos Pred Value : 0.7692
         Neg Pred Value : 0.7619
             Prevalence : 0.5319
         Detection Rate : 0.4255
   Detection Prevalence : 0.5532
      Balanced Accuracy : 0.7636
```

Figura 38 Rendimiento de los modelados realizados con distintos algoritmos. De izquierda a derecha y de arriba a abajo: Cox-Ridge, GBM, árbol de decisión C5.0, árbol de decisión rpart, Random Forest ranger, Random Forest SRS

En general la eficacia de los algoritmos es de moderada a buena entre aproximadamente 0.6 y 0.8 según el modelo. El modelo de árbol de decisión con algoritmo C5.0 presenta unos valores de eficiencia y del estimador kappa (aproximadamente 0.6 y 0.2 respectivamente) algo bajos. Los algoritmos RandomForestSRS y Cox-Ridge presentan resultados similares y son los que muestran un rendimiento más elevado. Cox-Ridge es ligeramente "mejor" con una buena eficiencia de 0.77 y un estimador kappa mejor que aceptable de 0.53. Los tres modelos resultantes presentan resultados similares siendo el GBM muy similar al RandomForest ranger, y el árbol de decisión rpart ligeramente inferior. El sistema considera como evento positivo la clase 0 (es decir vivir), de ahí que todos los sistemas muestren una mejor sensibilidad (detectar el máximo de pacientes posibles que están vivos, aunque ello suponga etiquetar algunos erróneamente) que sensitividad (detectar el máximo de pacientes muertos, aunque ello suponga etiquetar algunos erróneamente). Como hemos visto anteriormente, el sistema tiene más problemas para identificar a los fallecidos. La precisión, también conocida como "Positive Predictive Value" hace referencia a cuan preciso es el algoritmo en su detección. Es decir, de los sujetos clasificados como vivos ¿qué porcentaje de ellos están realmente vivos? En ese aspecto todos los modelos muestran dificultades.

6.3 Curvas Roc. Área bajo la curva

Al estimar el área bajo la curva para un tiempo t_0 fijado obtenemos:

Regresión de Cox-Bridge	0,89
GBM	0,81
Árbol de decisión C5.0	0,63
Árbol de decisión rpart	0,76
RandomForest ranger	0,83
RandomForest SRS	0,83

Para aquellos modelos que generan por defecto un listado de probabilidades de muerte/vida y no una clasificación entre grupo de muertos y grupo de vivos, se ha asimilado probabilidades de muerte mayores al 50% al grupo de los muertos y probabilidades menores al 50% al grupo de los vivos. Al estimar las predicciones del modelo para un tiempo fijo cercano a la esperanza de vida el modelo que mejor actúa es claramente el de la regresión de Cox-Ridge. Los modelados por RandomForest y el GBM muestran una AUC parecida. El único modelo que muestra resultados mediocres en cuanto a la AUC es el árbol de decisión C5.0.

Estos resultados quedan claramente reflejados en la visualización de la curva ROC para todos los algoritmos. El árbol de clasificación C5.0 refleja claramente que el modelo ha sido creado concretamente para una situación temporal concreta y no modelizado en general y aplicado a una situación en concreto.

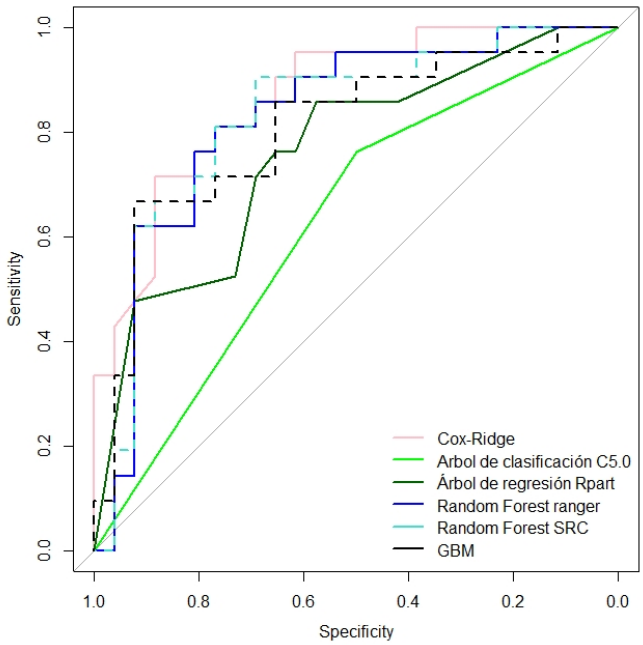


Figura 39 Curvas ROC para los distintos modelos considerados

6.4 Curvas de predicción

El paquete *Pec* permite para algunas modelizaciones en concreto como Cox-Ridge o RandomForest del paquete RandomForestSRS, generar las curvas de predicción de supervivencia para todos los sujetos del test de validación. Obviamente aquellos algoritmos usados como clasificadores en que no se generan probabilidades de supervivencia si no una clasificación en el grupo de muerto/vivo, no está disponible esa opción. En otros casos, simplemente la función pertinente no tiene implementada la opción de graficar las probabilidades para ese modelo en concreto.

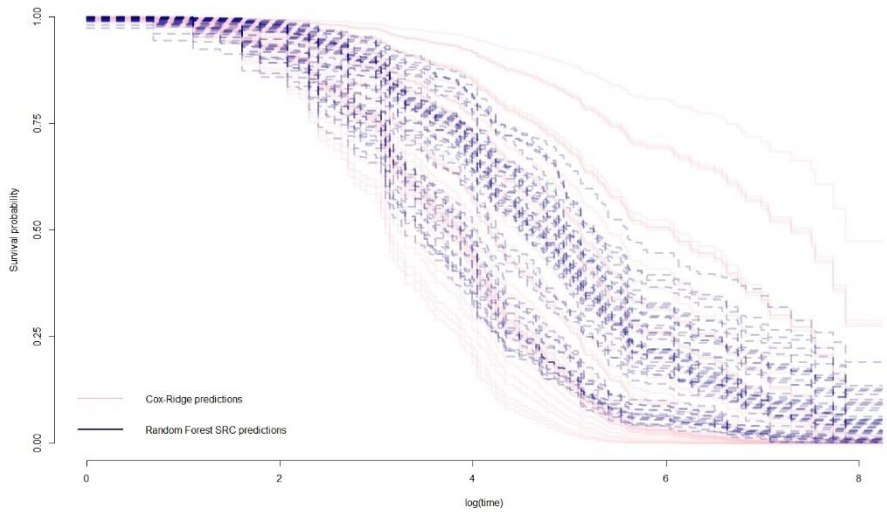


Figura 40 Curvas de predicción para Cox-Ridge en rosa y RandomForestSRS

Como podemos ver las curvas de supervivencia del modelado mediante RandomForestSRS son mucho más estrechas mientras que las predicciones para el modelado de Cox-Ridge se “desparrama” más en gráfico, generando individuos con predicciones de supervivencia muy cortas o muy largas. Ello ha sido reportado ya en la literatura.

7. Conclusiones y reflexión

- De acuerdo con la literatura previa existente, los pacientes con mayor índice de Rades muestran mejor supervivencia de manera significativa estadísticamente

- No se pueden establecer conclusiones significativas acerca de la supervivencia de los pacientes diagnosticados estrictamente previamente al 2018 o posteriormente

- Las covariables identificadas como pronóstico mediante el modelo de Cox de riesgos proporcionales son de mayor a menor significación: realizar tratamiento oncológico tras el tratamiento de la CM, pertenecer a grupos Rades elevados, gozar de un estado ambulatorio a los 7 días de finalizar el tratamiento, tener un tumor cuya histología sea mieloma múltiple o linfoma y no haber realizado previamente radioterapia descompresiva previa.

- El índice de Blisky, indicativo del grado de afectación de la CM, no se muestra significativo para la supervivencia, en discrepancia a lo que apuntaban las últimas guías clínicas al respecto

- El valor del biomarcador sólo correlaciona con la supervivencia mediante una regresión de Cox univariante, en discrepancia a lo que apuntaba la última guía clínica publicada

- Las covariables tratamiento oncológico tras la CM y estado ambulatorio del paciente, sólo son pronósticas de supervivencia en un primer intervalo de tiempo tras la finalización del tratamiento. El estado ambulatorio a los 7 días sólo resulta pronostico durante el primer mes, mientras que el tratamiento oncológico tras la CM lo hace durante los tres primeros meses

- Las variables más importantes en el modelado mediante árboles de decisión y random forest son las más significativas en la regresión de Cox: tratamiento tras CM, Rades, histología y estado ambulatorio a los 7 días.

- El modelado mediante árboles de decisión o random forest incorpora además otras variables de importancia distintas según el modelo usado que no son significativas para la regresión de Cox. Estas variables son: el conocimiento del tumor primario de la CM, el género, el índice SINS, el

índice Blisky, el centro de tratamiento, la opción terapéutica realizada o el número de vértebras afectas.

- El modelado mediante árboles de decisión muestra resultados peores que los demás modelos, de acuerdo con la literatura, independientemente de su tarea de aprendizaje (regresión o clasificación)

- Las predicciones del modelo de Cox-Bridge y del modelo RadomForestSRS son las más acertadas, siendo ligeramente mejor la primera.

- El modelo de Cox-Bridge predice un rango de probabilidades de supervivencia más ancho que el modelo de Cox-Bridge

En general, aunque se han logrado los objetivos y el enfoque metodológico que se ha seguido en el TFM ha sido adecuado, la planificación ha sido errónea. Se tuvo en cuenta un intervalo de tiempo suficientemente largo hasta la obtención de una base de datos llena, pero se infraestimó completamente la cantidad de tiempo necesaria para “limpiarla”.

Para un futuro TFM habría que intervenir directamente en el diseño de la base de datos y habría que prever un margen de tiempo suficientemente largo para la limpieza de las erratas. Ello permitiría evitar errores. Por ejemplo, para esta base de datos se habría introducido en primer lugar la identificación de cada observación mediante dos variables, ello evitaría la búsqueda posterior en la historia clínica de si el sujeto ha tenido más de una CM o si es un error. Además, se incorporaría la posibilidad de definir como CM múltiple una CM que no sólo sucede en localizaciones distintas si no que afecta de manera distinta al canal medular. Tal como estaba la base de datos, se podía definir una CM como múltiple 2CM que aparecían simultáneamente en por ejemplo la zona dorsal y cervical y eran ambas medulares. Pero no se podía definir como CM múltiple un par de CM que aparecían en las mismas zonas siendo una medular y la otra radicular. Estas CM se introducían por separado. Han debido luego de “perseguirse” para homogeneizar los criterios. En tercer lugar, se hubiera calculado directamente y exportado como variable el valor de la puntuación SINS y RADES para que no se hubiera tenido que calcular luego mediante una des tabulación de las variables tabuladas. En cuarto lugar, se hubiera obligado en los campos en que al menos debe existir una opción verdadera, a ser rellenada. Los campos vacíos son exportados en Access como “Falso” lo que genera confusiones de si son realmente falsos o están vacíos. En quinto lugar, se hubiera definido la variable de “biomarcadores” de manera más precisa. La definición actual no permite, salvo existencia de una correlación muy y muy fuerte, un buen análisis estadístico. Los valores de “no existe biomarcador para ese tumor”, “el biomarcador existe, pero clínicamente no se ha solicitado o no se puede solicitar”, “el biomarcador existe y es negativo” caen en el mismo saco, lo que dificulta establecer conclusiones sólidas al respecto. Finalmente, y para ir acabando, se habrían introducido alarmas en las correlaciones de tiempo avisando de fechas ilógicas o raras.

En un futuro, me gustaría estudiar varios aspectos. En un plano más práctico y cercano, me gustaría repasar uno por uno con los médicos especialistas los distintos outliers y puntos extremales para decidir si suponen una realidad del sistema (gestión defectuosa, errores humanos, casos clínicamente infrecuentes pero existentes) o son errores. Sería interesante volver a realizar todo el estudio con los outliers introducidos y observar si cambian mucho las conclusiones obtenidas. También me gustaría redefinir la variable de biomarcadores y rellenar esa información para todos los pacientes, para estudiar si los resultados obtenidos varían o son similares. Otro punto interesante sería seguir reclutando pacientes unos 3-4 años más, esperar 7 años para cerrar el ensayo clínico y rehacer el trabajo. Una comparativa de reclutamiento de pacientes entre 2009 a 2017, contra únicamente pacientes de 2018 no es legítima.

En un plano más amplio, me gustaría seguir investigando el paquete *Pec* para obtener los errores en las curvas de predicción y el índice de Brier como explica algún artículo (38). Los comandos que parecían pertinentes daban error. Me gustaría seguir investigando por qué daban ese error, así como otras funcionalidades del paquete. Además, pienso que analizar los datos mediante un estudio jerárquico de clústeres para detectar pacientes con patrones similares, aportaría información pertinente. Por último, sería interesante escoger un t_0 muy pequeño y otro muy grande y comparar el rendimiento de Cox-Bridge y el de RandomForestSRS.

8. Glosario

AUC: área bajo la curva
CM: compresión medular
CR: cirugía
GBM: Gradient Boosting Modelling
RF: Random Forest
ROC: receiver operating characteristic
RT: radioterapia
TFM: trabajo final de máster
Tto: tratamiento

9. Bibliografia

1. Loblaw DA. Systematic Review of the Diagnosis and Management of Malignant Extradural Spinal Cord Compression: The Cancer Care Ontario Practice Guidelines Initiative's Neuro-Oncology Disease Site Group. Journal of Clinical Oncology. 2005 Feb 7;23(9):2028–37.
2. Boriani S, Gasbarrini A, Bandiera S, Ghermandi R, Lador R. En bloc resections in the spine—the experience of 220 cases over 25 years. World Neurosurg 2017; 98: 217–29. 5 Sakaura H, Hosono N, Mukai Y, Ishii T, Yonenobu K, Yoshikawa H.
3. Outcome of total en bloc spondylectomy for solitary metastasis of the thoracolumbar spine. J Spinal Disord Tech 2004; 17: 297–300. 6
4. Tomita K, Kawahara N, Baba H, Tsuchiya H, Nagata S, Toribatake Y. Total en bloc spondylectomy for solitary spinal metastases. Int Orthop 1994; 18: 291–98
5. Maranzano E, Bellavita R, Rossi R, et al. Short-course versus split-course radiotherapy in metastatic spinal cord compression: results of a phase III, randomized, multicenter trial. J Clin Oncol 2005; 23: 3358–65. 8
6. Katagiri H, Takahashi M, Inagaki J, et al. Clinical results of nonsurgical treatment for spinal metastases. Int J Radiat Oncol Biol Phys 1998; 42: 1127–32.
7. Rades et al, The first Score predicting overall survival in patients with metastasic spinal cord compression, Cancer vol 112 nº1, 2008
8. Stephen Lutz MD a,∗, Tracy Balboni MD MPH b , Joshua Jones MD c , Simon Lo MB ChB d , Joshua Petit MD e , Shayna E. Rich MD PhD f , Rebecca Wong MB ChB g , Carol Hahn MD h Palliative radiation therapy for bone metastases: Update of an ASTRO Evidence-Based Guideline. Practical Radiation Oncology (2017) 7, 4-12
9. Patchell RA, Tibbs PA, Regine WF, Payne R, Saris S, Kryscio RJ, et al. Direct decompressive surgical resection in the treatment of spinal cord compression caused by metastatic cancer: a randomised trial. The Lancet. 2005 Aug;366(9486):643–8.
10. Loblaw DA, Mitera G. The Optimal Dose Fractionation Schema for Malignant Extradural Spinal Cord Compression. The Journal of Supportive Oncology. 2011 Jul;9(4):121–4.
11. Loblaw DA, Mitera G, Ford M, Laperriere NJ. A 2011 Updated Systematic Review and Clinical Practice Guideline for the Management of Malignant Extradural Spinal Cord Compression. International Journal of Radiation Oncology*Biology*Physics. 2012 Oct;84(2):312–7.
12. Rades D, Stalpers LJ, Veninga T, et al: Evaluation of five radiation schedules and prognostic factors for metastatic spinal cord compression. J Clin Oncol 23: 3366-3375, 2005
13. Hoskin P, Misra V, Hopkins K, et al: SCORAD III: Randomized noninferiority phase III trial of single-dose radiotherapy (RT) compared to multifraction RT in patients (pts) with metastatic spinal canal compression (SCC). J Clin Oncol 35, 2017 (suppl, abstr LBA10004)

14. Spratt DE, Beeler WH, de Moraes FY, Rhines LD, Gemmete JJ, Chaudhary N, et al. An integrated multidisciplinary algorithm for the management of spinal metastases: an International Spine Oncology Consortium report. The Lancet Oncology. 2017 Dec;18(12):e720–30.
15. Lawton AJ, Lee KA, Cheville AL, Ferrone ML, Rades D, Balboni TA, et al. Assessment and Management of Patients With Metastatic Spinal Cord Compression: A Multidisciplinary Review. JOURNAL OF CLINICAL ONCOLOGY. 2018 Sept; 37:61-71.
16. Fisher et al, A novel classification system for spinal instability in neoplastic disease: an evidence-based approach and expert consensus from the Spine Oncology Group 2010 Oct 15; 35(22):E1221-1229
17. Bilsky MH, Laufer I, Fourney DR, Groff M, Schmidt MH, Varga PP, et al. Reliability analysis of the epidural spinal cord compression scale. Journal of Neurosurgery: Spine. 2010 Sep;13(3):324–8.
18. Switlyk MD, Kongsgaard U, Skjeldal S, et al. Prognostic factors in patients with symptomatic spinal metastases and normal neurological function. Clin Oncol (R Coll Radiol) 2015; 27: 213–2
19. Bollen L, van der Linden YM, Pondaag W, et al. Prognostic factors associated with survival in patients with symptomatic spinal bone metastases: a retrospective cohort study of 1,043 patients. Neuro Oncol 2014; 16: 991–98.
20. Leithner A, Radl R, Gruber G, et al: Predictive value of seven preoperative prognostic scoring systems for spinal metastases. Eur Spine J 17:1488-1495, 2008
21. Schoenfeld AJ, Le HV, Marjoua Y, et al: Assessing the utility of aclinical prediction score regarding 30-day morbidity and mortality following metastatic spinal surgery: The New England Spinal Metastasis Score (NESMS). Spine J 16:482-490, 2016
22. Fehlings MG, Nater A, Tetreault L, et al: Survival and clinical outcomes in surgically treated patients with metastatic epidural spinal cord compression: Results of the prospective multicenter AOSpine study. J Clin Oncol 34:268-276, 2016 (Díficil estimar la prognosis)
23. Verlaan JJ, Choi D, Versteeg A, et al: Characteristics of patients who survived , 3 months or . 2 years after surgery for spinal metastases: Can we avoid inappropriate patient selection? J Clin Oncol 34:3054-3061, 2016 62.
24. Nater A, Martin AR, Sahgal A, et al: Symptomatic spinal metastasis: A systematic literature review of the preoperative prognostic factors for survival, neurological, functional and quality of life in surgically treated patients and methodological recommendations for prognostic studies. PLoS One 12:e0171507, 2017
25. Laufer I, Rubin DG, Lis E, et al: The NOMS framework: Approach to the treatment of spinal metastatic tumors. Oncologist 18:744-751, 2013 (NOMS)
26. Barzilai O, Laufer I, Yamada Y, et al: Integrating evidence-based medicine for treatment of spinal metastases into a decision framework: Neurologic, oncologic, mechanicals stability, and systemic disease. J Clin Oncol 35:2419-2427, 2017 (NOMS)
27. Paton GR, Frangou E, Fourney DR. Contemporary treatment strategy for spinal metastasis: the “LMNOP” system. Can J Neurol Sci 2011; 38: 396–403. (LMNOP)

28. Rades D, Huttenlocher S, Dunst J, et al: Matched pair analysis comparing surgery followed by radiotherapy and radiotherapy alone for metastatic spinal cord compression. J Clin Oncol 28:3597-3604, 2010 73
29. Rades D, Huttenlocher S, Bajrovic A, et al: Surgery followed by radiotherapy versus radiotherapy alone for metastatic spinal cord compression from unfavorable tumors. Int J Radiat Oncol Biol Phys 81:e861-e868, 2011
30. Gandhi L, Rodríguez-Abreu D, Gadgeel S, Esteban E, Felip E, De Angelis F, et al. Pembrolizumab plus Chemotherapy in Metastatic Non–Small-Cell Lung Cancer. New England Journal of Medicine [Internet]. 2018 Apr 16 [cited 2018 Apr 26]
31. Loblaw DA, Laperriere NJ, Mackillop WJ: A population-based study of malignant spinal cord compression in Ontario. Clin Oncol (R Coll Radiol) 15:211-217, 2003
32. Bach F, Larsen BH, Rohde K, et al: Metastatic spinal cord compression. Occurrence, symptoms, clinical presentations and prognosis in 398 patients with spinal cord compression. Acta Neurochir (Wien) 107:37-43, 1990
33. Sperduto PW, Yang TJ, Beal K, Pan H, Brown PD, Bangdiwala A, et al. Estimating Survival in Patients With Lung Cancer and Brain Metastases: An Update of the Graded Prognostic Assessment for Lung Cancer Using Molecular Markers (Lung-molGPA). JAMA Oncology. 2017 Jun 1;3(6):827.
34. Sperduto PW, Kased N, Roberge D, Xu Z, Shanley R, Luo X, et al. Summary Report on the Graded Prognostic Assessment: An Accurate and Facile Diagnosis-Specific Tool to Estimate Survival for Patients With Brain Metastases. Journal of Clinical Oncology. 2012 Feb 1;30(4):419–25.
35. Sfumato P, Filleron T, Giorgi R, Cook RJ, Boher JM. Goftte: A R package for assessing goodness-of-fit in proportional (sub) distributions hazards regression models. Computer Methods and Programs in Biomedicine. 2019 May 30 Aug;177:269-275
36. Mogensen UB, Ishwaran H, Gerds TA
Evaluating Random Forests for Survival Analysis using Prediction Error Curves. J Stat Softw. 2012 Sep;50(11):1-23.
37. Lantz, Machine Learning with R, Packt Publishing 2015
38. Mogensen et al, Evaluating Random Forests for Survival Analysis Using Prediction Error Curves, Journal of Statistical Software September 2012, Volume 50, Issue 11.
39. https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf, enero 2021

Anexo

Tabla1_SINS

Element of SINS	Score
Location	
Junctional (occiput-C2, C7–T2, T11–L1, L5–S1)	3
Mobile spine (C3–C6, L2–L4)	2
Semi-rigid (T3–T10)	1
Rigid (S2–S5)	0
Pain relief with recumbency and/or pain with movement/loading of the spine	
Yes	3
No (occasional pain but not mechanical)	1
Pain free lesion	0
Bone lesion	
Lytic	2
Mixed (lytic/blastic)	1
Blastic	0
Radiographic spinal alignment	
Subluxation/translation present	4
De novo deformity (kyphosis/scoliosis)	2
Normal alignment	0
Vertebral body collapse	
>50% collapse	3
<50% collapse	2
No collapse with >50% body involved	1
None of the above	0
Posterolateral involvement of the spinal elements (facet, pedicle or CV joint fracture or replacement with tumor)	
Bilateral	3
Unilateral	1
None of the above	0

Tabla2_Escala Bilsky

TABLE 3. The Epidural Spinal Cord Compression Scale

Grade	Description
0	Bone-only disease
1a	Epidural impingement, without deformation of the thecal sac
1b	Deformation of the thecal sac, without spinal cord abutment
1c	Deformation of the thecal sac, with spinal cord abutment, without cord compression
2	Spinal cord compression, with CSF visible around the cord
3	Spinal cord compression, with no CSF visible around the cord

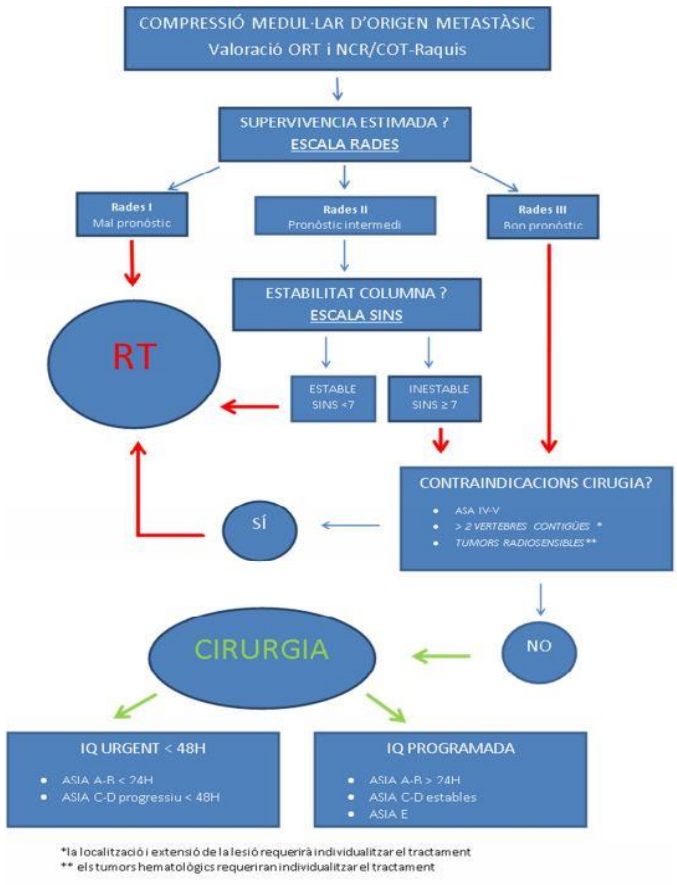
Tabla3_ Escala Rades

Table 1. Significant Prognostic Factors and Corresponding Scores

Prognostic Factor	Score
Type of primary tumor	
Breast cancer	8
Prostate cancer	7
Myeloma/lymphoma	9
Lung cancer	3
Other tumors	4
Other bone metastases at the time of RT	
Yes	5
No	7
Visceral metastases at the time of RT	
Yes	2
No	8
Interval from tumor diagnosis to MSCC, mo	
≤15	4
>15	7
Ambulatory status before RT	
Ambulatory	7
Nonambulatory	3
Time of developing motor deficits before RT, d	
1-7	3
8-14	6
>14	8

RT indicates radiotherapy; MSCC, metastatic spinal cord compression.

Figura1_ Algoritmo de decisión ICO



Código R

```{r setup, include=FALSE, echo=FALSE}
library(knitr)
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE,
                      comment = NA, prompt = TRUE, tidy = FALSE,
                      fig.width = 7, fig.height = 7, fig_caption = TRUE,
                       cache=TRUE, error = TRUE)
Sys.setlocale("LC_TIME", "C")
```

```{r echo=FALSE}
if(!(require(printr))) {
  install.packages(
    'printr',
    type = 'source',
    repos = c('http://yihui.name/xran', 'http://cran.rstudio.com')
  )
}
```

```{r installPackages, message=FALSE, warning=FALSE, eval=FALSE}
install.packages("knitr")
install.packages("Rcmdr")
install.packages("plotrix")
install.packages("gplots")
install.packages("ggplot2")
install.packages("DMwR")
install.packages("htmlTable")
install.packages("prettydoc")
install.packages("plyr")
install.packages("car")
install.packages("e1071")
install.packages("GGally")
install.packages("rcompanion")
install.packages("EnvStats")
install.packages("vcd")
install.packages("scales")
install.packages("survival")
install.packages("gtsummary")
install.packages("survminer")
install.packages("mlr3proba")
install.packages("ranger")
install.packages("pROC")
install.packages("gbm")
install.packages("C50")
install.packages("partykit")
install.packages("gmodels")
install.packages("caret")
install.packages("pec")
install.packages("randomForestSRC")
install.packages("rpart")
install.packages("rpart.plot")
install.packages("dlookr")
install.packages("tidyverse")
```

```{r decargaPaquetes, message=FALSE, warning=FALSE}
library(knitr)
library(Rcmdr)
library(plyr)
library(plotrix)
library(DMwR)
library(ggplot2)
library(gplots)
library(prettydoc)
library(htmlTable)
library(car)
library(e1071)
library(rcompanion)
library(EnvStats)
library(GGally)
library(vcd)
```

```
library(scales)
library(survival)
library(gtsummary)
library(survminer)
library(mlr3proba)
library(ranger)
library(pROC)
library(gbm)
library(C50)
library(partykit)
library(gmodels)
library(caret)
library(pec)
library(randomForestSRC)
library(rpart)
library(rpart.plot)
library(dlookr)
library(tidyverse)
```
```{r message=FALSE, warning=FALSE}
CM0Access <- readXL("./Datos/Base de datos_v0_1.xlsx",
  rownames=FALSE, header=TRUE, na="", sheet="Registro TFM", stringsAsFactors=TRUE)
```

```{r message=FALSE, warning=FALSE}
CM0 <-
  readXL("./Datos/Base de datos_v3.xlsx",
  rownames=FALSE, header=TRUE, na="", sheet="Data", stringsAsFactors=TRUE)
```

Base de datos manipulada eliminando las variables originales (usadas para crear otras) y que ya no son de interés
```{r message=FALSE, warning=FALSE}
#creación de un nuevo CMframe con sólo las variables de interés
CM <- subset(CM0[,-c(3,9:11,13:15,18:19,21:22,24:27,33:84,88:94,96,106,109:116,125:128,143)])
```

Base de datos eliminando las variables que tienen un porcentaje excesivo de NA's
```{r message=FALSE, warning=FALSE}
#creación de un nuevo CMframe con sólo las variables de interés
CM <- subset(CM[,-c(41,42,43,44,48)]) #aprovechamos para eliminar también el début de clínica asintomático que se nos había "colado"
```

Definimos una base de datos con sólo las variables numéricas, substituimos los valores NA por primeros vecinos. Redondeamos para mantener los mismo decimales que en el resto de los datos.
```{r message=FALSE, warning=FALSE}
#definimos el nuevo subset con sólo las variables numéricas
```

```
CM_num <- subset(CM, select = c(ID, edad, n_vert_afect, n_vert_comp, survival, t_ac, t_itto, t_sc, puntuacion_RADES, puntuacion_SINS))
#subsituimos los valores de NA que faltan con el método de los primeros vecinos knnImputation
cleanCM_num <- round(knnImputation(CM_num, k= 10, scale = TRUE, meth = "weighAvg", distData = NULL))
```

Definimos la base de datos numérica transformada y la general con las variables numéricas transformadas, ambas con los valores NA substituidos.

```{r}
# Definimos las nuevas variables transformadas
log.survival <- round(log(cleanCM_num$survival),2)
log.n_vert_afect <- round(log(cleanCM_num$n_vert_afect),2)
sqrt.n_vert_comp <- round(sqrt(cleanCM_num$n_vert_comp),2)
sqrt.t_sc <- round(sqrt(cleanCM_num$t_sc),2)
sqrt.t_ac <- round(sqrt(cleanCM_num$t_ac),2)
sqrt.t_itto <- round(sqrt(cleanCM_num$t_itto),2)

# creamos un nuevo data.frame transformada y "limpo" de NA y eliminamos los sujetos que en sus transformaciones logarítmicas tienen valores iniciales 0, ya que no existe la transformada. En el caso de las vértebras afectas no es posible que sean 0 (no habría CM).
cleanCM_numT <- data.frame(cleanCM_num$ID, log.survival, cleanCM_num$edad,cleanCM_num$puntuacion_RADES,
cleanCM_num$puntuacion_SINS, log.n_vert_afect, sqrt.n_vert_comp, sqrt.t_sc, sqrt.t_ac, sqrt.t_itto)
cleanCM_numT <- subset(cleanCM_numT[-which(cleanCM_num$survival == 0),])
apply(cleanCM_numT, 2, skewness)

# creamos también el dataframe que contenga todas las variables categóricas y las variables numéricas sin NA, y otro en que no haya NA pero las variables numéricas sean transformadas
CM$SAP <- as.numeric(CM$SAP)
cleanCM <- round(knnImputation(CM))
# creamos el nuevo dataframe y remplazamos las variables
cleanCM_T <- data.frame(cleanCM)
cleanCM_T$survival <- log(cleanCM$survival)
cleanCM_T$n_vert_afect <- sqrt(cleanCM$n_vert_afect)
cleanCM_T$n_vert_comp <- sqrt(cleanCM$n_vert_comp)
cleanCM_T$t_ac <- sqrt(cleanCM$t_ac)
cleanCM_T$t_sc <- sqrt(cleanCM$t_sc)
cleanCM_T$t_itto <- sqrt(cleanCM$t_itto)
# Eliminamos los sujetos con sueprvivencia e 0 días para los que no existe log
cleanCM_T <- subset(cleanCM_T[-c(which(cleanCM$survival == 0)),])
```

Definimos los outliers:

```{r}
# "boxplot.stats(cleanCM_num$survival)" nos da una série de estadísticas de las cuales la función outl corresponde el valor del outlier. Com la función which encontramos la hilera dónde se produce ese valor
```

```
outSurv <- which(cleanCM_num$survival %in% c(boxplot.stats(cleanCM_num$survival)$out))
outSurvT <- which(cleanCM_numT$log.survival %in% c(boxplot.stats(cleanCM_numT$log.survival)$out))

outVertAf <- which(cleanCM_num$n_vert_afect %in% c(boxplot.stats(cleanCM_num$n_vert_afect)$out))
outVertAfT <- which(cleanCM_numT$log.n_vert_afect %in% c(boxplot.stats(cleanCM_numT$log.n_vert_afect)$out))

outVertC <- which(cleanCM_num$n_vert_comp %in% c(boxplot.stats(cleanCM_num$n_vert_comp)$out))
outVertCT <- which(cleanCM_numT$sqrt.n_vert_comp %in% c(boxplot.stats(cleanCM_numT$sqrt.n_vert_comp)$out))

outSINS <- which(cleanCM_num$puntuacion_SINS %in% c(boxplot.stats(cleanCM_num$puntuacion_SINS)$out))
cleanCM_num$ID[outSINS]
outSINST <- which(cleanCM_numT$puntuacion_SINS %in% c(boxplot.stats(cleanCM_numT$cleanCM_num.puntuacion_SINS)$out))

outt_sc <- which(cleanCM_num$t_sc %in% c(boxplot.stats(cleanCM_num$t_sc)$out))
outt_scT <- which(cleanCM_numT$sqrt.t_sc %in% c(boxplot.stats(cleanCM_numT$sqrt.t_sc)$out))

outt_ac <- which(cleanCM_num$t_ac %in% c(boxplot.stats(cleanCM_num$t_ac)$out))
outt_acT <- which(cleanCM_numT$sqrt.t_ac %in% c(boxplot.stats(cleanCM_numT$sqrt.t_ac)$out))

outt_itto <- which(cleanCM_num$t_itto %in% c(boxplot.stats(cleanCM_num$t_itto)$out))
outt_ittoT <- which(cleanCM_numT$sqrt.t_itto %in% c(boxplot.stats(cleanCM_numT$sqrt.t_itto)$out))
```

Creamos dos bases de datos. Ambas sin NA y sin los ouliers, una con las variables numéricas transformadas y la otra sin transformar.

```{r}
cleanCM_Tout <- subset(cleanCM_T[-c(outSINST, outVertCT, outt_ittoT, outt_acT, outt_scT), ])
dim(cleanCM_Tout)
# En la base de datos no transformada existen dos pacientes más por lo que el número de la hilera del outlier no coincide. Detectamos de qué outlier se trata y luego calculamos su hilera para borrarlo.
IDout <- cleanCM_T$ID[c(outSINST, outVertCT, outt_acT, outt_ittoT, outt_scT)]
cleanCM_KM <- subset(cleanCM, !(cleanCM$ID %in% IDout))
```

Creamos la base de datos según la cual realizaremos la regresión de Cox y el modelizaje con machine learning

```{r}
```

```
# Eliminamos todas las variables que dependen unas de otras y sólo dejamos la que creemos más representativa
CMcox <-subset(cleanCM_Tout[,-c(14,15,22:29,31:38)])
CMcox <- select(CMcox, -c("estatus_lesion_UC","dosis_RT"))
CMcox$histologia <- as.factor(round(CMcox$histologia))
dim(CMcox)
```

Tarea3.1 Comprobar que efectivamente los pacientes tienen mejor o peor supervivencia según la puntuación de RADES obtenida

```{r}
KMrades <- survfit(Surv(time = cleanCM_KM$survival, event = cleanCM_KM$estatus_UC) ~ cleanCM_KM$RADES, conf.type="plain")
```

```{r fig22 Curvas Kapplan-Meier según RADES warning = FALSE, message=FALSE, eval=FALSE}
ggsurvplot(
  KMrades,
  data = cleanCM_KM,
  size = 1,                 # change line size
  palette =
    c("#E7B800", "#2E9FDF", "pink"),# custom color palettes
  conf.int = TRUE,          # Add confidence interval
  pval = TRUE,              # Add p-value
  risk.table = TRUE,        # Add risk table
  risk.table.col = "strata",# Risk table color by groups
  legend.labs =
    c("RADES1 20-30p", " RADES2 31-35p", "RADES3 36-45p"),    # Change legend labels
  risk.table.height = 0.25, # Useful to change when you have multiple groups
  ggtheme = theme_bw(),      # Change ggplot2 theme
  ncensor.plot = TRUE,      # plot the number of censored subjects at time t
  ncensor.plot.height = 0.25
)
```

Tarea3.2 Trazar e interpretar las curvas de Kaplan-Meier para ambos grupos y estimar la tasa de supervivencia a 6 meses

Vamos a contar con el paquete "survival" para ello. El cálculo del estimador de Kaplan-Meier no tiene en cuenta, obviamente, para el cálculo de la supervivencia los tiempos censurados (es decir, el de aquellos pacientes que el sujeto está vivo a cierre del estudio o ha habido una pérdida del seguimiento o por cualquier motivo ha salido del estudio).

El estimador Kaplan-Meier de toda la base de datos es:

```{r}
# Estimador Kapplan-Meier para nuestra base de datos sin distinguir épocas de tiempo
KMGlobal <- survfit(Surv(time = cleanCM_KM$survival, event = cleanCM_KM$estatus_UC) ~ 1, conf.type="plain")
summary(KMGlobal)
```

donde la primera columna es el tiempo (en años) en que estamos valorando la situación, la segunda el número de pacientes vivos al inicio de ese intérvalo, la tercera columna el nº de pacientes que fallecen entre el tiempo considerado y el siguiente, la cuarta la supervivencia o estimador de Kaplan-Meier, la quinta la desviación estándar de esa probabilidad y la sexta y la séptima los límites del intervalo de confianza del 95%.

Vamos a visualizar las estimaciones de la función de supervivencia según el método de Kaplan-Meier para cada grupo (anterior o posterior al 2018).

```{r}
# Estimador Kapplan-Meier del Grupo1 (anterior al 2018) y Grupo2 (posterior o igual 2018)
KMgrupos <- survfit(Surv(time = cleanCM_KM$survival, event = cleanCM_KM$estatus_UC) ~ cleanCM_KM$grupo, conf.type="plain")
summary(KMgrupos)
```

```{r fig17, warning = FALSE, message=FALSE, eval=FALSE}
plot(KMGlobal, col=c("black","red","red"),
xlab="Tiempo de supervivencia (días)", ylab="Probabilidades de supervivencia")
legend("bottomleft", c("Surv Global", "CI inf", "CI sup"), lty=c("solid","dashed","dashed"),
col=c("black","red","red"))
```

Gráficamente, visualizamos las curvas de supervivencia como

```{r fig.cap="Fig17_Curva de supervivencia global", echo=FALSE}
knitr::include_graphics("figures/fig17.jpeg")
```

```{r fig18, warning = FALSE, message=FALSE, eval=FALSE}
ggsurvplot(
  KMgrupos,
  data = cleanCM_KM,
  size = 1,                 # change line size
  palette =
    c("#E7B800", "#2E9FDF"),# custom color palettes
  conf.int = TRUE,          # Add confidence interval
  pval = TRUE,              # Add p-value
  risk.table = TRUE,        # Add risk table
  risk.table.col = "strata",# Risk table color by groups
  legend.labs =
    c("Anterior 2018", "2018 y posterior"),    # Change legend labels
  risk.table.height = 0.25, # Useful to change when you have multiple groups
  ggtheme = theme_bw(),      # Change ggplot2 theme
  ncensor.plot = TRUE,      # plot the number of censored subjects at time t
  ncensor.plot.height = 0.25
)
```

```{r}
KMGlobal
KMgrupos
KMrades
```

```

```{r}
cleanCM1_KM <- subset(cleanCM_KM, cleanCM$grupo == 1)
cleanCM2_KM <- subset(cleanCM_KM, cleanCM$grupo == 2)
censoredCM1 <- sum(length(which(cleanCM1_KM$estatus_UC == 0)))
censoredCM1
censoredCM1/sum(length(cleanCM1_KM$estatus_UC))*100
censoredCM2 <- sum(length(which(cleanCM2_KM$estatus_UC == 0)))
censoredCM2
censoredCM2/sum(length(cleanCM2_KM$estatus_UC))*100
```
```{r}
outgrupo <- cleanCM_T$grupo[c(outt_acT, outt_ittoT, outt_scT, outVertCT, outSINST)]
sum(length(which(outgrupo == 1)))
sum(length(which(outgrupo == 2)))
sum(length(which(outgrupo == 1)))/sum(length(cleanCM1_KM$estatus_UC))*100
sum(length(which(outgrupo == 2)))/sum(length(cleanCM2_KM$estatus_UC))*100
```

```{r}
chisq.test(table(cleanCM$RADES,cleanCM$dosis_RT))
```
Debemos refutar la hipótesis de independencia y por tanto eliminamos dosis_RT.

```{r}
# Eliminamos todas las variables que dependen unas de otras y sólo dejamos la que creemos más representativa
CMcox <-subset(cleanCM_Tout[,-c(14,15,22,31:38)])
CMcox <- select(CMcox, -c("estatus_lesion_UC","dosis_RT"))
CMcox$histologia <- as.factor(round(CMcox$histologia))
CMcox$tiempoDesarrollo_DefMotor_RADES <- as.factor(round(CMcox$tiempoDesarrollo_DefMotor_RADES))
CMcox$tumPrimario_RADES <- as.factor(round(CMcox$tumPrimario_RADES))
dim(CMcox)
```

```{r warning=FALSE}
set.seed(18734)

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$genero),300,replace=TRUE)), as.factor(sample(1:length(CMcox$genero),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$centro),300,replace=TRUE)), as.factor(sample(1:length(CMcox$centro),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$localizacion),300,replace=TRUE)), as.factor(sample(1:length(CMcox$localizacion),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$tipo_CM),300,replace=TRUE)), as.factor(sample(1:length(CMcox$tipo_CM),300,replace=TRUE)))))
```
```

```
chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$histologia),300,replace=
TRUE)), as.factor(sample(1:length(CMcox$histologia),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$valor_biomarcador),300,
replace=TRUE)),
as.factor(sample(1:length(CMcox$valor_biomarcador),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$RADES),300,replace=T
RUE)), as.factor(sample(1:length(CMcox$RADES),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$SINS),300,replace=TRU
E)), as.factor(sample(1:length(CMcox$SINS),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$Blisky),300,replace=TR
UE)), as.factor(sample(1:length(CMcox$Blisky),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$clinica_debut),300,repla
ce=TRUE)), as.factor(sample(1:length(CMcox$clinica_debut),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$dosis_RT),300,replace=
TRUE)), as.factor(sample(1:length(CMcox$dosis_RT),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$opcion_terapeutica),300
,replace=TRUE)),
as.factor(sample(1:length(CMcox$opcion_terapeutica),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$tto_trasCM),300,replace
=TRUE)), as.factor(sample(1:length(CMcox$tto_trasCM),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$RT_descomprPrevia),30
0,replace=TRUE)),
as.factor(sample(1:length(CMcox$RT_descomprPrevia),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$CM_PrimarioConocido),
300,replace=TRUE)),
as.factor(sample(1:length(CMcox$CM_PrimarioConocido),300,replace=TRUE)))))

chisq.test(table(data.frame(as.factor(sample(1:length(CMcox$test_diagnostico),300,re
place=TRUE)),
as.factor(sample(1:length(CMcox$test_diagnostico),300,replace=TRUE)))))
```

```{r fig19, distribución de la variable SINS al comparar los dos grupos}
ggplot(CMcox, aes(grupo,  fill = SINS)) +
  geom_bar(position="dodge",colour="black")+
  labs(x= "Grupos",  y="casos de CM", fill="SINS")+
  ylim(c(0,400))+
  geom_text(aes(label=..count..),stat='count',position=position_dodge(0.9),
            vjust=-0.5, size=5.0)+
  facet_wrap(~"Casos de CM por grado de SINS en cada grupo")+theme_bw(base_size
= 15)+theme(legend.position = c(0.7, 0.8))
```

```{r}
survCMcox <- Surv(time = CMcox$survival, event = CMcox$estatus_UC)
survdiff(survCMcox ~ CMcox$grupo)
```

```{r}
survdiff(survCMcox ~ CMcox$RADES)
```

```{r fig24}
coxhisto <- coxph(survCMcox ~ histologia, data = CMcox)
coxhisto %>% gtsummary::tbl_regression(exp = TRUE)
summary(coxhisto)
```

```{r}
covariates <- c("edad", "genero",  "centro", "grupo", "t_ac", "t_sc", "t_itto", "localizacion",
"tipo_CM",  "n_vert_afect",  "n_vert_comp",  "CM_PrimarioConocido",
"valor_biomarcador",  "RADES",  "SINS",  "Blisky",  "clinica_debut",
"RT_descomprPrevia",  "empeoramiento_neurologico",  "estatusAmbulatorio7",
"fijacionPrevia_SegmIrradiar", "test_diagnostico", "opcion_terapeutica", "tto_trasCM")

univ_formulas <- sapply(covariates,
                        function(x) as.formula(paste('survCMcox ~', x)))

univ_models <- lapply( univ_formulas, function(x){coxph(x, data = CMcox)})

# Extract data
univ_results <- lapply(univ_models,
                       function(x){
                         x <- summary(x)
                         p.value<-signif(x$wald["pvalue"], digits=2)
                         wald.test<-signif(x$wald["test"], digits=2)
                         beta<-signif(x$coef[1], digits=2);#coeficient beta
                         HR <-signif(x$coef[2], digits=2);#exp(beta)
                         HR.confint.lower <- signif(x$conf.int[,"lower .95"], 2)
                         HR.confint.upper <- signif(x$conf.int[,"upper .95"],2)
                         HR <- paste0(HR, " (",
                                      HR.confint.lower, "-", HR.confint.upper, ")")
                         res <- c(beta, HR, wald.test, p.value)
                         names(res)<- c("beta", "HR (95% CI for HR)",  "wald.test",
"p.value")
                         return(res)
                         #return(exp(cbind(coef(x),confint(x))))
                        })
res <- t(as.data.frame(univ_results, check.names = FALSE))
as.data.frame(res)
```

```{r}
set.seed(1234)
lapply(univ_models, function(x){cox.zph(x)})
```

```

```{r}
CMcoxStep <- survSplit(Surv(survival, estatus_UC) ~ ., data= subset(CMcox[-which(CMcox$survival == 0),]), cut=c(log(30), log(90)), episode= "tgroup", id="id", zero = 0)

coxStepgenero <- coxph(Surv(tstart, survival, estatus_UC) ~ genero:strata(tgroup), data=CMcoxStep, x = TRUE, y = TRUE)
summary(coxStepgenero)
```

```{r}
coxSteptto <- coxph(Surv(tstart, survival, estatus_UC) ~ tto_trasCM:strata(tgroup), data=CMcoxStep, x = TRUE, y = TRUE)
summary(coxSteptto)
```

```{r}
set.seed(1234)
cox.zph(coxStepgenero)
cox.zph(coxSteptto)
```

```{r}
coxmulti <- coxph(Surv(survival, estatus_UC) ~ edad + genero + centro + t_ac + t_sc + t_itto + localizacion + tipo_CM + n_vert_afect + n_vert_comp + histologia + CM_PrimarioConocido + valor_biomarcador + RADES + SINS + Blisky + clinica_debut + RT_descomprPrevia + empeoramiento_neurologico + estatusAmbulatorio7 + tto_trasCM + fijacionPrevia_SegmIrradiar + test_diagnostico + opcion_terapeutica, data = CMcox)
summary(coxmulti)
```

```{r}
coxmulti <- coxph(Surv(survival, estatus_UC) ~ genero + t_ac + t_sc + histologia + CM_PrimarioConocido + RADES + RT_descomprPrevia + tto_trasCM + estatusAmbulatorio7, data = CMcox)
summary(coxmulti)
```
```{r}
coxmulti <- coxph(Surv(survival, estatus_UC) ~ t_ac + t_sc + histologia + CM_PrimarioConocido + RADES + RT_descomprPrevia + tto_trasCM + estatusAmbulatorio7, data = CMcox)
summary(coxmulti)
```
```{r}
coxmultiF <- coxph(Surv(survival, estatus_UC) ~ genero + t_ac + t_sc + CM_PrimarioConocido + RADES + RT_descomprPrevia + tto_trasCM + estatusAmbulatorio7 + histologia + opcion_terapeutica, data = CMcox, x = TRUE, y = TRUE)
summary(coxmultiF)
```
```

```
```{r}
testPH_multi <- cox.zph(coxmultiF)
testPH_multi
```
```{r}
coxmultiEst <- coxph(Surv(survival, estatus_UC) ~ genero + t_ac + t_sc + CM_PrimarioConocido + RADES + RT_descomprPrevia + strata(tto_trasCM) + strata(estatusAmbulatorio7) + histologia + opcion_terapeutica, data =  CMcox)
summary(coxmultiEst)
cox.zph(coxmultiEst)
```
```{r}
CMcoxStep <- survSplit(Surv(survival, estatus_UC) ~ ., data= subset(CMcox[-which(CMcox$survival == 0),]), cut=c(log(30), log(90)), episode= "tgroup", id="id", zero = 0)

coxmultiStep <- coxph(Surv(tstart, survival, estatus_UC) ~ genero + t_ac + t_sc + CM_PrimarioConocido + RADES + RT_descomprPrevia + tto_trasCM:strata(tgroup) + estatusAmbulatorio7:strata(tgroup) + histologia + opcion_terapeutica, data=CMcoxStep, x = TRUE, y = TRUE)
summary(coxmultiStep)
```
```{r}
test <- cox.zph(coxmultiStep)
test
ggcoxzph(test, point.size = 0.5)

```{r fig27 Residuos desviados message=FALSE, warning=FALSE}
ggcoxdiagnostics(coxmultiStep, type = "deviance", linear.predictions = FALSE, ggtheme = theme_bw(), point.size = 0.5)
```


```{r}
round(summary(coxmultiStep)$coefficients[,c(1,2,5)],3)
summary(coxmultiStep)$concordance
summary(coxmultiStep)$waldtest
summary(coxmultiStep)$logtest
coxStepPron <- coxph(Surv(tstart, survival, estatus_UC) ~ genero + t_ac + t_sc + CM_PrimarioConocido + RT_descomprPrevia + tto_trasCM:strata(tgroup) + estatusAmbulatorio7:strata(tgroup) + histologia + opcion_terapeutica + tiempoDesarrollo_DefMotor_RADES + estAmbulatorio_RADES + tumPrimario_RADES + metOseas_RADES + metVisc_RADES + intervalo_diagCM_RADES, data=CMcoxStep, x = TRUE, y = TRUE)
round(summary(coxStepPron)$coefficients[,c(1,2,5)],3)
summary(coxStepPron)$concordance
summary(coxStepPron)$waldtest
summary(coxStepPron)$logtest
```


```{r}
set.seed(1234)
```

```
random_splits <- runif(nrow(CMcox))
CM_train <- CMcox[random_splits < .8,]
dim(CM_train)
CM_test <- CMcox[random_splits >= .8,]
dim(CM_test)
prop.table(table(CM_train$grupo))
prop.table(table(CM_test$grupo))
```
```{r}
table(CM_train$survival)
```
El valor más parecido es 4.51085950651685, que es el *unique_death_time* 61 de los 150 en el que se avalúa el modelo.

```{r}
t0 <- 4.51085950651685
```

Creamos los dos datasets de validación y modelización con una variable adicional *ReachedEvent* que contabiliza para cada sujeto si dicho sujeto ha fallecido o ha pasado cualquier otra cosa para el tiempo especificado.
```{r}
CM_train_classif  <- CM_train
CM_train_classif$ReachedEvent <- ifelse((CM_train_classif$estatus_UC==1 & CM_train_classif$survival <= 4.51085950651685), 1, 0)
summary(CM_train_classif$ReachedEvent)
CM_test_classif  <- CM_test
CM_test_classif$ReachedEvent <- ifelse((CM_test_classif$estatus_UC==1 & CM_test_classif$survival <= 4.51085950651685), 1, 0)
summary(CM_test_classif$ReachedEvent)
```

Creamos una variable con todas las variables de estudio excluyendo la creada "ReachedEvent", la variable de salida (survival) y la variable de censura (estatus_UC).
```{r}
feature_names <- setdiff(names(CM_train_classif), c('ReachedEvent', 'survival', 'estatus_UC'))
```
```{r}
# Miramos cómo sería la concordancia del modelo con la opción Ridge
coxRidge0 <- coxph(Surv(survival,estatus_UC) ~ genero + t_ac + t_sc + CM_PrimarioConocido + RADES + RT_descomprPrevia + ridge(tto_trasCM) + ridge(estatusAmbulatorio7) + histologia + opcion_terapeutica, data = CMcox, x=TRUE, y=TRUE)
summary(coxRidge0)$coefficients[,c(1,2,5)]
summary(coxRidge0)$concordance
summary(coxRidge0)$logtest
```

```{r}
# Generamos la opción de Cox con la opción Ridge para la base de datos de entrenamiento
```

```
coxRidge <- coxph(Surv(survival,estatus_UC) ~ genero + t_ac + t_sc + RADES + RT_descomprPrevia + ridge(tto_trasCM) + ridge(estatusAmbulatorio7) + histologia + opcion_terapeutica, data = CM_train, x=TRUE, y=TRUE)
summary(coxRidge0)$coefficients[,c(1,2,5)]
summary(coxRidge0)$concordance
summary(coxRidge0)$logtest
cox.zph(coxRidge)
```

```{r}
cox_pred <- predictSurvProb(coxRidge, newdata = CM_test, times = t0, x= TRUE, y=TRUE)
t(cox_pred)
```

```{r}
# Creamos las formula de lo que queremos considerar
tree_formula <- formula(paste('ReachedEvent ~ ','edad + genero + centro + t_ac + t_sc + t_itto + localizacion + tipo_CM + n_vert_afect + n_vert_comp + histologia + CM_PrimarioConocido + valor_biomarcador + RADES + SINS + Blisky + clinica_debut + RT_descomprPrevia + empeoramiento_neurologico + estatusAmbulatorio7 + tto_trasCM + fijacionPrevia_SegmIrradiar + test_diagnostico + opcion_terapeutica'))
```

El algoritmo rpart genera un vector con las probabilidades de fallecimiento de cada paciente en el instante de tiempo t0.

```{r}
# Creamos el modelado del árbol de decisión
treepart_model <- rpart(tree_formula, CM_train_classif, x= TRUE, y = TRUE)
#Ejecutamos nuestras predicciones
treepart_pred <- predict(treepart_model, CM_test_classif, type = "vector", x= TRUE, y = TRUE)
```

```{r}
# Creamos el modelado del árbol de decisión
CM_train_classif$ReachedEvent <- as.factor(CM_train_classif$ReachedEvent)
tree_model <- C5.0(tree_formula, CM_train_classif)
#Ejecutamos nuestras predicciones
tree_pred <- predict(tree_model, CM_test_classif[,..feature_names])
```

Si ejecutamos *tree_pred* observamos un factor de 47 componentes con dos níveles. Nos informa del estado vivo (0) / muerto (1) de cada uno de los pacientes en el instante de tiempo que queremos evaluar. Para *treepart_pred* obtenemos la pobabilidad de fallecimiento para cada uno de los componenetes en un tiempo t0.

```{r}
length(tree_pred)
tree_pred
treepart_pred
```

```{r fig23 árbol de decisión por rpart, warning=FALSE, message=FALSE}
rpart.plot(treepart_model)
```

```
```{r warning=FALSE, message=FALSE}
# Definimos el modelo: outcome y
random_formula <- formula(paste('Surv(', 'survival', ',', 'estatus_UC', ') ~ ','edad + genero + centro + t_ac + t_sc + t_itto + localizacion + tipo_CM + n_vert_afect + n_vert_comp + histologia + CM_PrimarioConocido + valor_biomarcador + RADES + SINS + Blisky + clinica_debut + RT_descomprPrevia + empeoramiento_neurologico + estatusAmbulatorio7 + tto_trasCM + fijacionPrevia_SegmIrradiar + test_diagnostico + opcion_terapeutica'))

random_model <- ranger(random_formula,
            data = CM_train,
            seed = 1234,
            importance = 'permutation',
            mtry = 2,
            verbose = TRUE,
            num.trees = 1000,
            write.forest=TRUE,
            x = TRUE,
            y = TRUE)
# Inspeccionamos nuestro modelo
random_model

set.seed(1234)
random_model1 <- rfsrc(random_formula, data = CM_train, ntree = 1000)
```

Vamos a visualizar cuáles de nuestras variables son más significativas
```{r fig20 Importancia de las variables en el modelado random forest}
sort(random_model$variable.importance)
barplot(sort(random_model$variable.importance), names.arg = " ", legend.text = TRUE, horiz = TRUE, col = terrain.colors(random_model$num.independent.variables))
```
Generamos las predicciones sobre el grupo de test
```{r}
random_pred <- predict( random_model, CM_test[, c('edad', 'genero', 'centro', 't_ac', 't_sc', 't_itto', 'localizacion', 'tipo_CM', 'n_vert_afect', 'n_vert_comp', 'histologia', 'CM_PrimarioConocido', 'valor_biomarcador', 'RADES', 'SINS', 'Blisky', 'clinica_debut', 'RT_descomprPrevia', 'empeoramiento_neurologico', 'estatusAmbulatorio7', 'tto_trasCM', 'fijacionPrevia_SegmIrradiar', 'test_diagnostico', 'opcion_terapeutica')] )

random_pred1 <- predictSurvProb(random_model1,newdata = CM_test, times = t0, x=TRUE, y=TRUE)
```
```{r}
classification_formula <- formula(paste('ReachedEvent ~ ','edad + genero + centro + t_ac + t_sc + t_itto + localizacion + tipo_CM + n_vert_afect + n_vert_comp + histologia + CM_PrimarioConocido + valor_biomarcador + RADES + SINS + Blisky + clinica_debut + RT_descomprPrevia + empeoramiento_neurologico + estatusAmbulatorio7 + tto_trasCM + fijacionPrevia_SegmIrradiar + test_diagnostico + opcion_terapeutica'))

set.seed(1234)
```
```

```
# redefinimos el conjunto CM_train_classif porque lo necesitamos como numérico y no como factor el ReachedEvent, si sólo lo reconvertieramos quedarían los niveles 1 y 2, no 0 y 1
CM_train_classif  <- CM_train
CM_train_classif$ReachedEvent <- ifelse((CM_train_classif$estatus_UC==1 & CM_train_classif$survival <= 4.51085950651685), 1, 0)

gbm_model = gbm(classification_formula,
                data =  CM_train_classif,
                distribution='bernoulli',
                n.trees=1000,
                interaction.depth=3,
                shrinkage=0.01,
                bag.fraction=0.5,
                keep.data=FALSE,
                cv.folds=5)
```

Generamos las predicciones del modelo:

```{r}
nTrees <- gbm.perf(gbm_model)
gbm_pred <- predict(gbm_model, newdata=CM_test_classif[,..feature_names], type="response", n.trees=nTrees)
```

```{r Tabla coxRidge}
CrossTable(CM_test_classif$ReachedEvent, round(1-cox_pred),
prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('realidad', 'predicción GBM'))
```

```{r tabla árboles de clasificación c5.0}
CrossTable(CM_test_classif$ReachedEvent, tree_pred, prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('realidad', 'predicción Árbol C5.0'))
```

```{r tabla árboles de regresión rpart}
CrossTable(CM_test_classif$ReachedEvent, round(treepart_pred), prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('realidad', 'predicción Árbol rpart'))
```

```{r tabla Random Forest ranger}
CrossTable(CM_test_classif$ReachedEvent, round(1-random_pred$survival[,which(random_pred$unique.death.times==4.51085950651685)]), prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
dnn = c('realidad', 'predicción random forest RANGER'))
```

```{r tabla Random forest SRC}
CrossTable(CM_test_classif$ReachedEvent, round(1-random_pred1), prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('realidad', 'predicción random forestSRC'))
```

```{r Tabla GBM}
CrossTable(CM_test_classif$ReachedEvent, round(gbm_pred),
prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('realidad', 'predicción GBM'))
```

Miramos la eficiencia, la sensibilidad, la especificidad, el estadístico kappa mediante *ConfusionMatrix* del paquete caret.
```{r tabla de confusión coxRidge}
confusionMatrix(as.factor(CM_test_classif$ReachedEvent),as.factor(round(1-cox_pred)), dnn = c('realidad', 'predicción Cox Ridge'))
```

```{r tabla de confusión arbol C5.0}
confusionMatrix(as.factor(CM_test_classif$ReachedEvent),as.factor(tree_pred), dnn = c('realidad', 'predicción Árbol de decisión'))
```

```{r tabla de confusión arbol rpart}
confusionMatrix(as.factor(CM_test_classif$ReachedEvent),as.factor(round(treepart_pred)), dnn = c('realidad', 'predicción Árbol de decisión'))
```

```{r tabla de confusión random forest}
confusionMatrix(as.factor(CM_test_classif$ReachedEvent),as.factor(round(1-random_pred$survival[,which(random_pred$unique.death.times==4.51085950651685)])), dnn = c('realidad', 'predicción random forest Ranger'))
```

```{r tabla de confusión random forestSRC}
confusionMatrix(as.factor(CM_test_classif$ReachedEvent),as.factor(round(1-random_pred1)), dnn = c('realidad', 'predicción random forest SRC'))
```

```{r tabla de confusión GBM}
confusionMatrix(as.factor(CM_test_classif$ReachedEvent),as.factor(round(gbm_pred)), dnn = c('realidad', 'predicción GBM'))
```
```{r Área bajo la curva en un modelado por cox ridge, message=FALSE, warning=FALSE}
roccox <- roc(response = CM_test_classif$ReachedEvent, predictor = as.numeric(1-cox_pred))
```

```{r Area bajo la curva en un modelado por árboles C5.0, message=FALSE, warning=FALSE}
roctree <- roc(response = CM_test_classif$ReachedEvent, predictor = as.numeric(tree_pred))
```

```{r Area bajo la curva en un modelado por árboles rpart, message=FALSE, warning=FALSE}
```

```
roctreepart <- roc(response = CM_test_classif$ReachedEvent, predictor = as.numeric(treepart_pred))
```

```{r Área bajo la curva en un modelado por random forest, message=FALSE, warning=FALSE}
rocforest <- roc(response = CM_test_classif$ReachedEvent, predictor = 1-random_pred$survival[,which(random_pred$unique.death.times==4.51085950651685)])
```

```{r Area bajo la curva en un modelado por Random Forest SRC, message=FALSE, warning=FALSE, message=FALSE, warning=FALSE}
rocforestSRC <- roc(response = CM_test_classif$ReachedEvent, predictor = as.numeric(1- random_pred1))
```

```{r Area bajo la curva en un GBM, message=FALSE, warning=FALSE, message=FALSE, warning=FALSE}
rocgbm <- roc(response = CM_test_classif$ReachedEvent, predictor = gbm_pred)
```

```{r fig21 curvas ROC}
# Visualizamos las tres áreas bajo la curva
plot(roccox, col = "pink")
plot(roctree, add = TRUE, col= "green")
plot(roctreepart, add = TRUE, col= "dark green")
plot(rocforest, add = TRUE, col = "blue")
plot(rocforestSRC, add = TRUE, col = "turquoise", lty = 2)
plot(rocgbm, add = TRUE, col = "black", lty = 2)
legend("bottomright", legend=c("Cox-Ridge", "Arbol de clasificación C5.0", "Árbol de regresión Rpart", "Random Forest ranger", "Random Forest SRC", "GBM"), col=c("pink", "green", "dark green", "blue", "turquoise", "black"), lwd=2, bty = "n")
```

```{r}
plotPredictSurvProb(coxRidge, newdata = CM_test, lty = 1, col = "pink", xlab = "log(time)")
plotPredictSurvProb(random_model1, newdata = CM_test, add = TRUE, lty = 2, col = "dark blue")
legend("bottomleft", legend=c("Cox-Ridge predictions", "Random Forest SRC predictions"), col=c("pink", "dark blue"), lwd=2, bty = "n")
```

```{r message=FALSE, warning=FALSE,eval=FALSE}
CM$genero <- factor(CM$genero, labels = c("hombre","mujer"))
CM$centro <- factor(CM$centro, labels = c("ICO Badalona", "ICO Hospitalet"))
CM$grupo <- factor(CM$grupo, labels = c("<2018", ">=2018"))
CM$opcion_terapeutica <- factor(CM$opcion_terapeutica, labels = c("CR", "CR+RT", "RT+CR", "RT", "SOPORTE"))
CM$localizacion <- factor(CM$localizacion, labels = c("cervical", "dorsal", "lumbar", "sacra", "múltiple"))
CM$tipo_CM <- factor(CM$tipo_CM, labels = c("medular", "radicular", "intramedular"))
CM$vert_afect <- factor(CM$vert_afect, labels = c("1","1-5",">5"))
CM$vert_comp <- factor(CM$vert_comp, labels = c("0","1",">1"))
```

```
CM$CM_PrimarioConocido <- factor(CM$CM_PrimarioConocido, labels = c("no", "si"))
CM$valor_biomarcador <- factor(CM$valor_biomarcador, labels = c("negativo/no disponible", "positivo"))
CM$RADES <- factor(CM$RADES, labels = c("Grupo1", "Grupo2", "Grupo3"))
CM$SINS <- factor(CM$SINS, labels = c("Estable", "Potencialmente inestable", "inestable"))
CM$Blisky <- factor(CM$Blisky, labels = c("0", "A1", "A2", "A3", "B", "C"))
CM$RT_descomprPrevia <- factor(CM$RT_descomprPrevia, labels = c("no", "si"))
CM$empeoramiento_neurologico <- factor(CM$empeoramiento_neurologico, labels = c("no", "si"))
CM$tto_trasCM <- factor(CM$tto_trasCM, labels = c("no inicia", "no continúa", "inicia", "continúa"))
CM$estatusAmbulatorio7 <- factor(CM$estatusAmbulatorio7, labels =c("no", "si"))
CM$estatus_lesion_UC <- factor(CM$estatus_lesion_UC, labels = c("estable","progresión en PTV","progresión fuera de PTV","RC radiológica","RC metabólica", "No evaluada"))
CM$clinica_debut <- factor(CM$clinica_debut, labels = c("asintomático","sensitivo/neuropático","somático","motor","somático+motor","sens/neurop+somático","sens/neurop+motor","todos"))
CM$dosis_RT <- factor(CM$dosis_RT, labels = c("8Gyx1s","5Gyx4s","4Gyx5s","3Gyx10s","Otros"))
CM$metOseas_RADES <- as.factor(CM$metOseas_RADES)
CM$metVisc_RADES <- as.factor(CM$metVisc_RADES)
CM$tumPrimario_RADES <- as.factor(CM$tumPrimario_RADES )
CM$intervalo_diagCM_RADES <- as.factor(CM$intervalo_diagCM_RADES)
CM$estAmbulatorio_RADES <- as.factor(CM$estAmbulatorio_RADES)
CM$tiempoDesarrollo_DefMotor_RADES <- as.factor(CM$tiempoDesarrollo_DefMotor_RADES)
CM$location_SINS <- as.factor(CM$location_SINS)
CM$dolor_SINS <- as.factor(CM$dolor_SINS)
CM$caract_lesion_SINS <- as.factor(CM$caract_lesion_SINS)
CM$alineacColumna_SINS <- as.factor(CM$alineacColumna_SINS)
CM$colapsoVertebral_SINS <- as.factor(CM$colapsoVertebral_SINS)
CM$afectPosterolat_SINS <- as.factor(CM$afectPosterolat_SINS)
CM$histologia <- as.factor(CM$histologia)
CM$fijacionPrevia_SegmIrradiar <- as.factor(CM$fijacionPrevia_SegmIrradiar)
CM$test_diagnostico <- factor(CM$test_diagnostico, labels = c("RMN","TC", "otros"))
```

```{r}
summary(CM)
min(sort(CM0$fecha_diagnostico))
max(sort(CM0$fecha_diagnostico))
max(sort(CM0$fecha_UC))
CM1 <- select(CM, -c(location_SINS, alineacColumna_SINS, colapsoVertebral_SINS, afectPosterolat_SINS, caract_lesion_SINS, SINS_Calc, RADES_Calc))
str(CM1)
```

```{r fig1, message=FALSE, warning=FALSE, eval=FALSE}
#par(mar=c(1,1,1,1))
par(mfrow=c(1,4))
```

```
op <- par(cex = 0.8)

barplot(table(CM$grupo), col=terrain.colors(13), legend.text = paste(round((prop.table(table(CM$grupo)))*100,1),"% ", c("Sin AP", "Adenocarcinoma", "Carcinoma escamoso", "Carcinoma de célula pequeña", "Células claras", "Mieloma múltiple", "Carcinoma no célula pequeña", "Carcinoma ductal infiltrante", "Carcinoma lobulillar infiltrante", "Carcinoma urotelial", "Hepatocarcinoma", "Linfoma", "Otros")), args.legend = list(x = 3,5, bty = "n"), main = "Histología")

barplot(table(CM$valor_biomarcador),col = terrain.colors(2), legend.text = paste(round((prop.table(table(CM$valor_biomarcador)))*100,1),"%"), args.legend = list(x = 2,5, bty = "n"), main = "Biomarcador")

barplot(table(CM$CM_PrimarioConocido),col=terrain.colors(2), legend.text = paste(round((prop.table(table(CM$CM_PrimarioConocido)))*100,1),"% "), args.legend = list(x = 3,5, bty = "n"), main = "Tumor primario conocido")

barplot(table(CM$RT_descomprPrevia),col=terrain.colors(2), legend.text = paste(round((prop.table(table(CM$RT_descomprPrevia)))*100,1),"% "), args.legend = list(x = 3,5, bty = "n"), main = "RT descompresiva previa")
```

```{r fig.cap="Fig1_Episodios de CM por centro y grupo", echo=FALSE}
knitr::include_graphics("figures/fig1.jpeg")
```

```{r fig2 Genero edad supervivencia y estatus, message=FALSE, warning=FALSE, eval=FALSE}
#Repartimos el número de gráficos que vamos a visualizar
#layout(matrix(c(1,3,2,3), 2, 2, byrow = TRUE))
par(mfrow=c(2,2))

barplot(table(CM$genero),col=terrain.colors(2), legend.text = paste(round((prop.table(table(CM$genero)))*100,1),"%"), main = "Género", args.legend = list(x = 2.3,5, bty = "n"))

hist(CM$edad, breaks = 150, freq = TRUE, main = "Histograma de edad", col = "darkmagenta", xlab = "años", ylab = "número CM", xlim = c(30,110))

hist(CM$survival, breaks = 50, freq = TRUE, main = "Tiempo de supervivencia o censurado", col = "blue", xlab = "días", ylab = "número CM", xlim = c(0,4000))

barplot(table(CM$estatus_UC),col=terrain.colors(2), legend.text = paste(round((prop.table(table(CM$estatus_UC)))*100,1),"%"), main = "Estado del paciente", args.legend = list(x = 1.2,5, bty = "n"), names.arg = c("vivo", "muerto"))
```

En la siguiente figura observamos el histograma de edad, la proporción por géneros y como el histograma de la supervivencia en días.

```{r fig.cap="Fig2_Género, edad, supervivencia y estatus_UC", echo=FALSE}


```

```
knitr::include_graphics("figures/fig2.jpeg")
```

```{r fig3 Datos relativos a la CM, message=FALSE, warning=FALSE, eval=FALSE}
# Distribuimos los gráficos en una imagen
par(mfrow=c(2,2))
op <- par(cex = 0.8)

barplot(table(CM$localizacion),col=terrain.colors(5), legend.text = paste(round((prop.table(table(CM$localizacion)))*100,1),"%"), args.legend = list(x = "topright", bty = "n"), main = "Localización de la CM")

barplot(table(CM$tipo_CM),col=terrain.colors(3), legend.text = paste(round((prop.table(table(CM$tipo_CM)))*100,1),"%"), args.legend = list(x = "topright", bty = "n"), main = "Tipo de CM")

hist(CM$n_vert_afect, breaks = 40, main = "Histograma vertebras afectas", col = "green", xlab = "nº vértebras afectas por CM", ylab = "Frecuencia")

hist(CM$n_vert_comp, breaks = 40, main = "Histograma vértebras comprimidas", col = "brown", xlab = "nº vértebras comprimidas por CM", ylab = "Frecuencia")
```

```{r fig.cap="Fig3_Datos relativos a la CM", echo=FALSE}
knitr::include_graphics("figures/fig3.jpeg")
```
```{r fig4 Datos relativos al tumor, message=FALSE, warning=FALSE, eval=FALSE}
#par(mfrow=c(2,2))
layout(matrix(c(1,1,1,2,3,4), 2, 3, byrow = TRUE))
op <- par(cex = 0.9)

barplot(table(CM$histologia), col=terrain.colors(13), legend.text = paste(round((prop.table(table(CM$histologia)))*100,1),"% ", c("Sin AP", "Adenocarcinoma", "Carcinoma escamoso", "Carcinoma de célula pequeña", "Células claras", "Mieloma múltiple", "Carcinoma no célula pequeña", "Carcinoma ductal infiltrante", "Carcinoma lobulillar infiltrante", "Carcinoma urotelial", "Hepatocarcinoma", "Linfoma", "Otros")), args.legend = list(x = 14,7, bty = "n"), main = "Histología")

barplot(table(CM$valor_biomarcador),col = terrain.colors(2), legend.text = paste(round((prop.table(table(CM$valor_biomarcador)))*100,1),"%"), args.legend = list(x = 2,5, bty = "n"), main = "Biomarcador")

barplot(table(CM$CM_PrimarioConocido),col=terrain.colors(2), legend.text = paste(round((prop.table(table(CM$CM_PrimarioConocido)))*100,1),"% "), args.legend = list(x = 1,5, bty = "n"), main = "Tumor primario conocido")

barplot(table(CM$RT_descomprPrevia),col=terrain.colors(2), legend.text = paste(round((prop.table(table(CM$RT_descomprPrevia)))*100,1),"% "), args.legend = list(x = 2,5, bty = "n"), main = "RT descompresiva previa")
```

```{r fig.cap="Fig4_Datos relativos al tumor", echo=FALSE}
knitr::include_graphics("figures/fig4.jpeg")
```

```
```{r fig5 Datos relativos,message=FALSE, warning=FALSE, eval=FALSE}
# Distribuimos los gráficos en una imagen
layout(matrix(c(1,1,1,2,3,4), 2, 3, byrow = TRUE))
op <- par(cex = 0.8)

barplot(table(CM$clinica_debut),col = terrain.colors(8), legend.text = paste(round((prop.table(table(CM$clinica_debut)))*100,1),"%"), args.legend = list(x = 9,3, bty = "n"), main = "Début clínico")

barplot(table(CM$tto_trasCM),col= terrain.colors(4), legend.text = paste(round((prop.table(table(CM$tto_trasCM)))*100,1),"%"), args.legend = list(x = 6,5, bty = "n"), names.arg = NULL, main = "Tratamiento oncológico tras CM")

barplot(table(CM$estatusAmbulatorio7),col = terrain.colors(2), legend.text = paste(round((prop.table(table(CM$estatusAmbulatorio7)))*100,1),"%"), args.legend = list(x = 3,5, bty = "n"), main = "Estatus ambulatorio a los 7días")

barplot(table(CM$empeoramiento_neurologico),col = terrain.colors(2), legend.text = paste(round((prop.table(table(CM$empeoramiento_neurologico)))*100,1),"%"), args.legend = list(x = 3,5, bty = "n"), main = "Empeoramiento neurológico a los 7 días")
```

```{r fig.cap="Fig5_Datos relativos al tumor", echo=FALSE}
knitr::include_graphics("figures/fig5.jpeg")
```

```

```{r fig.cap="Fig6_Datos relativos al paciente pre y post tratamiento", echo=FALSE}
knitr::include_graphics("figures/fig6.jpeg")
```
En referencia a los índices tenemos una muestra dónde casi 2/3 estan clasificados como RADES 1 (entre 20 y 30 puntos), más de la mitad muestra un SINS catalogado como potencialmente inestable (con un 13% de NA) y en referencia a la escalera Blisky una mayoría de B y C (existe un 12% en este caso de NA).

```{r fig7, message=FALSE, warning=FALSE, eval=FALSE}
# calculamos el porcentaje del total que representa cada categoría en su variable respectiva. Usamos la opción count para contar las frecuencias de cada categoría
pct_rades <- round(count(CM$RADES)[1:3,2]/sum(count(CM$RADES)[1:3,2])*100)
pct_sins <- round(count(CM$SINS)[1:4,2]/sum(count(CM$SINS)[1:4,2])*100)

#Repartimos el número de gráficos que vamos a visualizar
layout(matrix(c(1,3,2,3), 2, 2, byrow = TRUE))

#ploteamos en gráfico quesito el % por centro y por grupo
pie3D(count(CM$RADES)[1:3,2],labels = paste(count(CM$RADES)[1:3,1], ": ", pct_rades,"%",sep=""), explode=0.3, main="RADES")
```

```
pie3D(count(CM$SINS)[1:4,2],labels = paste(count(CM$SINS)[1:4,1], ": ", pct_sins,"%",sep=""), explode=0.7, main="SINS")

barplot(table(CM$Blisky),col= rainbow(7), legend.text = paste(round((prop.table(table(CM$Blisky)))*100,1),"% ",levels(CM$Blisky)), main = "Escalera Blisky")
```

```{r fig.cap="Fig7_Escaleras pronósticas: RADES, SINS y Blisky", echo=FALSE}
knitr::include_graphics("figures/fig7.jpeg")
```
```{r}
describe(CM_num[,-1])[,c(1,3:5,7:9,12,18,21,25)]
describe(cleanCM_num[,-1])[,c(1,3:5,7:9,12,18,21,25)]
```
```{r fig8, message=FALSE, warning=FALSE}
# Función de panel diagonal: histograma y densidad > dgp.fn <- function(x,...) {
dgp.fn <- function(x,...) {
  par(new=TRUE)
  hist(x, col="lightblue", probability=TRUE, axes=FALSE, main="")
  lines(density(x), col="navy", lwd=2)
  rug(x)
}

# Función de panel no diagonal: diagrama dispersión y recta de regresión > pn.fn <- function(x,y,...){
pn.fn <- function(x,y,...){
  points(x,y)
  abline(lm(y~x), col="navy", lwd=2);
}
```

```{r fig8 fig8a, message=FALSE, warning=FALSE, eval=FALSE}
#Matriz de diagramas de dispersión
pairs(cleanCM_num[,2:10], panel=pn.fn,
            diag.panel=dgp.fn,
            label.pos=0.3,
            cex.labels=1.5)
ggpairs(cleanCM_num[,2:10])
```

```{r fig.cap="Fig8_Scatter bivariado con histograma de las distribuciones en la diagonal y ajuste por regresión líneal usando las variables numéricas de la base de datos", echo=FALSE}
knitr::include_graphics("figures/fig8.jpeg")
knitr::include_graphics("figures/fig8a.jpeg")
```

```{r fig 8b Coeficiente de correlación entre las variables}
plot_correlate(cleanCM_num[,-1])
```

```{r message=FALSE, warning=FALSE,eval=FALSE}
cleanCM_T$genero <- factor(cleanCM_T$genero, labels = c("hombre","mujer"))
cleanCM_T$centro <- factor(cleanCM_T$centro, labels = c("ICO Badalona", "ICO Hospitalet"))
cleanCM_T$grupo <- factor(cleanCM_T$grupo, labels = c("<2018", ">=2018"))
cleanCM_T$opcion_terapeutica <- factor(cleanCM_T$opcion_terapeutica, labels = c("CR", "CR+RT", "RT+CR", "RT", "SOPORTE"))
cleanCM_T$localizacion <- factor(cleanCM_T$localizacion, labels = c("cervical", "dorsal", "lumbar", "sacra", "múltiple"))
cleanCM_T$tipo_CM <- factor(cleanCM_T$tipo_CM, labels = c("medular", "radicular", "intramedular"))
cleanCM_T$vert_afect <- factor(cleanCM_T$vert_afect, labels = c("1","1-5",">5"))
cleanCM_T$vert_comp <- factor(cleanCM_T$vert_comp, labels = c("0","1",">1"))
cleanCM_T$CM_PrimarioConocido <- factor(cleanCM_T$CM_PrimarioConocido, labels = c("no", "si"))
cleanCM_T$valor_biomarcador <- factor(cleanCM_T$valor_biomarcador, labels = c("negativo/no disponible", "positivo"))
cleanCM_T$RADES <- factor(cleanCM_T$RADES, labels = c("Grupo1", "Grupo2", "Grupo3"))
cleanCM_T$SINS <- factor(cleanCM_T$SINS, labels = c("Estable", "Potencialmente inestable", "inestable"))
cleanCM_T$Blisky <- factor(cleanCM_T$Blisky, labels = c("0", "A1", "A2", "A3", "B", "C"))
cleanCM_T$RT_descomprPrevia <- factor(cleanCM_T$RT_descomprPrevia, labels = c("no", "si"))
cleanCM_T$empeoramiento_neurologico <- factor(cleanCM_T$empeoramiento_neurologico, labels = c("no", "si"))
cleanCM_T$tto_trasCM <- factor(cleanCM_T$tto_trasCM, labels = c("no inicia", "no continua", "inicia", "continua"))
cleanCM_T$estatusAmbulatorio7 <- factor(cleanCM_T$estatusAmbulatorio7, labels =c("no", "si"))
cleanCM_T$estatus_lesion_UC <- factor(cleanCM_T$estatus_lesion_UC, labels = c("estable","progresión en PTV","progresión fuera de PTV","RC radiológica","RC metabólica", "No evaluada"))
cleanCM_T$clinica_debut <- factor(cleanCM_T$clinica_debut, labels = c("asintomático","sensitivo/neuropático","somático","motor","somático+motor","sens/neurop+somático","sens/neurop+motor","todos"))
cleanCM_T$dosis_RT <- factor(cleanCM_T$dosis_RT, labels = c("8Gyx1s","5Gyx4s","4Gyx5s","3Gyx10s","Otros"))
cleanCM_T$metOseas_RADES <- as.factor(cleanCM_T$metOseas_RADES)
cleanCM_T$metVisc_RADES <- as.factor(cleanCM_T$metVisc_RADES)
cleanCM_T$tumPrimario_RADES <- as.factor(cleanCM_T$tumPrimario_RADES )
cleanCM_T$intervalo_diagCM_RADES <- as.factor(cleanCM_T$intervalo_diagCM_RADES)
cleanCM_T$estAmbulatorio_RADES <- as.factor(cleanCM_T$estAmbulatorio_RADES)
cleanCM_T$tiempoDesarrollo_DefMotor_RADES <- as.factor(cleanCM_T$tiempoDesarrollo_DefMotor_RADES)
cleanCM_T$location_SINS <- as.factor(cleanCM_T$location_SINS)
cleanCM_T$dolor_SINS <- as.factor(cleanCM_T$dolor_SINS)
cleanCM_T$caract_lesion_SINS <- as.factor(cleanCM_T$caract_lesion_SINS)
cleanCM_T$alineacColumna_SINS <- as.factor(cleanCM_T$alineacColumna_SINS)
```

```
cleanCM_T$colapsoVertebral_SINS <- as.factor(cleanCM_T$colapsoVertebral_SINS)
cleanCM_T$afectPosterolat_SINS <- as.factor(cleanCM_T$afectPosterolat_SINS)
cleanCM_T$histologia <- as.factor(cleanCM_T$histologia)
cleanCM_T$fijacionPrevia_SegmIrradiar <- as.factor(cleanCM_T$fijacionPrevia_SegmIrradiar)
cleanCM_T$test_diagnostico <- factor(cleanCM_T$test_diagnostico, labels = c("RMN","TC", "otros"))
cleanCM_T$estatusAmbulatorio7 <- factor(cleanCM_T$estatusAmbulatorio7, labels = c("no","si"))
```

En cuanto a las variables categóricas es interesante estudiar la supervivencia mediante una caja de dispersión en función de ellas.

```{r fig9, message=FALSE, warning=FALSE, eval=FALSE}
#par(mar=c(1,1,1,1))
par(mfrow=c(2,2))
boxplot(cleanCM_T$survival ~ cleanCM_T$centro , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "centro")
boxplot(cleanCM_T$survival ~ cleanCM_T$genero , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "genero")
boxplot(cleanCM_T$survival ~ cleanCM_T$grupo , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "grupo")
boxplot(cleanCM_T$survival ~ cleanCM_T$CM_PrimarioConocido , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "Primario Conocido")
```

```{r fig9a, message=FALSE, eval=FALSE}
boxplot(cleanCM_T$survival ~ cleanCM_T$valor_biomarcador , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "Biomarcador")
boxplot(cleanCM_T$survival ~ cleanCM_T$empeoramiento_neurologico , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "Empeoramiento neurológico")
boxplot(cleanCM_T$survival ~ cleanCM_T$estatusAmbulatorio7 , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "Estatus ambulatorio a 7 días de CM")
boxplot(cleanCM_T$survival ~ cleanCM_T$tto_trasCM , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "tratamiento oncológico tras CM")
```

```{r fig.cap="Fig9/9a_Caja de dispersión de la supervivencia en función de las distintas variables categóricas I", echo=FALSE}
knitr::include_graphics("figures/fig9.jpeg")
knitr::include_graphics("figures/fig9a.jpeg")
```

Este análisis deberá realizarse tras el tratamiento de los outliers ya que escalan de tal manera el gráfico que no puede observarse con claridad.

```{r fig 10, eval=FALSE, message=FALSE, warning=FALSE}
par(mfrow=c(2,2))
```

```
boxplot(cleanCM_T$survival ~ cleanCM_T$localizacion , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "localización")
boxplot(cleanCM_T$survival ~ cleanCM_T$tipo_CM, col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "tipo de CM")
boxplot(cleanCM_T$survival ~ cleanCM_T$vert_afect , col = "#fffbb2", varwidth = TRUE, notch = TRUE, nn = TRUE, ylab = "supervivencia", xlab = "vértebras afectas")
boxplot(cleanCM_T$survival ~ cleanCM_T$vert_comp , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "vértebras comprimidas")
```

```{r fig 10a, eval=FALSE, message=FALSE, warning=FALSE}
par(mfrow=c(2,2))
boxplot(cleanCM_T$survival ~ cleanCM_T$RADES , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "localización")
boxplot(cleanCM_T$survival ~ cleanCM_T$SINS , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "localización")
boxplot(cleanCM_T$survival ~ cleanCM_T$Blisky , col = "#fffbb2", varwidth = TRUE, notch = FALSE, ann = TRUE, ylab = "supervivencia", xlab = "localización")
boxplot(cleanCM_T$survival ~ cleanCM_T$RT_descomprPrevia , col = "#fffbb2", varwidth = TRUE, notch = TRUE, ann = TRUE, ylab = "supervivencia", xlab = "localización")
```

```{r fig.cap="Fig10/10a_Caja de dispersión de la supervivencia en función de las distintas variables categóricas II", echo=FALSE}
knitr::include_graphics("figures/fig10.jpeg")
knitr::include_graphics("figures/fig10a.jpeg")
```

````{r fig11, message=FALSE, warning=FALSE, eval=FALSE}
par(mfrow=c(3,3))
boxplot(CM$survival, col = "#fcd1bf", main = "Boxplot supervivencia", varwidth = TRUE, notch = TRUE)
boxplot(CM$edad, col = "#fcd1bf", main = "Boxplot edad", varwidth = TRUE, notch = TRUE)
boxplot(CM$n_vert_afect, col = "#fcd1bf", main = "Boxplot vértebras afectas", varwidth = TRUE, notch = TRUE)
boxplot(CM$n_vert_comp, col = "#fcd1bf", main = "Boxplot vértebras comprimidas", varwidth = TRUE, notch = TRUE)
boxplot(CM$puntuacion_RADES, col = "#fcd1bf", main = "Boxplot puntuación RADES", varwidth = TRUE, notch = TRUE)
boxplot(CM$puntuacion_SINS, col = "#fcd1bf", main = "Boxplot puntuación SINS", varwidth = TRUE, notch = TRUE)
boxplot(CM$t_sc, col = "#fcd1bf", main = "Boxplot tiempo sospecha clínica a diagnóstico", varwidth = TRUE, notch = FALSE)
boxplot(CM$t_ac, col = "#fcd1bf", main = "Boxplot tiempo diagnóstico a aviso de compresión", varwidth = TRUE, notch = FALSE)
boxplot(CM$t_itto, col = "#fcd1bf", main = "Boxplot tiempo diagnóstico a inicio de tratamiento", varwidth = TRUE, notch = TRUE)
```

```{r fig.cap="Fig11_Cajas de dispersión de las distintas variables numéricas", echo=FALSE}
knitr::include_graphics("figures/fig11.jpeg")
```
```{r}
summary(cleanCM_num)
apply(cleanCM_num, 2, sd)
normality(cleanCM_num[,-1])
```

```{r fig12, message=FALSE, warning=FALSE}
par(mfrow=c(3,3))
library("car")
qqPlot(cleanCM_num$edad, main = "edad")
qqPlot(cleanCM_num$puntuacion_RADES, main = "Puntuación RADES")
qqPlot(cleanCM_num$puntuacion_SINS, main = "Puntuación SINS")
qqPlot(cleanCM_num$t_sc, main = "tiempo sospecha clínica")
qqPlot(cleanCM_num$t_ac, main = "tiempo aviso CM")
qqPlot(cleanCM_num$t_itto, main = "tiempo inicio tratamiento")
qqPlot(cleanCM_num$survival, main = "supervivencia")
qqPlot(cleanCM_num$n_vert_afect, main = "vértebras afectas")
qqPlot(cleanCM_num$n_vert_comp, main = "vértebras comprimidas")
```
```{r}
apply(cleanCM_num,2,shapiro.test)
apply(cleanCM_num, 2, skewness)
```
```{r fig.cap="Fig13_Valores de lambda y potencias correspondiente de la escalera de Tukey", echo=FALSE}
knitr::include_graphics("figures/fig13.JPG")
```

```{r}
transformTukey(cleanCM_num$survival, quiet = TRUE, verbose = FALSE, plotit = FALSE, returnLambda = TRUE)
transformTukey(cleanCM_num$puntuacion_RADES, plotit = FALSE, returnLambda = TRUE, quiet = TRUE,)
transformTukey(cleanCM_num$puntuacion_SINS, plotit = FALSE, returnLambda = TRUE, quiet = TRUE,)
transformTukey(cleanCM_num$n_vert_afect, plotit = FALSE, returnLambda = TRUE, quiet = TRUE,)
transformTukey(cleanCM_num$n_vert_comp, plotit = FALSE, returnLambda = TRUE, quiet = TRUE,)
transformTukey(cleanCM_num$t_sc, plotit = FALSE, returnLambda = TRUE, quiet = TRUE,)
transformTukey(cleanCM_num$t_ac, plotit = FALSE, returnLambda = TRUE, quiet = TRUE,)
transformTukey(cleanCM_num$t_itto, plotit = FALSE, returnLambda = TRUE, quiet = TRUE)
```
```{r}
names (cleanCM_numT)[1] = "ID"
```

```
names (cleanCM_numT)[3] = "edad"
names (cleanCM_numT)[4] = "puntuacion_RADES"
names (cleanCM_numT)[5] = "puntuacion_SINS"
```

```{r}
normality(cleanCM_numT)
```

```{r fig14, warning=FALSE, message=FALSE, eval=FALSE}
par(mfrow=c(3,3))
library(car)
qqPlot(cleanCM_numT$edad, main = "edad", pch = 0.2, ylab = "cuantiles edad", xlab = "cuantiles normal")
qqPlot(cleanCM_numT$puntuacion_RADES, main = "Puntuación RADES", pch = 0.2, ylab = "cuantiles RADES", xlab = "cuantiles normal")
qqPlot(cleanCM_numT$puntuacion_SINS, main = "Puntuación SINS", pch = 0.2, ylab = "cuantiles SINS", xlab = "cuantiles normal"))
qqPlot(cleanCM_numT$sqrt.t_sc, main = "tiempo sospecha clínica", pch = 0.2, ylab = "cuantiles edad", xlab = "cuantiles normal"))
qqPlot(cleanCM_numT$sqrt.t_ac, main = "tiempo aviso CM", pch = 0.2, ylab = "cuantiles edad", xlab = "cuantiles normal"))
qqPlot(cleanCM_numT$sqrt.t_itto, main = "tiempo inicio tratamiento", pch = 0.2, ylab = "cuantiles edad", xlab = "cuantiles normal"))
qqPlot(cleanCM_numT$log.survival, main = "supervivencia")
qqPlot(cleanCM_numT$log.n_vert_afect, main = "vértebras afectas")
qqPlot(cleanCM_numT$sqrt.n_vert_comp, main = "vértebras comprimidas")
```

```{r}
describe(cleanCM_num[,-1])[,c(3:5,7:9,12,18,21,25)]
apply(cleanCM_numT,2,shapiro.test)
```

```{r fig.cap="Fig14_qqPlot para las distintas variables numéricas transformadas según la escalera de potencias de Tukey", echo=FALSE}
knitr::include_graphics("figures/fig14.jpeg")
```

```{r fig15, warning=FALSE, message=FALSE, eval=FALSE}
pairs(cleanCM_numT, panel=pn.fn,
      diag.panel=dgp.fn,
      label.pos=0.3,
      cex.labels=1.5)
```

```{r fig16, warning=FALSE, message=FALSE, eval=FALSE}
ggpairs(cleanCM_numT)
```

```{r fig16a Correlación de las variables}
plot_correlate(cleanCM_numT[,-1])
```

```
```{r fig.cap="Fig15_Gráfico de dispersión con histograma en la diagonal y recta de regresión lineal", echo=FALSE}
knitr::include_graphics("figures/fig15.jpeg")
knitr::include_graphics("figures/fig16.jpeg")
```
```{r}
# Miramos cuantos valores NA tenemos en la base de datos en total y qué porcentaje significa en la globalidad de los datos
sum(is.na(CM))
mean(is.na(CM))
# Miramos cuantos valores NA tenemos en la base de datos que contiene sólo las variables numéricas y qué porcentaje significa en la globalidad de los datos
sum(is.na(CM_num))
mean(is.na(CM_num))
```
```{r warning=FALSE}
na_count<- round(sapply(CM, function(y)
sum(length(which(is.na(y))))/sum(length(y)))*100,1)
na_count
mean(is.na(cleanCM))
```
```{r fig.cap="Fig11_Cajas de dispersión de las distintas variables numéricas", echo=FALSE}
knitr::include_graphics("figures/fig11.jpeg")
```

```{r fig11a boxplot de las variables numéricas transformadas, message=FALSE, warning=FALSE}
par(mfrow=c(3,3))
boxplot(cleanCM_numT$edad , col = "#fcd1bf", main = "Boxplot edad", varwidth = TRUE, notch = TRUE)
boxplot(cleanCM_numT$log.survival, col = "#fcd1bf", main = "Boxplot supervivencia", varwidth = TRUE, notch = TRUE)
boxplot(cleanCM_numT$log.n_vert_afect, col = "#fcd1bf", main = "Boxplot vértebras afectas", varwidth = TRUE, notch = TRUE)

boxplot(cleanCM_numT$sqrt.n_vert_comp, col = "#fcd1bf", main = "Boxplot vértebras comprimidas", varwidth = TRUE, notch = TRUE)
mtext(paste("Outliers: ", paste(boxplot.stats(cleanCM_numT$sqrt.n_vert_comp)$out, collapse = ", ")))

boxplot(cleanCM_numT$puntuacion_RADES, col = "#fcd1bf", main = "Boxplot puntuación RADES", varwidth = TRUE, notch = TRUE)

boxplot(cleanCM_numT$puntuacion_SINS, col = "#fcd1bf", main = "Boxplot puntuación SINS", varwidth = TRUE, notch = TRUE)
mtext(paste("Outliers: ",
paste(boxplot.stats(cleanCM_numT$cleanCM_num.puntuacion_SINS)$out, collapse = ", ")))
```
```

```
boxplot(cleanCM_numT$sqrt.t_sc, col = "#fcd1bf", main = "Boxplot tiempo des de la sospecha clínica al diagnóstico", varwidth = TRUE, notch = FALSE)
mtext(paste("Outliers: ", paste(boxplot.stats(cleanCM_numT$sqrt.t_sc)$out, collapse = ", ")))

boxplot(cleanCM_numT$sqrt.t_ac, col = "#fcd1bf", main = "Boxplot tiempo des del diagnóstico al aviso de compresión", varwidth = TRUE, notch = FALSE)
mtext(paste("Outliers: ", paste(boxplot.stats(cleanCM_numT$sqrt.t_ac)$out, collapse = ", ")))

boxplot(cleanCM_numT$sqrt.t_itto, col = "#fcd1bf", main = "Boxplot tiempo des del diagnóstico al inicio de tratamiento", varwidth = TRUE, notch = TRUE)
mtext(paste("Outliers: ", paste(boxplot.stats(cleanCM_numT$sqrt.t_itto)$out, collapse = ", ")))
```
```{r fig.cap="Fig11a_Cajas de dispersión de las distintas variables numéricas transformadas con el valor de los outliers impreso", echo=FALSE}
knitr::include_graphics("figures/fig11a.jpeg")
```
```{r}
# "boxplot.stats(cleanCM_num$survival)" nos da una série de estadísticas de las cuales la función outl corresponde el valor del outlier. Com la función which encontramos la hilera dónde se produce ese valor
cleanCM_num$ID[outSurv]
cleanCM_numT$cleanCM_num.ID[outSurvT]

cleanCM_num$ID[outVertAf]
cleanCM_numT$cleanCM_num.ID[outVertAfT]

cleanCM_num$ID[outVertC]
cleanCM_numT$cleanCM_num.ID[outVertCT]

cleanCM_num$ID[outSINS]

cleanCM_num$ID[outt_sc]
cleanCM_numT$cleanCM_num.ID[outt_scT]

cleanCM_num$ID[outt_ac]
cleanCM_numT$cleanCM_num.ID[outt_acT]

cleanCM_num$ID[outt_itto]
cleanCM_numT$cleanCM_num.ID[outt_ittoT]
```
```{r}
shapiro.test(cleanCM_numT$sqrt.n_vert_comp[-c(which(cleanCM_numT$sqrt.n_vert_comp %in% c(boxplot.stats(cleanCM_numT$sqrt.n_vert_comp)$out)))])

shapiro.test(cleanCM_numT$sqrt.t_sc[-c(which(cleanCM_numT$sqrt.t_sc %in% c(boxplot.stats(cleanCM_numT$sqrt.t_sc)$out)))])
```

```
shapiro.test(cleanCM_numT$sqrt.t_ac[-c(which(cleanCM_numT$sqrt.t_ac %in% c(boxplot.stats(cleanCM_numT$sqrt.t_ac)$out)))])

shapiro.test(cleanCM_numT$sqrt.t_itto[-c(which(cleanCM_numT$sqrt.t_itto %in% c(boxplot.stats(cleanCM_numT$sqrt.t_itto)$out)))])
```

```{r}
# comparamos
gf = goodfit(cleanCM_numT$sqrt.t_itto,type= "poisson",method= "ML")
gf1 = goodfit(cleanCM_num$t_itto,type= "poisson",method= "ML")
# miramos el p-valor
gf.summary = capture.output(summary(gf))[[5]]
pvalue = unlist(strsplit(gf.summary, split = " "))
pvalue = as.numeric(pvalue[length(pvalue)]); pvalue

gf1.summary = capture.output(summary(gf1))[[5]]
pvalue1 = unlist(strsplit(gf1.summary, split = " "))
pvalue1 = as.numeric(pvalue1[length(pvalue1)]); pvalue1
```

```{r fig11b boxplot de las variables numéricas transformadas sin ouliers, message=FALSE, warning=FALSE}
par(mfrow=c(3,3))
boxplot(cleanCM_Tout$edad , col = "#fcd1bf", main = "Boxplot edad", varwidth = TRUE, notch = TRUE)
boxplot(cleanCM_Tout$survival, col = "#fcd1bf", main = "Boxplot supervivencia", varwidth = TRUE, notch = TRUE)
boxplot(cleanCM_Tout$vert_afect, col = "#fcd1bf", main = "Boxplot vértebras afectas", varwidth = TRUE, notch = TRUE)
boxplot(cleanCM_Tout$vert_comp, col = "#fcd1bf", main = "Boxplot vértebras comprimidas", varwidth = TRUE, notch = TRUE)
boxplot(cleanCM_Tout$puntuacion_RADES, col = "#fcd1bf", main = "Boxplot puntuación RADES", varwidth = TRUE, notch = TRUE)
boxplot(cleanCM_Tout$puntuacion_SINS, col = "#fcd1bf", main = "Boxplot puntuación SINS", varwidth = TRUE, notch = TRUE)

boxplot(cleanCM_Tout$t_sc, col = "#fcd1bf", main = "Boxplot tiempo des de la sospecha clínica al diagnóstico", varwidth = TRUE, notch = FALSE)
boxplot(cleanCM_Tout$t_ac, col = "#fcd1bf", main = "Boxplot tiempo des del diagnóstico al aviso de compresión", varwidth = TRUE, notch = FALSE)

boxplot(cleanCM_Tout$t_itto, col = "#fcd1bf", main = "Boxplot tiempo des del diagnóstico al inicio de tratamiento", varwidth = TRUE, notch = TRUE)
```

```{r fig.cap="Fig11b_Cajas de dispersión de las distintas variables numéricas transformadas y con los outliers erradicado", echo=FALSE}
knitr::include_graphics("figures/fig11b.jpeg")
```