NCCTG Lung Cancer Data

Introducción

Los datos “NCCTG Lung Cancer Data” provienen del National Cancer Institute de los Estados Unidos. Este conjunto de datos contiene información sobre pacientes con cáncer de pulmón que participaron en un estudio clínico realizado por el North Central Cancer Treatment Group (NCCTG). El principal propósito de estos datos es realizar análisis de supervivencia para estimar la probabilidad de que un paciente con cáncer de pulmón sobreviva más allá de cierto tiempo. Proporciona información sobre características demográficas, índices de estado funcional y otros factores que pueden influir en la supervivencia de los pacientes.

Estructura de datos

La base de datos utilizada en este estudio contiene información sobre pacientes con cáncer de pulmón avanzado. Las variables incluidas se describen a continuación:

Variables:

- **inst:**Número de identificación de la institución médica donde se realizó el tratamiento.
- **time:**Tiempo en días desde el inicio del estudio hasta la muerte o el último contacto.
- **status:**Estado vital del paciente en el último contacto:
    - 0: Censurado (el paciente seguía vivo en el último contacto).
    - 1: Muerte (el paciente falleció durante el estudio).
- **age:**Edad del paciente en años en el momento del diagnóstico.
- **sex:**Género del paciente:
    - 0: Femenino
    - 1: Masculino
- **ph.ecog:**Índice de estado funcional del Eastern Cooperative Oncology Group (ECOG):

    - 0: Asintomático, capaz de realizar actividades normales.
    - 1: Síntomas leves, capaz de realizar actividades ligeras.
    - 2: Síntomas que limitan las actividades normales.
    - 3: Encamado más del 50% del día.
    - 4: Encamado el 100% del día.
    - 5: Fallecido.
- **ph.karno:**Índice de Karnofsky, una escala de 0 a 100 que evalúa la capacidad funcional de los pacientes con cáncer.
- **pat.karno:**Índice de Karnofsky del paciente al ingresar al estudio (KPS).
- **meal.cal:**Número de calorías consumidas por el paciente por día.
- **wt.loss:**Pérdida de peso en libras en los últimos seis meses.

Análisis Exploratorio

Características de la base de datos

variable	statistic	Total
Sample	N	228
age	Mean (SD)	62.4 (9.07)
age	Median (IQR)	63 (56-69)
age	Min-max	39-82
age	Missing	0 (0%)
sex	Female	90 (39.5%)
sex	Male	138 (60.5%)

Al examinar *Table 1*, se destaca que el conjunto comprende un total de 228 observaciones. La edad promedio de los pacientes es de 62.4 años, según la *Organización Mundial de la Salud*, los rangos de edad para cáncer son:

Rangos de edad para el cáncer:

- **Niños:** 0-14 años
- **Adolescentes y adultos jóvenes:** 15-39 años

- **Adultos:** 40-64 años
- **Adultos mayores:** 65 años o más

Para el cáncer, 62.4 años se considera en el rango **adultos mayores**, por lo cual el promedio de los pacientes son adultos mayores. Los valores extremos oscilan entre 39 y 82 años, lo que indica una amplia variabilidad en la edad de los pacientes.

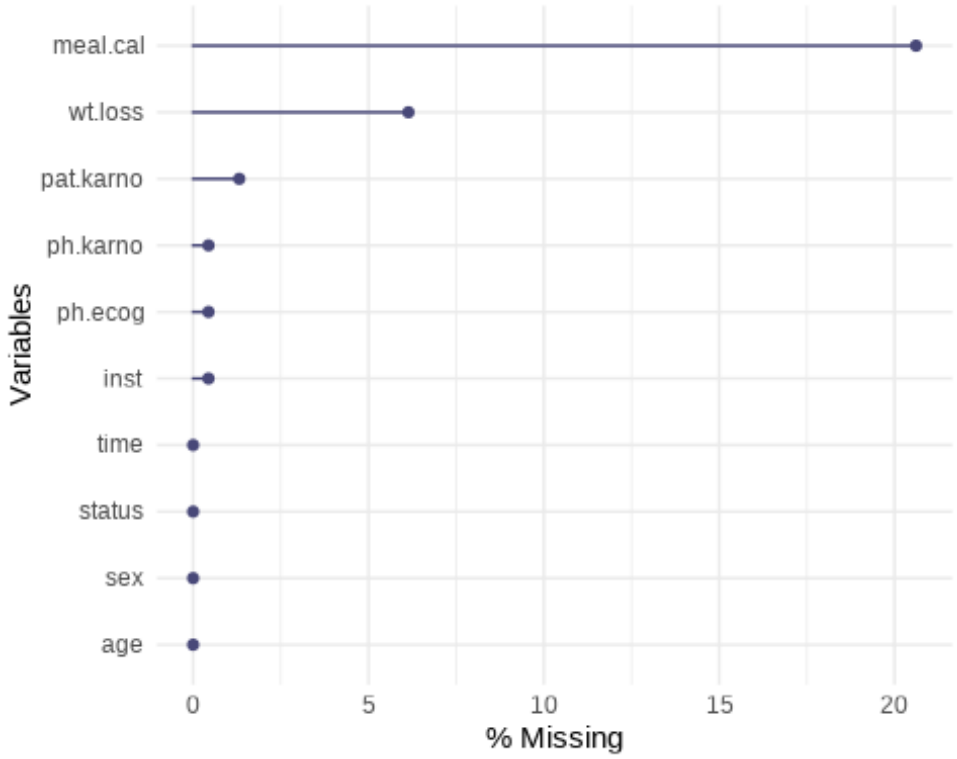
Es importante resaltar que no se observan valores perdidos en la variable de edad, lo que fortalece la integridad de los datos en este aspecto. Al explorar la distribución por género, se identifica que el sexo femenino representa el 39.5% de las observaciones, con un total de 90 casos. Por otro lado, el sexo masculino predomina en el estudio, con 138 observaciones, equivalente al 60.5% del total.

Este análisis evidencia una proporción mayor de hombres que de mujeres en la muestra estudiada.

variable	statistic	Total	Female	Male
Sample	N	228	90	138
age	Mean (SD)	62.4 (9.07)	61.1 (8.85)	63.3 (9.14)
age	Median (IQR)	63 (56-69)	61 (55-68)	64 (57-70)
age	Min-max	39-82	41-77	39-82
age	Missing	0 (0%)	0 (0%)	0 (0%)

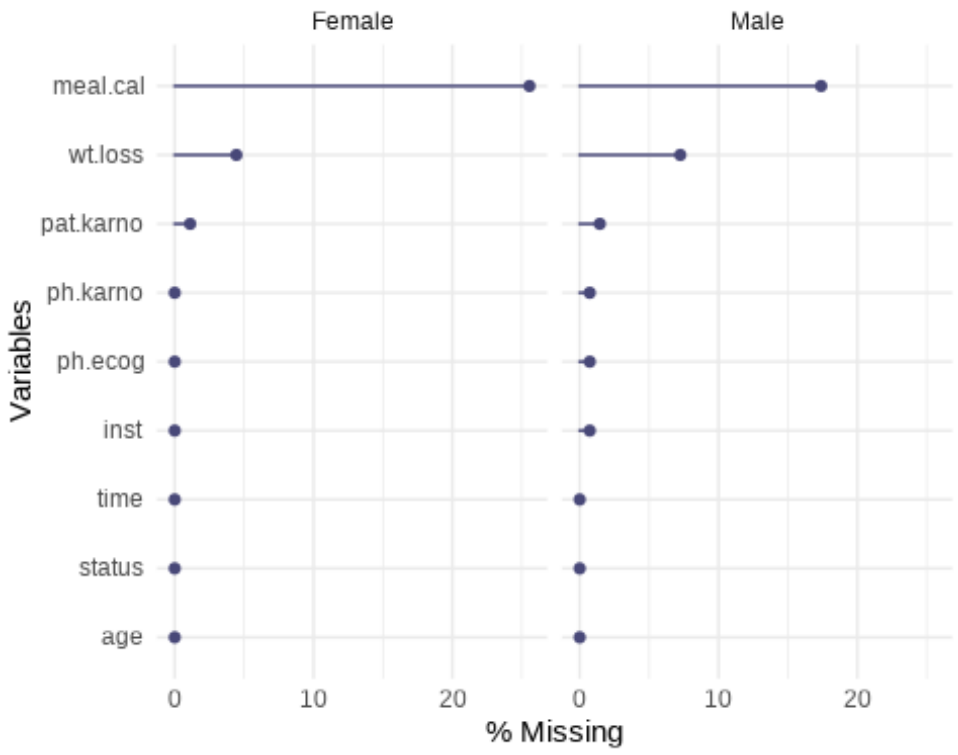
La *Table 2.* muestra que para las pacientes femeninas, la edad promedio es de 61.1 años, con una desviación estándar de 8.85. En contraste, los pacientes masculinos que presentan una edad promedio ligeramente mayor, con un valor de 63.3 años. Esto sugiere una tendencia hacia edades más avanzadas en el grupo de hombres en comparación con las mujeres.

La mediana de edad para las pacientes femeninas se sitúa en 61 años, mientras que para los pacientes masculinos se eleva a 64 años, evidenciando una diferencia de 3 años entre ambos grupos. Además, se observa una mayor variabilidad en la edad de los pacientes masculinos, con un rango que se extiende desde los 39 hasta los 82 años, en contraste con el rango de edad de las pacientes femeninas, que va desde los 41 hasta los 77 años. Es importante destacar que no se han identificado valores perdidos en la variable de edad para ninguno de los dos grupos de género, lo que confirma la integridad de los datos en este aspecto.



Porcentaje de valores faltantes en la base de datos 'cancer'

La *Figure 1*. presenta el porcentaje de valores faltantes o "missing" en la base de datos. Se observa que la variable con el mayor número de valores faltantes es meal.cal, con más del 20% de sus observaciones sin datos registrados. Le sigue wt.loss, con más del 5% de sus valores ausentes.



Porcentaje de valores faltantes por sexo en la base de datos 'cancer'

Al segmentar los datos por sexo en la *Figure 2.*, se revela una disparidad en la distribución de valores faltantes. En el caso de las mujeres, se observa que la variable meal.cal exhibe un preocupante porcentaje de más del 25% de valores faltantes, mientras que en los hombres este porcentaje es inferior, situándose por debajo del 20% para la misma variable. Por otro lado, la situación se invierte al analizar la variable wt.loss, donde los hombres muestran un porcentaje más alto de valores faltantes, alcanzando el 7%, mientras que las mujeres presentan un 4%

Resumen de indicadores de salud pulmonar

variable	statistic	Total
Sample	N	228
pat.karno	Mean (SD)	80 (14.6)
pat.karno	Median (IQR)	80 (70-90)
pat.karno	Min-max	30-100
pat.karno	Missing	3 (1.32%)
meal.cal	Mean (SD)	929 (402)
meal.cal	Median (IQR)	975 (635-1150)

variable	statistic	Total
meal.cal	Min-max	96-2600
meal.cal	Missing	47 (20.6%)
wt.loss	Mean (SD)	9.83 (13.1)
wt.loss	Median (IQR)	7 (0-15.8)
wt.loss	Min-max	-24-68
wt.loss	Missing	14 (6.14%)
time	Mean (SD)	305 (211)
time	Median (IQR)	256 (167-396)
time	Min-max	5-1022
time	Missing	0 (0%)

Tabla 3. La ingesta calórica mostró una amplia variabilidad, con un rango de 96 a 2600 calorías por día y una media de 928 calorías. Sin embargo, esta variable presenta la mayor tasa de datos faltantes (20.6%).

En cuanto al estado funcional, la mayoría de los pacientes se encontraban en buen estado según el índice de Karnofsky (1949), con una concentración de datos en el valor 80. El coeficiente nominal de Cohen (1960) también respalda esta conclusión, con puntuaciones cercanas a 80 y una gran proporción de pacientes capaces de realizar actividades normales o ligeras.

Al final del estudio, la mayoría de los pacientes seguían vivos. Los tiempos de seguimiento variaron considerablemente, con una media de 305 días. La pérdida de peso promedio en los últimos seis meses fue de 9.8 libras.

Análisis de la relación entre el estado del paciente y el tiempo de seguimiento

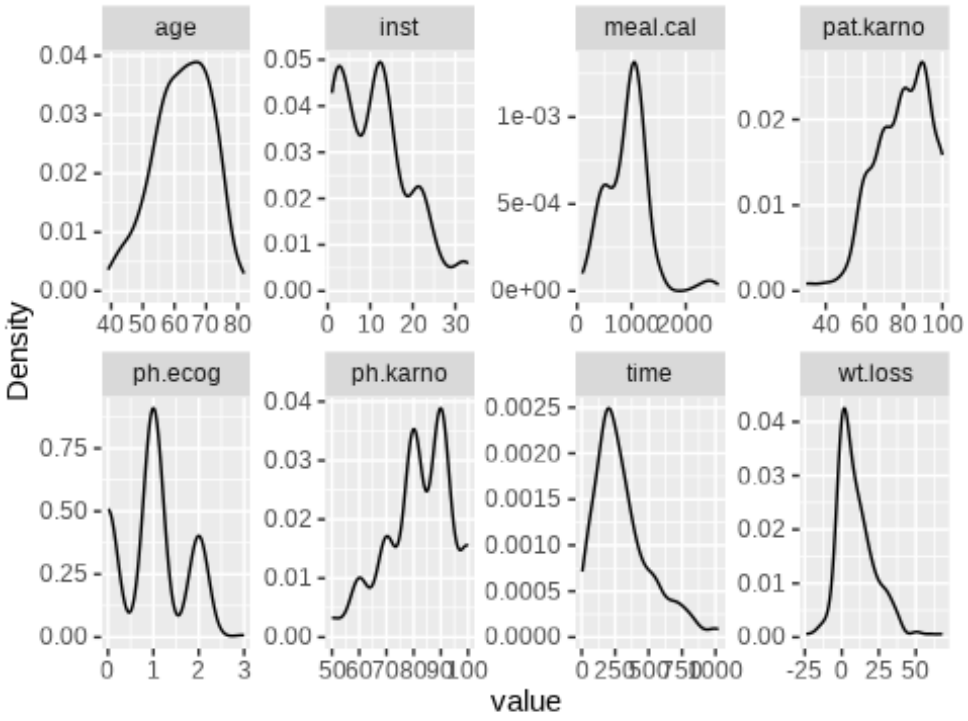
Relación entre el estado del paciente y el tiempo de seguimiento

name	desc	1 (N=63)	2 (N=165)	p	id
time	Mean ± SD	363.5 ± 221.1	283.0 ± 202.8	.010	time

Para aquellos pacientes cuyo estado final fue clasificado como *“Level 1: Censurado (el paciente seguía vivo en el último contacto)”*, el tiempo medio de seguimiento fue de aproximadamente 363.5 días, con una variabilidad alta, lo que sugiere que estos pacientes podrían haber fallecido durante el estudio. En contraste, para aquellos pacientes clasificados como *“Level 2: Muerte (el paciente*

falleció durante el estudio).”, el tiempo medio de seguimiento fue más corto, alrededor de 283.0 días, lo que podría indicar que estos pacientes estaban vivos o en mejores condiciones de salud al final del estudio. Además, el análisis estadístico mostró que esta diferencia en los tiempos de seguimiento entre los dos estados fue significativa, con un valor p de 0.010. Esto respalda la idea de que hay una diferencia real en los tiempos de seguimiento entre los pacientes que fallecieron y los que sobrevivieron, lo que podría ser relevante para comprender la progresión de la enfermedad y el resultado final de los pacientes con cáncer.

Distribución de Densidad de las Variables

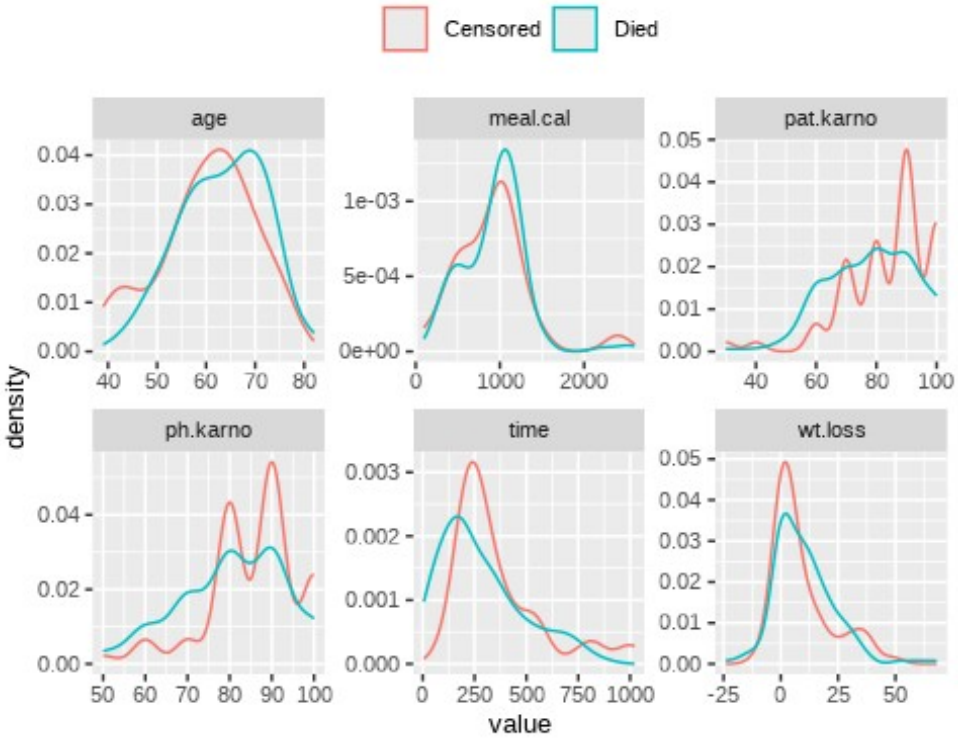


Gráficos de densidad de las variables

Al analizar detenidamente las gráficas de densidad presentadas en la Figure 3, podemos obtener información valiosa sobre la distribución de nuestras variables. En primer lugar, la variable ‘edad’ exhibe una distribución en forma de campana, indicando que la mayoría de los pacientes tienen entre 50 y 75 años de edad. Por otro lado, al examinar ‘meal.cal’, que representa la cantidad de calorías consumidas diariamente, observamos que la mayoría de los pacientes ingieren menos de 1500 calorías al día, mientras que solo unos pocos alcanzan un consumo entre 1500 y 2500.

En relación con ‘pat.karno’, una medida de la capacidad funcional, notamos que la mayoría de los pacientes obtienen una puntuación entre 60 y 100, con una mejora gradual desde 60 hasta 90. Similarmente, ‘ph.karno’, otro indicador de la capacidad funcional, sigue una tendencia ascendente desde valores bajos hasta alcanzar un máximo alrededor de 90 antes de disminuir nuevamente.

La variable ‘tiempo’, que representa el tiempo transcurrido desde el inicio del estudio hasta la muerte o el último contacto, parece concentrarse principalmente en los primeros 500 días, con menos pacientes seguidos por períodos más largos. Por último, ‘wt.loss’, que mide la pérdida de peso, muestra que la mayoría de los pacientes no experimentan una pérdida significativa de peso, ya que la mayoría reporta cero pérdida.



Distribución de Densidad de las Variables por estatus

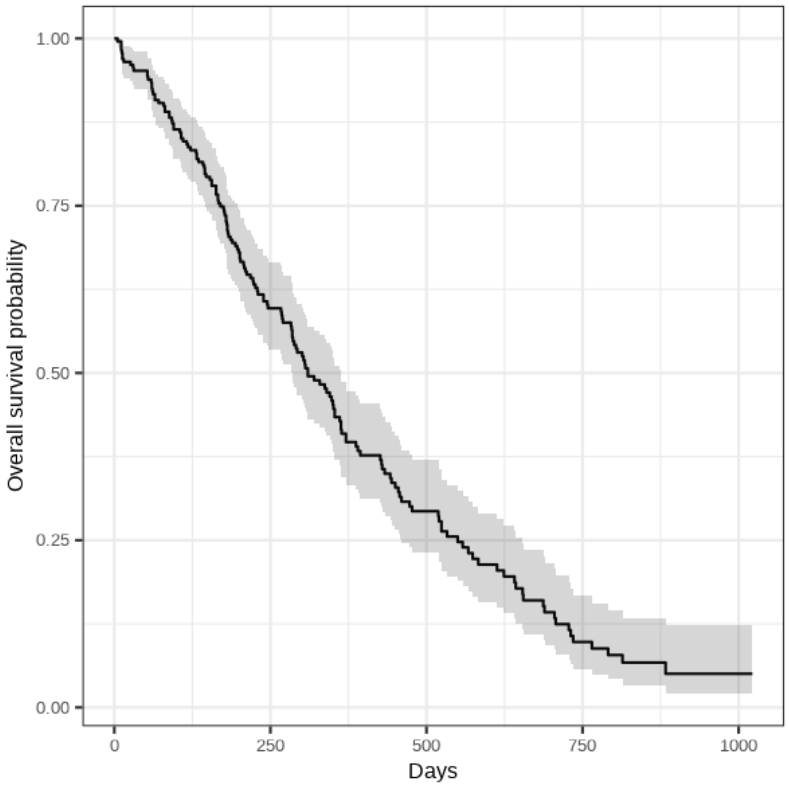
Al examinar detalladamente las gráficas de densidad presentes en la *Figure 4.* para la variable ‘edad’, podemos distinguir dos distribuciones claramente definidas: una para los datos censurados y otra para los casos de muerte. En el caso de los datos censurados, observamos una distribución normal con una media aproximada de 65 años, mientras que para los casos de muerte, la distribución parece estar ligeramente sesgada hacia la derecha, con una media alrededor de los 70 años.

Al analizar la variable ‘meal.cal’, notamos que las distribuciones de censura y muerte son bastante similares, con una ligera ventaja en la densidad de muerte. Por otro lado, al observar ‘pat.karno’, las distribuciones muestran diferencias significativas: los datos censurados presentan un aumento progresivo desde 40 hasta 75, con un pico máximo en 90, mientras que para los casos de muerte, la curva es más suave, alcanzando su punto máximo en 80.

En cuanto a ‘ph.karno’, los datos censurados exhiben múltiples picos de densidad, especialmente en 80 y 90, mientras que para los casos de muerte, la curva es más suave y menos pronunciada. En relación con la variable ‘tiempo’, los datos censurados muestran un punto máximo alrededor de los 250 días, mientras que la distribución para los casos de muerte está ligeramente sesgada hacia la izquierda.

Por último, al analizar ‘wt.loss’, observamos que ambas distribuciones de censura y muerte son bastante similares, con la densidad de censura ligeramente más alta que la de muerte.

Función de sobrevivencia de los pacientes



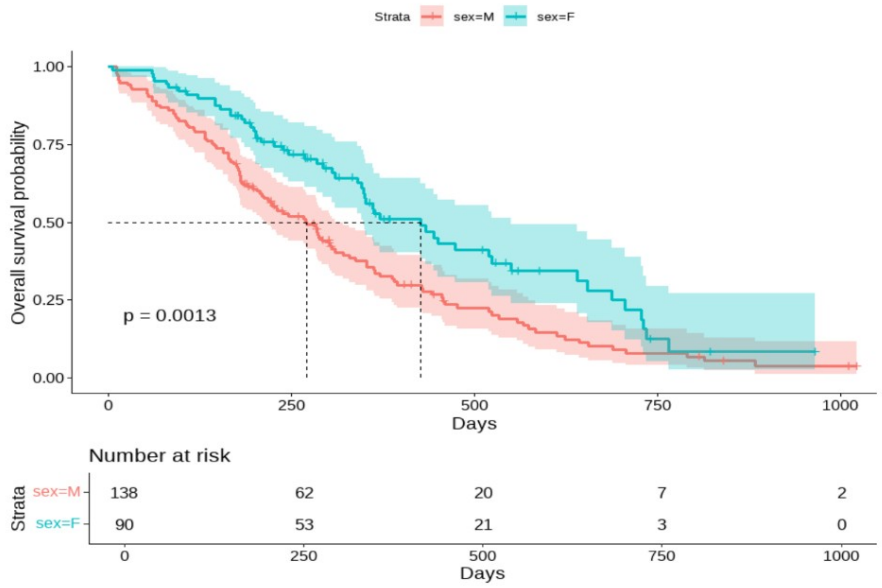
Gráfica de Función de Sobrevivencia de Pacientes con Cáncer de Pulmón

En esta gráfica, estamos viendo la probabilidad de supervivencia de todos los pacientes a lo largo del tiempo. Es como mirar cómo cambia la posibilidad de que un paciente siga con vida a medida que pasa el tiempo. Observamos lo siguiente:

- Al inicio del estudio, en el día cero, tenemos a todos los 228 pacientes en la gráfica. En este momento, no se ha registrado ningún evento de pérdida de vida.
- A medida que avanzamos en el tiempo, vemos que la línea central, que representa la probabilidad de sobrevivir, comienza a disminuir gradualmente. Esto significa que, en general, la posibilidad de que un paciente siga vivo va disminuyendo con el tiempo.
- La franja gris que rodea a la línea central muestra el intervalo de confianza del 95%. Es como una “zona de confianza” que nos dice dónde probablemente se encuentra la verdadera curva de supervivencia.

A medida que nos alejamos del día cero, este intervalo de confianza se va haciendo más ancho. Esto puede indicar que nuestra certeza sobre la probabilidad de supervivencia disminuye a medida que pasa el tiempo. esta gráfica nos muestra cómo cambia la probabilidad de supervivencia de los pacientes con cáncer de pulmón a lo largo del tiempo. A medida que pasan los días, vemos que esta probabilidad tiende a disminuir, y el intervalo de confianza nos ayuda a entender dónde podría estar la verdadera línea de supervivencia en función de nuestros datos.

Función de supervivenciaestratificada por sexo para pacientes con cáncer de pulmón



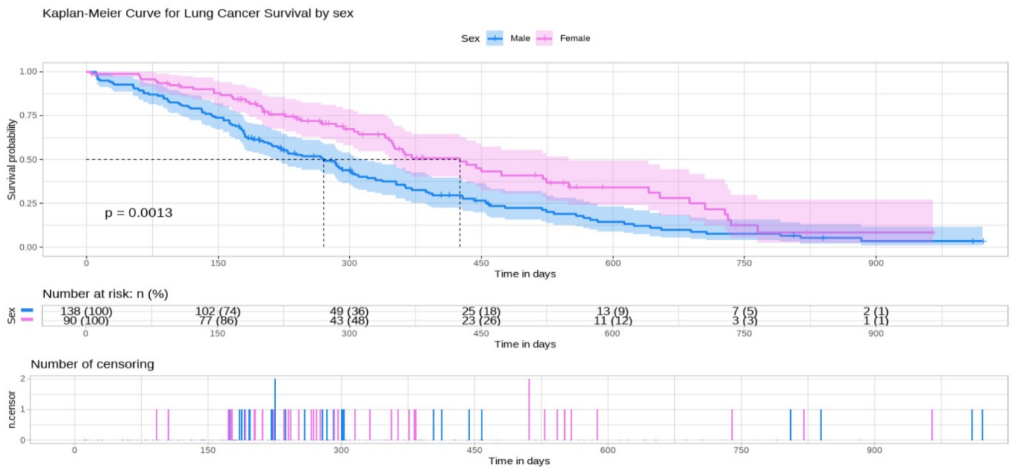
Gráfica de función de supervivencia estratificada por sexo para pacientes con cáncer de pulmón.

En esta gráfica de función de supervivencia estratificada por sexo para pacientes con cáncer de pulmón, podemos extraer importantes conclusiones:

- Se observa una clara tendencia de las mujeres a tener una mayor probabilidad de sobrevivir en comparación con los hombres.
- Las mujeres tienen una probabilidad mediana de supervivencia significativamente mayor que los hombres, indicando una mejor tasa de supervivencia.
- Los datos también respaldan estadísticamente esta observación, con un p-valor menor que 0.05, lo que indica una diferencia significativa en la supervivencia entre hombres y mujeres con cáncer de pulmón.

Este análisis resalta la importancia del sexo como un factor determinante en la supervivencia de los pacientes con cáncer de pulmón, lo que puede ser crucial para la toma de decisiones clínicas y el enfoque de tratamiento personalizado para cada paciente.

Función de sobrevivencia de los pacientes con cancer y el número de censuras estratificado por sexo.



Gráfica de la función de sobrevivencia de los pacientes con cáncer y el número de censuras estratificado por sexo.

En esta figura se puede observar la función de supervivencia y las censuras estratificadas por sexo en pacientes con cáncer de pulmón, notamos lo siguiente:

- Existe una tendencia generalizada en la duración de la supervivencia, con la mayoría de los pacientes presentando un periodo mediano consistente.
- El sexo del paciente no parece tener un efecto significativo en la distribución de los casos censurados y la duración de la supervivencia.
- La distribución de los casos censurados a lo largo del tiempo muestra un incremento especialmente marcado entre los días 150 y 300, independientemente del sexo del paciente.

Este análisis visual nos proporciona información valiosa sobre cómo se comporta la supervivencia de los pacientes con cáncer de pulmón en relación con el sexo, así como la distribución de los casos censurados a lo largo del tiempo. La consistencia en la duración mediana de la supervivencia y la falta de asociación clara entre el sexo y la supervivencia son aspectos destacados de estos hallazgos.