



UNIVERSIDADE DA CORUÑA

Análisis de Supervivencia en Pacientes con Cáncer de Mama: Predicción de Cardiotoxicidad

Ajuste de un Modelo de Regresión para la Predicción de Cardiotoxicidad en Pacientes con Cáncer de Mama



Autor: Pablo Torreira Pardo

A Coruña, 9 de diciembre 2024

Índice

Índice de figuras	II
Índice de tablas	II
1. Introducción	1
2. Objetivos	1
3. Material y Métodos	2
3.1. Material	2
3.1.1. Datos	2
3.1.2. Software y bibliotecas	2
3.2. Métodos	3
3.2.1. Análisis exploratorio	3
3.2.2. Análisis de supervivencia	3
3.2.3. Selección de variables y validación del modelo	5
3.2.4. Representación gráfica	6
4. Metodología	7
4.1. Preparación de los datos	7
4.2. Significancia estadística	7
4.3. Análisis exploratorio	7
4.4. Análisis de supervivencia	7
4.5. Selección y validación del modelo	8
4.6. Representación gráfica	8
5. Resultados	9
5.1. Exploración inicial de los datos	9
5.1.1. Distribución de las variables	9
5.1.2. Correlación entre variables	10
5.2. Análisis de supervivencia	11
5.2.1. Modelos univariados	12
5.2.2. Modelos multivariados	14
5.2.3. Comprobación de Hipótesis Estructurales	14
6. Conclusiones	15

Bibliografía

Índice de figuras

1.	Histogramas de las variables cuantitativas	9
2.	Gráficos circulares de las frecuencias de CTRCD para los niveles de AC	10
3.	Diagramas de cajas de las variables cuantitativas	10
4.	Representación gráfica de matriz de correlaciones	11
5.	Gráfico de función de supervivencia	11
6.	Gráficos de AC por grupos	12
7.	Gráficos de heart_rate estratificado por la mediana	12
8.	Gráficos de LVEF estratificado por la mediana	13
9.	Gráficos de PWT estratificado por la mediana	13
10.	Graficos de la estimación de los coeficientes $\beta(t)$	14
11.	Graficos para detección de valores atípicos	15

Índice de tablas

1.	Medidas descriptivas para las variables cuantitativas	9
2.	Resumen de los modelos univariados para PWT, LAd, LVDd y LVSd	14
3.	Parámetros del modelo multivariado final	14

1 Introducción

La cardiotoxicidad es un efecto adverso asociado al uso de múltiples tratamientos oncológicos, caracterizado por un impacto negativo en la función cardíaca que puede manifestarse como disfunción sistólica, insuficiencia cardíaca u otras alteraciones cardiovasculares. Entre las herramientas para abordar este problema, los modelos predictivos juegan un papel crucial al identificar factores de riesgo y guiar intervenciones preventivas.

Este estudio emplea un conjunto de datos recopilado por el Complejo Hospitalario Universitario de A Coruña, que incluye información detallada de 531 pacientes. Las variables estudiadas abarcan características clínicas como la fracción de eyección ventricular izquierda, la frecuencia cardíaca, el grosor de la pared posterior ventricular, y diámetros auriculares y ventriculares, así como el historial de tratamiento con antraciclinas. Estos datos ofrecen una base robusta para modelar y predecir el desarrollo de cardiotoxicidad, con especial énfasis en la relación temporal entre las variables y los eventos cardíacos.

El análisis se centra en el ajuste de modelos de regresión y análisis de supervivencia, con el objetivo de identificar las variables más influyentes y reducir la complejidad del modelo sin perder capacidad predictiva. Este enfoque proporciona una herramienta analítica que contribuye a optimizar la gestión clínica y a mejorar el entendimiento de los efectos cardiotóxicos en pacientes oncológicos.

2 Objetivos

El objetivo principal de este estudio es analizar los factores clínicos que influyen en el desarrollo de cardiotoxicidad (CTRCD) en pacientes con cáncer de mama utilizando un modelo de análisis de supervivencia basado en la Regresión de Cox. Este método se ha elegido porque permite manejar datos censurados, es decir, casos en los que no todos los pacientes desarrollan cardiotoxicidad durante el periodo de estudio. Con este modelo, se podrá estudiar cómo variables como la fracción de eyección ventricular izquierda (LVEF), frecuencia cardíaca (heart_rate), grosor de la pared posterior ventricular (PWT), diámetros auriculares y ventriculares (LAd, LVDd y LVSd), y haber recibido tratamiento con antraciclinas (AC) influyen en el tiempo hasta la aparición de cardiotoxicidad.

El análisis también tiene como objetivo optimizar el modelo, identificando las variables más relevantes y reduciendo la complejidad sin perder precisión predictiva. Además, se calcularán intervalos de confianza para los resultados y se validará el modelo para asegurar que sea una herramienta fiable y útil en el contexto clínico.

3 Material y Métodos

3.1. Material

3.1.1. Datos

El presente estudio se basa en un conjunto de datos proporcionado por el Complejo Hospitalario Universitario de A Coruña (CHUAC), recopilado entre 2007 y 2021. Este conjunto de datos incluye información clínica y funcional de 531 pacientes con cáncer de mama sometidas a tratamientos oncológicos con un alto potencial de producir cardiotoxicidad. Las variables analizadas son:

- **CTRCD**: Indicador binario del desarrollo de cardiotoxicidad (1 = sí, 0 = no).
- **time**: Tiempo, en días, desde el inicio del tratamiento hasta el evento o final del seguimiento.
- **LVEF**: Fracción de eyección ventricular izquierda (%).
- **heart_rate**: Frecuencia cardíaca (lpm).
- **PWT**: Grosor de la pared posterior ventricular (cm).
- **LAd**: Diámetro de la aurícula izquierda (cm).
- **LVDd**: Diámetro diastólico del ventrículo izquierdo (cm).
- **LVSD**: Diámetro sistólico del ventrículo izquierdo (cm).
- **AC**: Uso de antraciclina (1 = sí, 0 = no).

Los datos fueron procesados para eliminar valores faltantes y garantizar la consistencia de las variables, resultando en un conjunto de datos sin valores ausentes y con un total de 430 casos sin cardiotoxicidad y 47 con cardiotoxicidad.

3.1.2. Software y bibliotecas

El análisis estadístico fue realizado en el entorno de programación **RStudio** [6] (versión 4.3.0). Para la implementación de los modelos y análisis de supervivencia, se utilizaron los siguientes paquetes:

- **survival** [7]: Para ajustar modelos de regresión de Cox y análisis de supervivencia.
- **survminer** [5]: Para la visualización de curvas de supervivencia.
- **car** [4]: Para evaluar multicolinealidad mediante el cálculo del factor de inflación de la varianza (VIF).
- **corrplot** [8]: Para representar gráficamente la matriz de correlación.
- **RcmdrMisc** [3]: Para el uso de la función numSummary().

Además, se emplearon herramientas gráficas para evaluar la distribución de las variables y detectar posibles valores atípicos.

3.2. Métodos

3.2.1. Análisis exploratorio

El análisis inicial incluyó la evaluación de la distribución y relaciones de las variables mediante herramientas gráficas y estadísticas. Se generaron histogramas para observar la distribución de las variables cuantitativas y diagramas de caja para detectar valores atípicos. La matriz de correlación se utilizó para medir la fuerza y dirección de las relaciones lineales entre las variables cuantitativas.

Matriz de correlación Cada elemento C_{ij} en la matriz de correlación representa el coeficiente de correlación de Pearson:

$$C_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sigma_{X_i} \sigma_{X_j}},$$

donde:

- X_i, X_j : variables cuantitativas,
- $\text{Cov}(X_i, X_j)$: covarianza entre X_i y X_j ,
- $\sigma_{X_i}, \sigma_{X_j}$: desviaciones estándar de X_i y X_j .

Interpretación:

- $C_{ij} > 0$: relación positiva (a mayor X_i , mayor X_j).
- $C_{ij} < 0$: relación negativa (a mayor X_i , menor X_j).
- $|C_{ij}| \rightarrow 1$: fuerte relación lineal.
- $|C_{ij}| \rightarrow 0$: relación lineal débil o inexistente.

La matriz de correlación permite identificar posibles problemas de multicolinealidad, que ocurren cuando las variables explicativas están altamente correlacionadas entre sí. Este fenómeno puede inflar los errores estándar de los coeficientes en los modelos de regresión, reduciendo la precisión de las estimaciones.

3.2.2. Análisis de supervivencia

El análisis de supervivencia se centra en estudiar el tiempo hasta la ocurrencia de un evento de interés, en este caso, el desarrollo de cardiotoxicidad (CTRC). Este enfoque es particularmente útil debido a la presencia de datos censurados [1], es decir, pacientes que no desarrollaron el evento durante el periodo de seguimiento.

Función de supervivencia La función de supervivencia $S(t)$ mide la probabilidad de que un individuo sobreviva más allá de un tiempo t :

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

En términos de la densidad de tiempos de fallo $f(t)$ y la función de riesgo acumulada $H(t)$, se tiene:

$$S(t) = e^{-H(t)}$$

Interpretación:

- $S(t)$ próximo a 1: alta probabilidad de que no ocurra el evento hasta t .
- $S(t)$ próximo a 0: alta probabilidad de que el evento ocurra antes de t .

Modelo de riesgos proporcionales de Cox [2] El modelo ajusta la relación entre las variables explicativas y la función de riesgo $h(t|X)$:

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

donde $h_0(t)$ es la función de riesgo basal. **Interpretación de β_j :** El coeficiente β_j asociado a la variable X_j mide el cambio logarítmico en el riesgo por unidad de incremento en X_j :

$$e^{\beta_j} = \text{HR}_j,$$

donde HR_j (razón de riesgos o *Hazard Ratio*) representa el multiplicador del riesgo asociado a X_j :

- $\text{HR}_j > 1$: incremento en X_j aumenta el riesgo.
- $\text{HR}_j < 1$: incremento en X_j reduce el riesgo.
- $\text{HR}_j = 1$: X_j no tiene efecto sobre el riesgo.

Funciones de riesgo La función de riesgo, $h(t)$, mide la probabilidad instantánea de que un individuo experimente el evento en un momento dado t , dado que ha sobrevivido hasta ese momento. Matemáticamente, se define como:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

Prueba log-rank La prueba log-rank se utiliza para comparar las curvas de supervivencia entre dos o más grupos, evaluando si las diferencias en supervivencia son estadísticamente significativas. La hipótesis nula (H_0) asume que las curvas de supervivencia son iguales entre los grupos.

Para cada intervalo de tiempo t_i , se utilizan las siguientes definiciones:

- $O_{i,j}$: Número de eventos observados en el grupo j en el tiempo t_i .
- $E_{i,j}$: Número de eventos esperados en el grupo j en t_i , calculado como:

$$E_{i,j} = d_i \cdot \frac{n_{i,j}}{\sum_j n_{i,j}}$$

donde d_i es el número total de eventos ocurridos en t_i , y $n_{i,j}$ es el número de individuos en riesgo en el grupo j justo antes de t_i .

- $O - E$: Diferencia entre eventos observados y esperados.

El estadístico log-rank se calcula como:

$$\chi^2 = \frac{(\sum_i (O_{i,j} - E_{i,j}))^2}{\text{Var}(O_{i,j})}$$

donde la varianza se calcula como:

$$\text{Var}(O_{i,j}) = \sum_i \left(d_i \cdot \frac{n_{i,j}}{\sum_j n_{i,j}} \cdot \left(1 - \frac{n_{i,j}}{\sum_j n_{i,j}} \right) \right)$$

Distribución e interpretación El estadístico χ^2 sigue una distribución chi-cuadrado con $g-1$ grados de libertad, siendo g el número de grupos. Si el p-valor asociado es menor que el nivel de significancia (α), se rechaza la hipótesis nula, indicando diferencias significativas entre las curvas de supervivencia.

3.2.3. Selección de variables y validación del modelo

Factor de inflación de la varianza (VIF) Para evaluar la redundancia entre variables explicativas, se utilizó el VIF:

$$\text{VIF}(X_j) = \frac{1}{1 - R_j^2},$$

donde R_j^2 es el coeficiente de determinación al predecir X_j como función de las demás variables.

Interpretación:

- $\text{VIF} = 1$: no hay correlación con otras variables.
- $1 < \text{VIF} \leq 5$: correlación moderada.
- $\text{VIF} > 10$: alta correlación, posible multicolinealidad.

Índice de concordancia (C-index) El C-index mide la capacidad del modelo para discriminar correctamente entre los tiempos de fallo:

$$C = \frac{\sum_{i < j} I(T_i < T_j) \cdot I(\hat{h}_i > \hat{h}_j)}{\sum_{i < j} I(T_i < T_j)},$$

donde $I(\cdot)$ es la función indicadora. **Interpretación:**

- $C = 0,5$: el modelo no tiene capacidad predictiva (aleatorio).
- $C = 1$: predicción perfecta.
- Valores entre 0.7 y 0.8 indican buena discriminación.

Test de Schoenfeld Se utilizó el test de Schoenfeld para verificar el supuesto de riesgos proporcionales. El supuesto de riesgos proporcionales implica que el cociente entre las funciones de riesgo de dos individuos i y j es constante en el tiempo:

$$\frac{h_i(t)}{h_j(t)} = \exp((\mathbf{x}_i - \mathbf{x}_j)\boldsymbol{\beta}).$$

El test de Schoenfeld verifica este supuesto evaluando los residuos de Schoenfeld, que se calculan como:

$$r_k = z_k - \bar{z}(t_k),$$

donde:

- z_k son los valores observados de las covariables para el individuo que experimentó el evento en el tiempo t_k ,
- $\bar{z}(t_k)$ es el promedio ponderado de las covariables de todos los individuos en riesgo en t_k .

Bajo el supuesto de riesgos proporcionales, estos residuos deben ser independientes del tiempo. El test evalúa esta independencia ajustando una regresión de los residuos sobre el tiempo. Si los residuos presentan una tendencia significativa en función del tiempo, se rechaza el supuesto de riesgos proporcionales.

En este estudio, se utilizó el test de Schoenfeld para garantizar que las estimaciones del modelo de Cox fueran válidas y consistentes a lo largo del tiempo.

3.2.4. Representación gráfica

Con el objetivo de realizar un análisis exploratorio, se graficaron diagramas de cajas, histogramas para las distribuciones de las variables y gráficos circulares. También se generaron gráficos de función de riesgos y de curvas de supervivencia con intervalos de confianza al 95 % para las variables estratificadas.

4 Metodología

El presente estudio empleó un enfoque sistemático para analizar el posible efecto de los datos clínicos de pacientes con cáncer de mama sobre el desarrollo de cardiotoxicidad. Se comenzó con la limpieza y preparación del conjunto de datos, seguida de un análisis exploratorio de las variables y sus relaciones. Posteriormente, se aplicaron métodos de análisis de supervivencia para modelar el tiempo hasta la aparición de cardiotoxicidad en función de las variables de interés, considerando la censura en los datos. El ajuste y validación de los modelos incluyó la selección iterativa de variables significativas, la evaluación de supuestos fundamentales y la representación gráfica para facilitar la interpretación de los resultados.

4.1. Preparación de los datos

El proceso comenzó con la limpieza y transformación del conjunto de datos. Las variables fueron revisadas para asegurar que estuvieran en el formato adecuado según su naturaleza (por ejemplo, numéricas para las cuantitativas). Los datos fueron ordenados por tiempo de seguimiento (para asegurar que no genere problemas al modelar y graficar) y se identificaron valores faltantes. Dado que los datos ausentes representaban un bajo porcentaje del total (10.73 %), se optó por su eliminación para mantener la consistencia en el análisis.

4.2. Significancia estadística

Se utilizó un nivel de significancia de $\alpha = 0,05$, considerado un estándar en análisis estadísticos, para asegurar la robustez de los resultados. Inicialmente, se consideró emplear un nivel de significancia más alto para aumentar la sensibilidad y detectar asociaciones potencialmente relevantes. Sin embargo, se optó por mantener el nivel convencional para minimizar el riesgo de errores tipo I y garantizar que los resultados reflejen asociaciones estadísticamente confiables.

4.3. Análisis exploratorio

Para comprender la estructura y relaciones de los datos, se realizó un análisis exploratorio. Se generaron histogramas para visualizar las distribuciones de las variables cuantitativas y diagramas de caja para detectar posibles valores atípicos. Asimismo, se calculó una matriz de correlación para evaluar las relaciones entre las variables cuantitativas y detectar posibles patrones de multicolinealidad. Esta exploración inicial permitió guiar las decisiones posteriores en el modelado estadístico.

4.4. Análisis de supervivencia

El análisis de supervivencia fue seleccionado debido a la presencia de censura por la derecha en los datos, es decir, casos en los que no se registró el evento de interés (desarrollo de cardiotoxicidad) durante el periodo de seguimiento. Este tipo de análisis permite integrar información parcial de individuos censurados, modelando adecuadamente los tiempos hasta el evento y evitando sesgos que podrían surgir al ignorar estas observaciones.

El análisis de supervivencia se empleó para modelar el tiempo hasta la aparición del evento de interés, considerando la naturaleza censurada de los datos. Se utilizó el método de Kaplan-Meier para estimar la función de supervivencia y generar curvas que describieran la probabilidad acumulada de no desarrollar el evento en función del tiempo. Estas curvas se compararon entre diferentes grupos mediante el test log-rank, una herramienta que evalúa la existencia de diferencias estadísticamente significativas entre las distribuciones de supervivencia.

Posteriormente, se ajustaron modelos de regresión de Cox para explorar la influencia de múltiples variables explicativas en el tiempo hasta el evento. Este modelo permitió incorporar simultáneamente varios factores y se validó utilizando el test de Schoenfeld, que verificó el cumplimiento

del supuesto de riesgos proporcionales.

4.5. Selección y validación del modelo

El proceso de selección del modelo multivariado se llevó a cabo de manera iterativa. Inicialmente, se incluyeron todas las variables, eliminando progresivamente aquellas no significativas en base a sus p-valores. Se evaluó la multicolinealidad mediante el cálculo del factor de inflación de la varianza (VIF), y se descartaron variables redundantes para garantizar la estabilidad y precisión del modelo.

La validación del modelo incluyó un análisis de residuos, como los de Martingale y deviance, para identificar posibles inconsistencias o valores atípicos. Adicionalmente, se calculó el índice de concordancia (C-index) para evaluar la capacidad del modelo de discriminar entre diferentes tiempos de aparición del evento.

4.6. Representación gráfica

Para facilitar la comunicación de los resultados del análisis, se generaron gráficos a lo largo de todo el proceso. Esto incluyó la representación de curvas de supervivencia con intervalos de confianza, así como visualizaciones específicas para evaluar las relaciones modeladas y los residuos generados por el ajuste.

5 Resultados

5.1. Exploración inicial de los datos

El conjunto de datos incluye 477 pacientes provenientes de un estudio clínico sobre cardiotoxicidad asociada al tratamiento de cáncer de mama. Cada paciente fue seguido durante un periodo variable, en el que se registraron diversas características clínicas y la aparición del evento de interés, definido como el desarrollo de cardiotoxicidad ($CTRC = 1$). De los pacientes incluidos, 47 (9.85 %) desarrollaron cardiotoxicidad, mientras que los 430 restantes (90.15 %) no presentaron el evento durante el periodo de seguimiento, siendo considerados como casos censurados.

5.1.1. Distribución de las variables

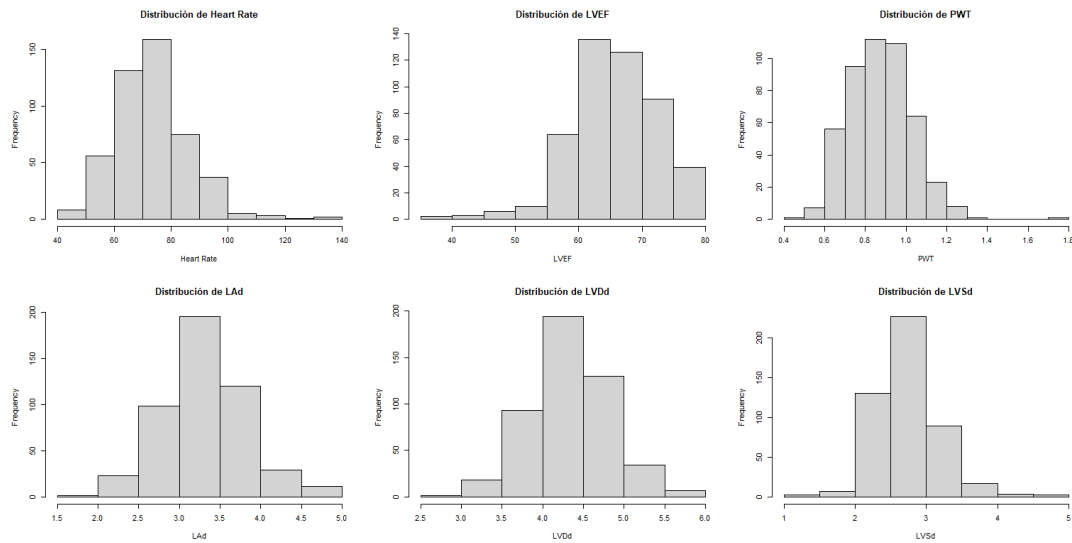


Figura 1: Histogramas de las variables cuantitativas

En cuanto a las variables cualitativas, las distribuciones presentadas muestran patrones coherentes con mediciones fisiológicas normales. La frecuencia cardíaca (Heart Rate) presenta una asimetría ligera hacia la derecha, la mayor parte de las muestras se distribuyen entre 60 y 80 latidos por minuto, con algunos valores más altos que podrían ser atípicos. La fracción de eyección del ventrículo izquierdo (LVEF), por otro lado, tiene una asimetría leve negativa, con la mayoría de los valores entre el 60 y 70 %. Variables como el espesor de la pared posterior del ventrículo izquierdo (PWT) y los diámetros de las estructuras cardíacas (LAd, LVDd, y LVSd) muestran distribuciones simétricas y concentradas alrededor de sus valores promedio, con rangos estrechos que indican baja variabilidad.

	Media	SD	IQR	Skewness	Kurtosis
heart_rate	74.2138	12.9836	16.0000	0.8438	2.0375
LVEF	65.6629	6.8425	9.4211	-0.5012	0.8342
PWT	0.8758	0.1538	0.2161	0.5023	1.2892
LAd	3.3282	0.4999	0.6347	0.2403	0.4220
LVDd	4.3455	0.4935	0.6458	0.1479	0.2565
LVSd	2.7444	0.4308	0.5202	0.5848	1.8164

Cuadro 1: Medidas descriptivas para las variables cuantitativas

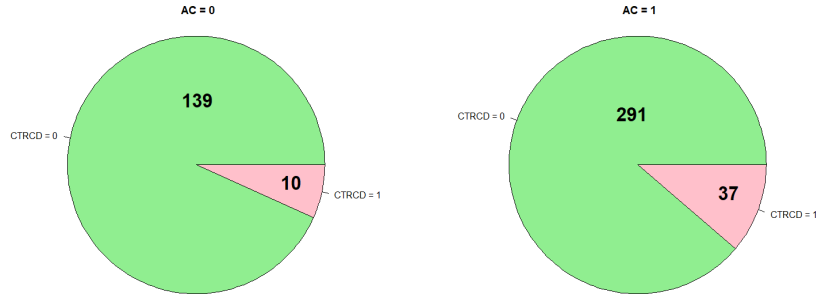


Figura 2: Gráficos circulares de las frecuencias de CTRCD para los niveles de AC

Los gráficos muestran que en ambos grupos, tratados con antraciclinas ($AC = 1$) y no tratados ($AC = 0$), la mayoría de los casos no presentan cardiotoxicidad ($CTRCD = 0$). Sin embargo, en el grupo tratado con antraciclinas, la proporción de casos con cardiotoxicidad ($CTRCD = 1$) es ligeramente mayor (12.71 % aprox.) en comparación con el grupo no tratado (7.19 % aprox.). Esto sugiere una posible relación entre el tratamiento con antraciclinas y la aparición de cardiotoxicidad, aunque la mayoría de los pacientes tratados no desarrollan esta condición.

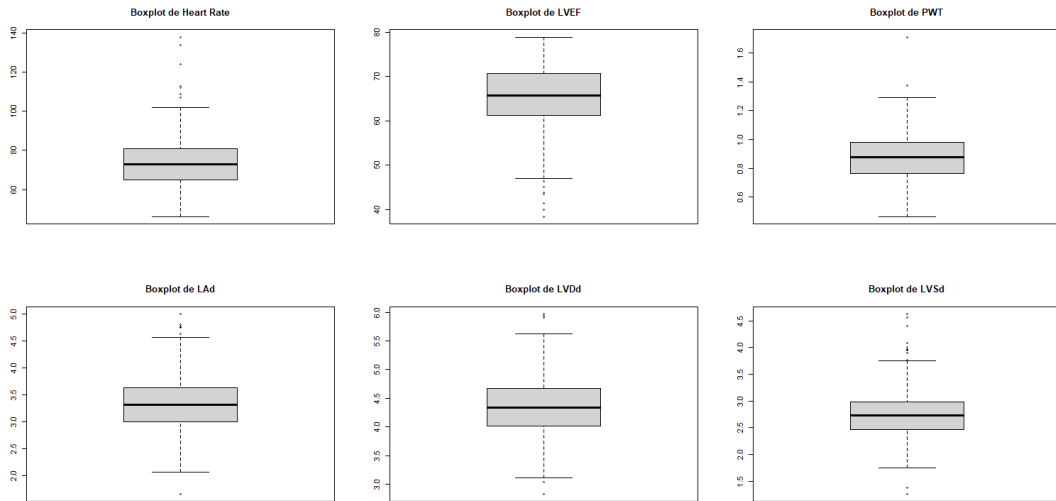


Figura 3: Diagramas de cajas de las variables cuantitativas

Los boxplots muestran cómo están distribuidos los datos de las variables y resaltan algunos valores que se alejan de lo normal. En todos los casos, los datos principales se concentran en una franja central bien definida, mientras que algunos puntos, considerados atípicos, se encuentran fuera de los extremos de los gráficos. Por ejemplo, en Heart Rate, la mayoría de los valores están alrededor del centro, pero hay algunos más altos que destacan. Algo similar ocurre con LVEF, mientras que PWT, LAd, LVDd y LVSD tienen menos valores extremos y están más uniformes en su rango central.

5.1.2. Correlación entre variables

El gráfico de correlación destaca una relación fuerte entre el diámetro diastólico (LVDd) y el sistólico (LVSD) del ventrículo izquierdo, así como una correlación negativa notable entre LVEF y LVSD, lo que sugiere que un aumento en el diámetro sistólico está asociado con una disminución en la fracción de eyección. En general, las mayores relaciones se dan entre las medidas estructurales del corazón.

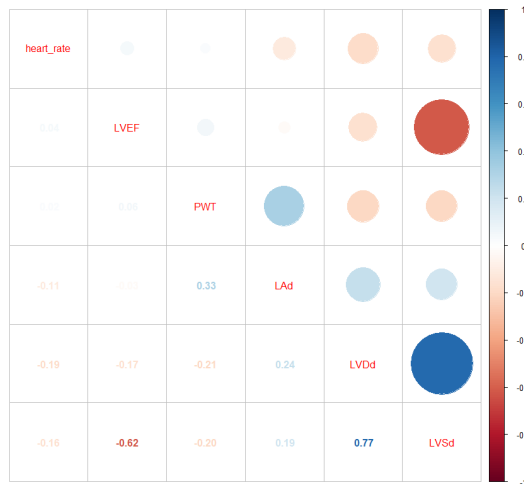


Figura 4: Representación gráfica de matriz de correlaciones

5.2. Análisis de supervivencia

Se analizó el tiempo hasta la aparición de cardiotoxicidad, mostrando las curvas de supervivencia y comparando los resultados entre los diferentes grupos. También se evaluó cómo factores como el tratamiento con antraciclinas y otras variables influyen en la probabilidad de desarrollar este evento.

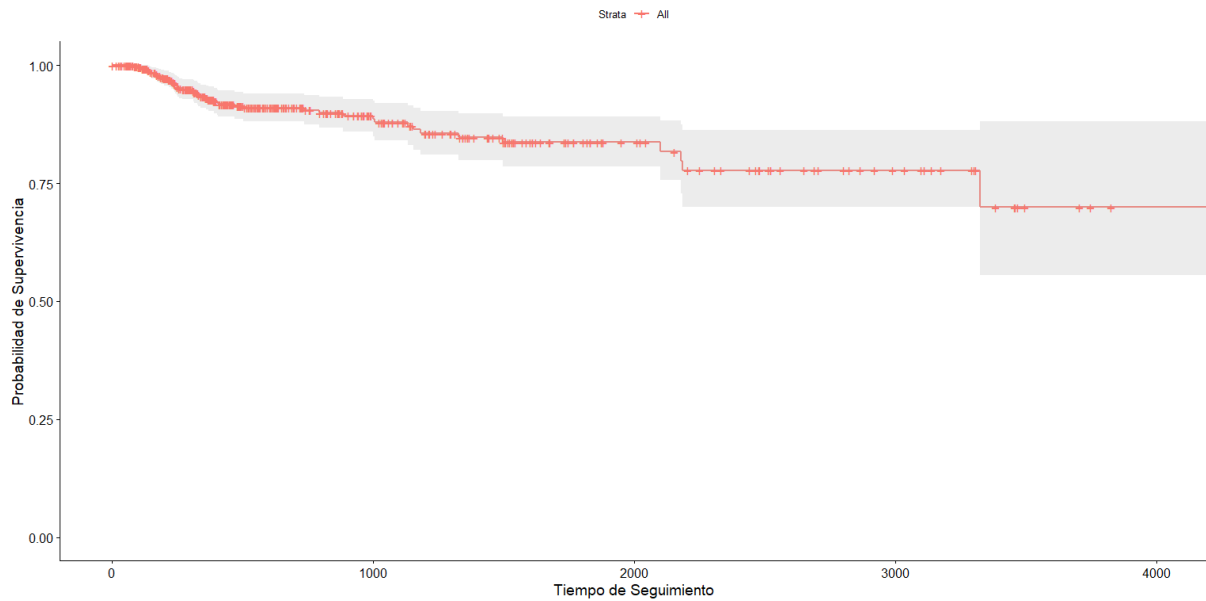


Figura 5: Gráfico de función de supervivencia

El gráfico muestra la función de supervivencia de los pacientes en relación con la cardiotoxicidad (CTRCD) a lo largo del tiempo. La probabilidad de no desarrollar cardiotoxicidad comienza alta (cerca de 1) y disminuye gradualmente con el tiempo, reflejando un aumento progresivo del riesgo. Las marcas en la línea indican casos censurados, es decir, pacientes que no experimentaron el evento dentro del período de observación. Las zonas grises alrededor de la línea representan los intervalos de confianza, mostrando mayor precisión al inicio del seguimiento y mayor incertidumbre hacia el final, probablemente debido a un menor número de pacientes en riesgo. En general, el gráfico refleja una supervivencia estable en los primeros períodos y una disminución más notable al final del seguimiento.

5.2.1. Modelos univariados

Los modelos de supervivencia univariados permiten evaluar de forma individual el impacto de cada variable en el tiempo hasta la aparición de cardiotoxicidad. Este análisis es esencial para identificar qué factores tienen una influencia significativa en la supervivencia, proporcionando una visión preliminar que guiará el desarrollo de modelos más complejos y ajustados.

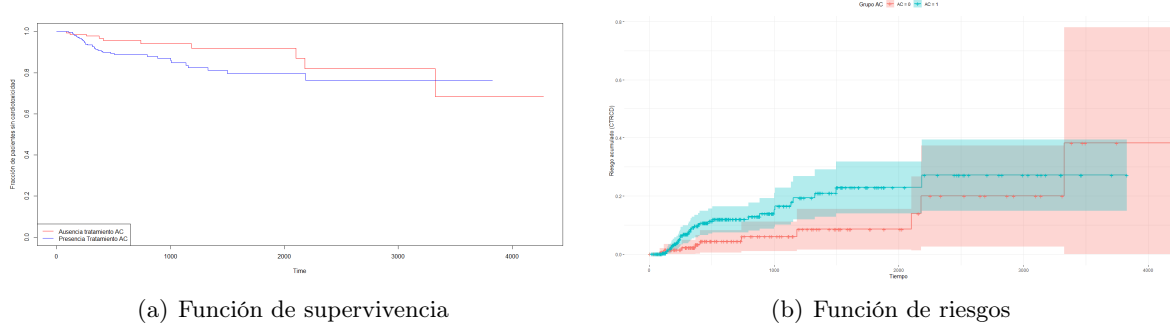


Figura 6: Gráficos de AC por grupos

Los resultados muestran que las pacientes tratadas con antraciclinas tienen un mayor riesgo acumulado de desarrollar cardiotoxicidad y una menor supervivencia libre del evento en comparación con las no tratadas. El modelo de Cox cuantifica esta tendencia, con una razón de riesgos (HR) de 1.9872 (IC 95 %: 0.9853-4.008), lo que indica casi el doble de probabilidad de cardiotoxicidad en las tratadas, aunque el resultado no alcanza significancia estadística ($p = 0.055$). Estos hallazgos, respaldados por los gráficos y el test de log-rank ($p = 0.05$), sugieren una asociación importante entre el tratamiento y un mayor riesgo, pero requieren análisis adicionales para confirmar su relevancia.

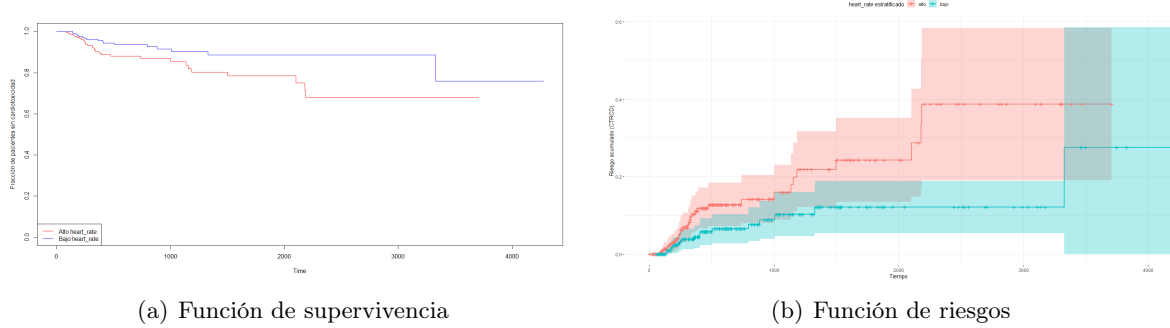


Figura 7: Gráficos de heart_rate estratificado por la mediana

Los resultados muestran que las mujeres con una frecuencia cardíaca elevada tienen un mayor riesgo acumulado de desarrollar cardiotoxicidad y una menor supervivencia libre del evento en comparación con aquellas con frecuencia cardíaca baja. El modelo de Cox cuantifica esta tendencia, indicando que por cada aumento unitario en la frecuencia cardíaca, el riesgo de cardiotoxicidad incrementa un 3.3 % ($HR = 1.033$; IC 95 %: 1.014-1.053; $p < 0.001$). La concordancia del modelo ($C = 0.639$, $se = 0.04$) sugiere una capacidad moderada para discriminar entre pacientes con diferentes niveles de riesgo. Estos hallazgos, respaldados por los gráficos y el test de log-rank ($p = 6e-04$), sugieren que la frecuencia cardíaca es un factor importante asociado a un mayor riesgo de cardiotoxicidad en mujeres, con resultados estadísticamente significativos que fortalecen esta asociación.

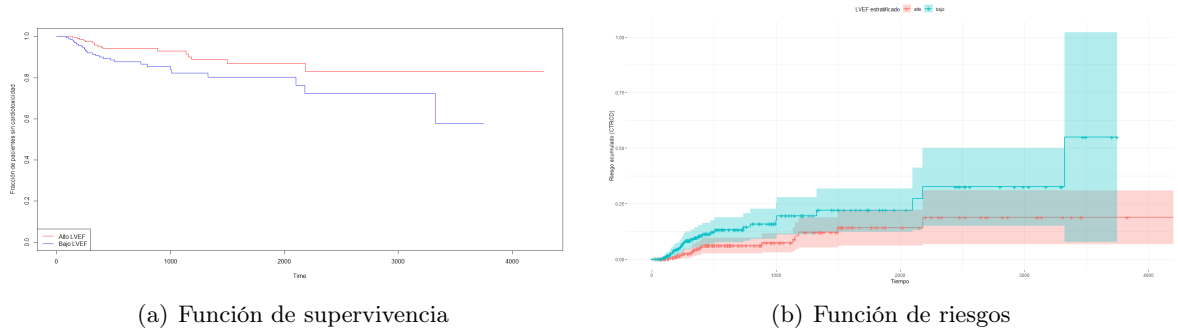


Figura 8: Gráficos de LVEF estratificado por la mediana

Los resultados muestran que las mujeres con una fracción de eyección ventricular izquierda (LVEF) baja tienen un mayor riesgo acumulado de desarrollar cardiotoxicidad y una menor supervivencia libre del evento en comparación con aquellas con LVEF alta. El modelo de Cox cuantifica esta tendencia, indicando que por cada incremento unitario en LVEF, el riesgo de cardiotoxicidad disminuye en un 4.4 % ($HR = 0.9557$; IC 95 %: 0.9201-0.9927; $p = 0.0192$). La concordancia del modelo ($C = 0.641$, $se = 0.04$) sugiere una capacidad moderada para discriminar entre pacientes con diferentes niveles de riesgo. Estas diferencias son estadísticamente significativas, como lo confirman el test de log-rank ($p = 0.02$) y el test de Wald ($p = 0.02$), subrayando la importancia de LVEF como un factor protector frente al desarrollo de cardiotoxicidad.

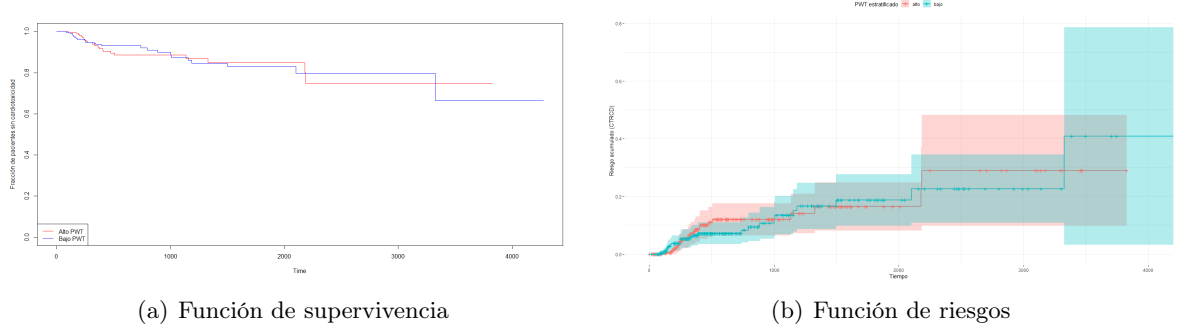


Figura 9: Gráficos de PWT estratificado por la mediana

Los resultados muestran que no hay diferencias significativas en el riesgo acumulado de desarrollar cardiotoxicidad ni en la supervivencia libre del evento entre mujeres con un grosor de la pared posterior ventricular (PWT) alto o bajo. El modelo de Cox cuantifica esta relación, con una razón de riesgos de 0.8735 (IC 95 %: 0.1298-5.877; $p = 0.889$), lo que indica que el grosor de la pared posterior no se asocia significativamente con el desarrollo de cardiotoxicidad. La concordancia del modelo ($C = 0.511$) sugiere que no tiene capacidad discriminativa. Además, el resultado del test de log-rank ($p = 0.9$) refuerza la falta de asociación entre esta variable y el riesgo de cardiotoxicidad.

Resto de variables: Los análisis univariados para el grosor de la pared posterior ventricular (PWT), el diámetro de la aurícula izquierda (LAd), el diámetro diastólico (LVDd) y el diámetro sistólico (LVSd) no mostraron asociaciones significativas con el desarrollo de cardiotoxicidad. Los valores p en todos los modelos fueron superiores a 0.2, y las razones de riesgos no evidenciaron un patrón consistente de aumento o disminución del riesgo. Además, las concordancias de los modelos oscilaron entre 0.474 y 0.575, lo que sugiere una baja capacidad discriminativa para estas variables en la predicción del riesgo de cardiotoxicidad. En conjunto, estos resultados indican que estas medidas no son factores relevantes para explicar el riesgo de desarrollar cardiotoxicidad en esta cohorte.

	HR	IC 95 %	p-valor	Concordancia	Test Log-rank p
PWT	0.8735	[0.1298, 5.877]	0.889	0.511	0.9
LAd	1.057	[0.5848, 1.909]	0.855	0.474	0.9
LVDd	0.9512	[0.531, 1.704]	0.866	0.488	0.9
LVSD	1.505	[0.7965, 2.843]	0.208	0.575	0.2

Cuadro 2: Resumen de los modelos univariados para PWT, LAd, LVDd y LVSD

5.2.2. Modelos multivariados

El modelo multivariado final se obtuvo mediante un proceso de selección progresiva, excluyendo iterativamente las variables con mayor p-valor hasta conservar únicamente aquellas significativamente asociadas al riesgo de cardiotoxicidad. El modelo final incluyó las variables `heart_rate` y `LVEF`, mostrando que una mayor frecuencia cardíaca está asociada con un aumento del 3.5 % en el riesgo de cardiotoxicidad por cada unidad adicional (HR = 1.035; IC 95 %: 1.0162-1.0550; $p < 0.001$), mientras que un mayor valor de `LVEF` actúa como un factor protector, disminuyendo el riesgo en un 4.8 % por cada unidad adicional (HR = 0.952; IC 95 %: 0.9162-0.9884; $p = 0.010$). Este modelo mostró un ajuste adecuado y resalta la importancia de estas variables en la predicción del riesgo de cardiotoxicidad.

	HR	IC 95 %	Coefficiente	p-valor
heart_rate	1.035	[1.0162, 1.0550]	0.0348	<0.001
LVEF	0.952	[0.9162, 0.9884]	-0.0496	0.010

Cuadro 3: Parámetros del modelo multivariado final

También se probó ajustar un modelo con interacción entre `heart_rate` y `LVEF`, pero los coeficientes de interacción y sus efectos principales no fueron significativos ($p > 0.6$). Esto sugiere que no hay evidencia de que el efecto de `heart_rate` sobre la cardiotoxicidad dependa de los niveles de `LVEF` o viceversa.

5.2.3. Comprobación de Hipótesis Estructurales

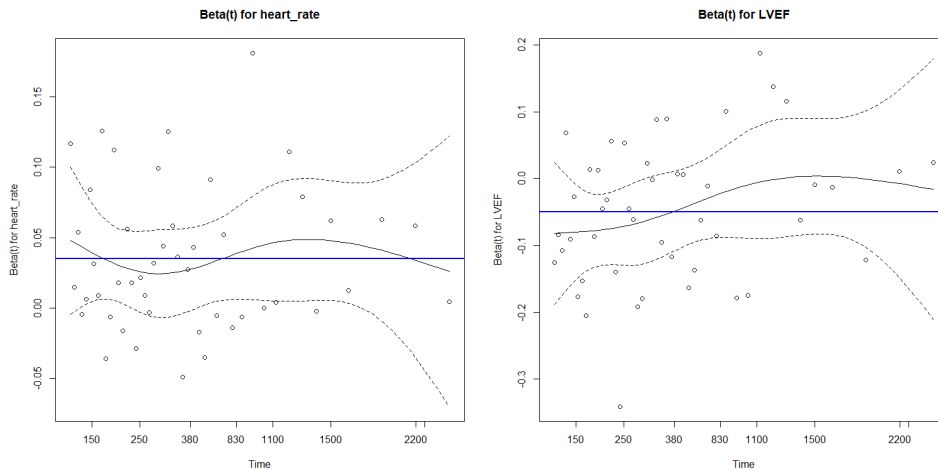


Figura 10: Graficos de la estimación de los coeficientes $\beta(t)$

Supuesto de Riesgos Proporcionales Se evaluó el supuesto de riesgos proporcionales utilizando gráficos de $\beta(t)$ para las variables `heart_rate` y `LVEF`. Los coeficientes estimados mostraron poca variación a lo largo del tiempo, indicando que el efecto de cada variable sobre el riesgo es constante. Esto fue confirmado estadísticamente mediante el test de riesgos proporcionales global ($p = 0.37$), lo que sugiere que el modelo de Cox ajustado cumple con este supuesto y no requiere ajustes adicionales.

Multicolinealidad La posible multicolinealidad entre las variables del modelo fue evaluada calculando el Factor de Inflación de la Varianza (VIF) mediante un modelo de regresión lineal. Los valores obtenidos fueron muy bajos ($VIF < 1,01$), lo que descarta problemas de colinealidad entre *heart_rate* y *LVEF*. Por lo tanto, ambas variables pueden ser incluidas en el modelo sin preocupación por redundancia o distorsión en las estimaciones.

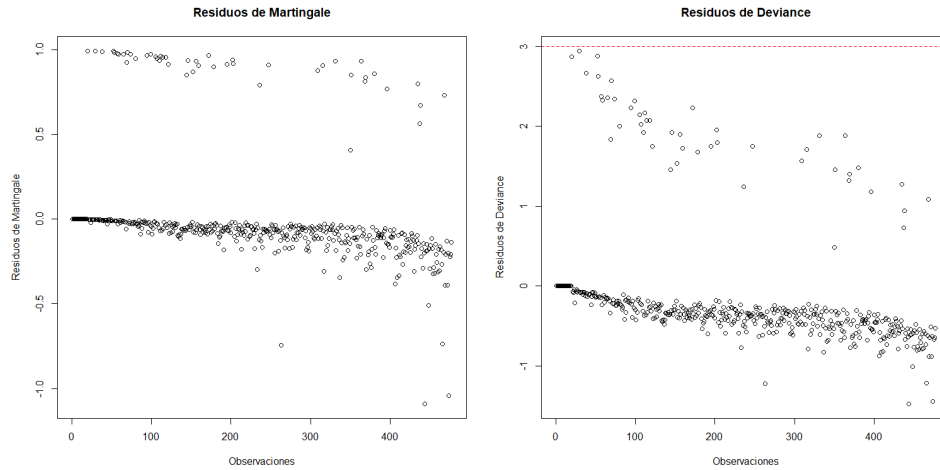


Figura 11: Graficos para detección de valores atípicos

Valores Atípicos Los residuos de Martingale y de Deviance fueron analizados para detectar posibles valores atípicos en los tiempos de supervivencia y las observaciones. Ningún residuo de Deviance excedió el umbral de 3, lo que indica que no se identificaron casos problemáticos o extremos en los datos. Esto sugiere que las estimaciones del modelo no están influidas por observaciones atípicas, asegurando la estabilidad y validez de los resultados.

6 Conclusiones

En este estudio se analizaron variables clínicas relacionadas con el riesgo de cardiotoxicidad en pacientes con cáncer de mama, identificando que la frecuencia cardíaca y la fracción de eyección ventricular izquierda son factores clave. Un aumento en la frecuencia cardíaca incrementa el riesgo, mientras que una mayor fracción de eyección ventricular actúa como un factor protector.

El modelo ajustado cumple con los supuestos de riesgos proporcionales y no mostró problemas de multicolinealidad ni valores atípicos significativos, lo que asegura la validez de los resultados. Estos hallazgos subrayan la importancia de estas variables en la monitorización clínica y abren la puerta a futuros estudios que amplíen este enfoque con muestras más grandes y otros factores relevantes.

Aún así, hay que tener en cuenta el gran desbalance que existe entre el número de pacientes que desarrollaron cardiotoxicidad y que no la desarrollaron. Podrían realizarse futuros estudios en los que se aumente el tamaño muestral, con el objetivo de obtener conclusiones más robustas sobre la influencia de las variables estudiadas en el desarrollo de cardiotoxicidad.

Bibliografía

- [1] DATAtab. Cox regression tutorial, 2024. Último acceso: 9 de diciembre de 2024.
- [2] Universitat de Barcelona. El modelo de cox de riesgos proporcionales, 2024. Último acceso: 9 de diciembre de 2024.
- [3] John Fox and Manuel Marquez. *RcmdrMisc: R Commander Miscellaneous Functions*, 2023. R package version 2.9-1.
- [4] John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition, 2019.
- [5] Alboukadel Kassambara, Marcin Kosinski, and Przemyslaw Biecek. *survminer: Drawing Survival Curves using 'ggplot2'*, 2024. R package version 0.5.0.
- [6] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020.
- [7] Terry M Therneau. *A Package for Survival Analysis in R*, 2024. R package version 3.7-0.
- [8] Taiyun Wei and Viliam Simko. *R package 'corrplot': Visualization of a Correlation Matrix*, 2024. (Version 0.95).