Artículo Original / Original Research

Ciencias Matemáticas, Vol. 38, No. 1, 2024, Pag. 69-76
Recibido (*Received*): 06-08-2024, Revisado (*Revised*): 14-11-2024
Aceptado (*Accepted*): 03-01-2025, En línea (*Online*): 18-03-2025

Ecosistema de ciencia de datos para el análisis de largos supervivientes en cáncer

Data science ecosystem for the analysis of long-term cancer survivors

Jorge Luis Palomino Hernández[1]*, Patricia Lorenzo-Luaces Álvarez[2], Lizet Sánchez Valdés[3]

Resumen La introducción de nuevas inmunoterapias ha incrementado la esperanza de vida en pacientes con cáncer, aunque la respuesta al tratamiento varía significativamente, observándose subpoblaciones con corta y larga supervivencia. Esta heterogeneidad subraya la necesidad de herramientas que permitan identificar y analizar estas diferencias en los estudios de supervivencia. Este trabajo tiene como objetivo presentar la implementación de un ecosistema de ciencia de datos basado en R para analizar la existencia de subpoblaciones de larga supervivencia en pacientes con cáncer. Se basa en una metodología que permite la identificación de subpoblaciones mediante pruebas de multimodalidad y el ajuste de modelos paramétricos de mezcla de supervivencia. El ecosistema utiliza diversos paquetes de R, como RMarkdown, y se aplica a un conjunto de 1245 pacientes con cáncer de pulmón avanzado tratados con CIMAvaxEGF en ensayos clínicos del Centro de Inmunología Molecular. Los resultados demuestran que el ecosistema es eficaz para realizar análisis completos, desde la carga de datos hasta la visualización de resultados. Se identificaron dos subpoblaciones: una con corta supervivencia (73%, mediana de 8,7 meses) y otra con larga supervivencia (27%, mediana de 20,9 meses). Esto muestra la utilidad del enfoque para caracterizar la heterogeneidad en la respuesta al tratamiento. En conclusión, el ecosistema desarrollado es una herramienta versátil, reproducible y abierta para el análisis de supervivencia en cáncer de pulmón avanzado. Futuras investigaciones podrían extender su aplicación a otros tipos de cáncer e incorporar metodologías adicionales para mejorar la caracterización de subpoblaciones.

Palabras Clave: cáncer, ciencia de datos, larga supervivencia.

Abstract *The introduction of new immunotherapies has increased the life expectancy of cancer patients, although treatment response varies significantly, with subpopulations showing short and long survival. This heterogeneity underscores the need for tools that enable the identification and analysis of these differences in survival studies. This work aims to present the implementation of an R-based data science ecosystem to analyze the existence of long-survival subpopulations in cancer patients. It is based on a methodology that allows the identification of subpopulations through multimodality tests and the fitting of parametric survival mixture models. The ecosystem utilizes various R packages, as RMarkdown, and is applied to a dataset of 1245 advanced lung cancer patients treated with CIMAvaxEGF in clinical trials at the Center of Molecular Immunology. The results demonstrate that the ecosystem is effective for conducting comprehensive analyses, from data loading to result visualization. Two subpopulations were identified: one with short survival (73%, median of 8.7 months) and another with long survival (27%, median of 20.9 months). This shows the utility of the approach for characterizing heterogeneity in treatment response. In conclusion, the developed ecosystem is a versatile, reproducible, and open tool for survival analysis in advanced lung cancer. Future research could extend its application to other cancer types and incorporate additional methodologies to improve the characterization of subpopulations.*

Keywords: *cancer, data science, long-term survival.*

Mathematics Subject Classification: 92-04, 62P10, 62N01, 62N02.

[1]Dirección de investigaciones clínicas, Centro de Inmunología Molecular, La Habana, Cuba. Email: jorge@cim.sld.cu.
[2]Dirección de investigaciones clínicas, Centro de Inmunología Molecular, La Habana, Cuba. Email: patricial@cim.sld.cu.
[3]Dirección de investigaciones clínicas, Centro de Inmunología Molecular, La Habana, Cuba. Email: lsanchez@cim.sld.cu.
***Autor para Correspondencia (*Corresponding Author*)**
Editado por (Edited by): Damian Valdés Santiago, Facultad de Matemática y Computación, Universidad de La Habana, Cuba.

Citar como: Palomino Hernández, J.L.; Lorenzo-Luaces Álvarez, P.;& Sánchez Valdés, L. (2024). Ecosistema de ciencia de datos para el análisis de largos supervivientes en cáncer. *Ciencias Matemáticas*, 38(1), 69–76. DOI: https://doi.org/10.5281/zenodo.15046727. Recuperado a partir de https://revistas.uh.cu/rcm/article/view/10012.

Introducción

El cáncer en estadio avanzado representa uno de los mayores desafíos en la oncología moderna, tanto por su complejidad biológica como por su impacto en la calidad de vida de los pacientes. Los pacientes de cáncer avanzado, por lo general, tienen menos oportunidades de tratamiento y un tiempo de supervivencia más limitado. Sin embargo, dentro de este grupo existe una subpoblación de pacientes que, a pesar de su diagnóstico, presentan una supervivencia significativamente más larga que la media. Estos casos ofrecen una ventana única para comprender los factores que pueden influir en la progresión de la enfermedad. El análisis de estas subpoblaciones puede contribuir al desarrollo de terapias más efectivas y personalizadas, lo que mejora el pronóstico de otros pacientes.

El grupo de investigación al que pertenecen los autores ha adoptado una metodología que explora la presencia de multimodalidad en los datos de supervivencia y el ajuste de un modelo de mezcla de supervivencia [15].

La multimodalidad y los modelos de mezcla de supervivencia ofrecen enfoques complementarios. Por una parte, las pruebas de multimodalidad permiten señalar la existencia de varios picos en la distribución de los datos, lo que puede sugerir la presencia de subpoblaciones con diferentes patrones de supervivencia [11]. Mientras que los modelos de mezcla de supervivencia, por otro lado, proporcionan una herramienta estadística para descomponer estos datos en subpoblaciones discretas, permitiendo un análisis más detallado y preciso [6].

Muchas de las herramientas que comúnmente se utilizan para el análisis estadístico están limitadas a procedimientos de análisis de supervivencia más básicos, que no tienen en cuenta la presencia de varias subpoblaciones. Herramientas de carácter comercial ampliamente utilizadas como el software de IBM SPSS Statistics [10] no incluyen métodos para el análisis de multimodalidad [9], ni modelos paramétricos de mezcla de distribuciones de supervivencia. Este tipo de herramientas comerciales son más restrictivas respecto a otros entornos de código abierto y libres de costos como R [7], ya que no permiten la adición de manera libre de nuevas funcionalidades.

Por otro lado, los ecosistemas de código abierto se encuentran en continuo crecimiento, con gran cantidad de bibliotecas que aportan nuevos métodos de análisis, gracias a la gran red de usuarios y desarrolladores que tienen a nivel global [2]. Estos sistemas son muy versátiles, lo que permite a los analistas añadir las funcionalidades necesarias, según el tipo de análisis que necesiten realizar. Además, permiten realizar en un mismo ambiente las distintas etapas del proceso de análisis; desde la preparación de los datos hasta la visualización de los resultados. Por su naturaleza de código abierto estos entornos fomentan la colaboración, reproducibilidad y la transparencia en los procesos de análisis.

Este trabajo se traza como objetivo presentar la implementación de un ecosistema de ciencia de datos basado en R para el análisis de supervivencia en presencia de subpoblaciones de supervivencia corta y larga en pacientes con cáncer, así como su aplicación sobre un conjunto de datos integrado de pacientes de cáncer de pulmón avanzado tratados con la vacuna CIMAVaxEGF [4], incluidos en ensayos clínicos promovidos por el Centro de Inmunología Molecular.

Relevancia del estudio

La investigación sobre un ecosistema de ciencia de datos para el análisis de largos supervivientes en cáncer resulta crucial para mejorar la comprensión de esta población específica. Contar con un ecosistema reproducible de análisis basado en un entorno abierto permite realizar mejores y más eficientes análisis sobre este tipo de pacientes. Este ecosistema puede ser actualizado con nuevos métodos y paquetes de manera sencilla, lo que potencia la versatilidad y reutilización del mismo. Esta herramienta ayudará a identificar patrones y factores que contribuyen a la supervivencia prolongada, facilitando la personalización de tratamientos y políticas de salud pública.

1. Materiales y métodos

1.1 Descripción del ecosistema de ciencia de datos

En esta sección se presenta el ecosistema de ciencia de datos implementado. Se describe cada etapa de análisis y las herramientas utilizadas para ello. Este ecosistema incluye la preparación de los datos, análisis descriptivos, métodos para el análisis de multimodalidad, el ajuste de modelos de mezcla de supervivencia y la visualización de los resultados. En la Figura 1 se muestra un diagrama del ecosistema implementado.

1.1.1 Preparación de los datos

La preparación de los datos puede incluir, entre otras tareas, la integración de distintos conjuntos de datos, la selección de variables relevantes para el análisis o la creación de nuevas variables. Dentro de las variables a seleccionar para el análisis se encuentran las fechas en que fue diagnosticado el paciente con cáncer, la fecha de inclusión en el ensayo (si se analizan pacientes incluidos en investigaciones clínicas), la última fecha conocida del paciente o su fecha de fallecimiento.

Entre las variables que se deben crear está el tiempo de supervivencia y el estado vital del paciente al momento de cierre del estudio, indicándose que el paciente falleció (estado = 1) o que el paciente continua vivo, o sea que está censurada la observación (estado = 0).

El tiempo de supervivencia se calcula tomando como referencia la fecha de diagnóstico o de inclusión en el estudio, la última fecha en la que se tuvo información del paciente y, en caso de fallecimiento, la fecha de este evento. Por otro lado, el estado vital del paciente puede determinarse a partir de la fecha de fallecimiento. Sin embargo, en algunos estudios clínicos, esta información ya se ha recopilado durante el desarrollo de la investigación. En tales casos, no es necesario volver a determinar esta variable.

Para la preparación de los datos se utiliza el paquete `dplyr` [21]. Este paquete proporciona una interfaz de programación para manipular y analizar datos en R, así como

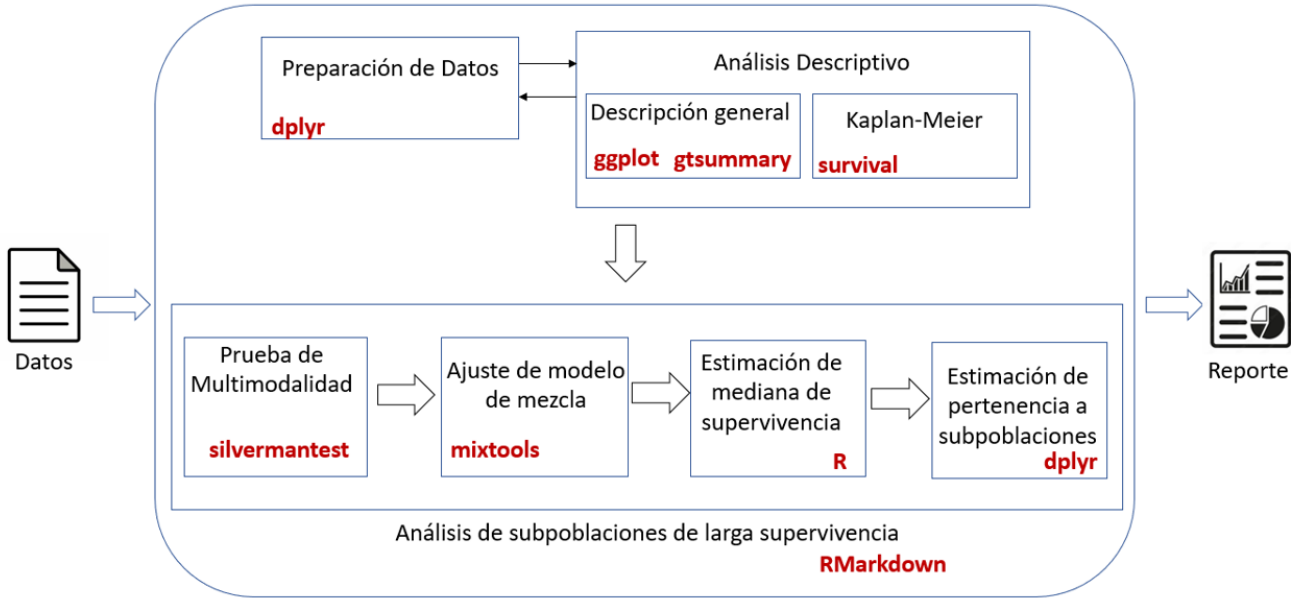

Figura 1. Ecosistema de ciencia de datos utilizado para el análisis de subpoblaciones de larga supervivencia [*Data science ecosystem used for long-term survival subpopulation analysis*].

permite realizar operaciones comunes de transformación de datos de forma eficiente.

1.1.2 Análisis descriptivo

El análisis descriptivo constituye el primer paso de análisis y tiene como objetivo proporcionar una visión general de los datos. En esta etapa se pueden visualizar, a través de gráficos y tablas, elementos clave como la distribución por edades, género, y otros factores demográficos y clínicos que son relevantes para el estudio.

Como parte del análisis descriptivo se realizan gráficos de densidad del tiempo de supervivencia de los pacientes y el ajuste de las curvas de supervivencia por el método de Kaplan-Meier [5]. Los gráficos de densidad ayudan a visualizar la distribución completa del conjunto de datos analizado y detectar la presencia de multimodalidad de manera visual. Por otra parte, el análisis de Kaplan-Meier permite visualizar la presencia de una meseta al final de la curva. Esto puede indicar la posible existencia de subgrupos de larga supervivencia.

Para implementar esta etapa se utilizan los siguientes paquetes: `gtsummary` [18] se usó para visualizar en forma de tabla las características del conjunto de datos analizado. `ggplot` [20] permite realizar diferentes tipos de gráficos y puede ser utilizado para realizar el gráfico de densidad del tiempo de supervivencia. El paquete `survival` [19] se utilizó para realizar el ajuste de las curvas de supervivencia por el método de Kaplan-Meier.

1.1.3 Prueba de multimodalidad

Antes de aplicar el modelo paramétrico de mezcla de distribuciones de supervivencia, se evalúa la existencia de multimodalidad en los datos, lo que justifica la aplicación del modelo. Para este análisis se aplicó el método propuesto por Silverman [17] y adaptado por Hall y York [8].

El objetivo es comprobar si la verdadera distribución de una variable en una población es unimodal o multimodal. Formalmente, dada una muestra de una variable aleatoria con función de densidad f (en este caso el tiempo de supervivencia), denotando por j el número de modos en f, la hipótesis que debe comprobarse viene dada por:

$$H_0 : j \leq k, \quad H1 : j > k, \tag{1}$$

donde k es el número de componentes del modelo de mezcla (o el número de modas de los datos).

Este análisis se realiza mediante el paquete `silverman-test` [14]. Este paquete verifica el número de modas en una densidad empírica usando el método de Silverman [17]. Este paquete también presenta otras funcionalidades de interés como la visualización de los resultados, la verificación de multimodalidad para varios parámetros, la visualización de múltiples densidades para un número determinado de modas y el cálculo del ancho de banda crítico [14].

La función `silverman.plot` permite visualizar los resultados de las pruebas de hipótesis implementadas, considerando valores de k, desde $k = 1$ hasta $k = 5$. El primer valor de k para el cual se rechaza la hipótesis coincide con el

número de modas de los datos.

1.1.4 Ajuste del modelo de mezcla de supervivencia

En este paso se ajusta un modelo paramétrico de mezcla de supervivencia al conjunto de datos con el propósito de estimar los parámetros de las subpoblaciones. Para ello, se utiliza la función `weibullRMM_SEM` del paquete `mixtools` [3]. Esta función asume la mezcla de subpoblaciones con distribución Weibull [12]. El resultado del algoritmo incluye estimaciones de las proporciones de mezcla y de los parámetros de escala y forma de las distribuciones Weibull, así como la probabilidad posterior de pertenencia a cada subpoblación. Para el uso de este algoritmo se fija el número de iteraciones, en este caso se utilizaron 10000 iteraciones.

1.1.5 Estimación de la mediana de supervivencia

En la interpretación clínica de los resultados, más importante que los parámetros de las distribuciones resulta la estimación de la mediana de supervivencia de cada subpoblación. Las medianas del tiempo de supervivencia para cada subpoblación se calcula utilizando los parámetros de forma (β) y escala (λ) estimados por el modelo de mezcla de supervivencia a partir de la siguiente fórmula:

$$\text{Mediana} = \lambda \cdot (\ln(2))^{\frac{1}{\beta}}.$$

1.1.6 Pertenencia a las subpoblaciones

La pertenencia a cada subpoblación se determina a partir de las probabilidades *a posteriori* de pertenecer a cada subpoblación que se obtienen del modelo de mezcla de supervivencia. Finalmente, se calcula el porcentaje de pacientes que pertenecen a cada subpoblación.

Para la implementación de todo este flujo de análisis y la presentación de los resultados se utilizó la herramienta RMarkdown [1]. RMarkdown es un entorno de trabajo que facilita la creación de documentos que combinan salidas gráficas y de texto con el código que las genera. De esta manera se puede realizar el análisis en un mismo entorno, documentarlo y exportar estos resultados en archivos en formato html, pdf y docx. El análisis realizado puede ejecutarse más de una vez obteniendo el mismo resultado, siempre que no cambien los datos fuentes [13, 22] (Figura 2).

1.2 Aplicación a los datos

El ecosistema descrito para el análisis de subgrupos de larga supervivencia se aplicó a un conjunto de datos integrado de pacientes incluidos en ensayos clínicos promovidos por el Centro de Inmunología Molecular. Este conjunto contiene 1245 pacientes en estadio IIIB o IV, tratados con la vacuna CIMAVaxEGF. Los pacientes han sido incluidos en ensayos clínicos entre los años 2002 y 2022.

2. Resultados

Las principales características demográficas y clínicas de los pacientes se presentan en la Tabla 1. En esta tabla se puede observar que en el conjunto de datos predominan pacientes masculinos (64 %), de piel blanca (71 %) y con estadio IV (52 %). Existen en el conjunto de datos 179 pacientes vivos, para un 14 % de datos censurados.

Característica	**N (%)**
Edad (años)	65 (58, 72) [Mediana (RIC)]
Sexo	
Femenino	434 (36 %)
Masculino	773 (64 %)
Desconocido	38
Color de Piel	
Blanca	840 (71 %)
Mestiza	192 (16 %)
Negra	153 (13 %)
Desconocido	60
Grupo Histológico	
Adenocarcinoma	429 (34 %)
Carcinoma Epidermoide	358 (29 %)
Otro	458 (37 %)
Estadio	
IIIB	596 (48 %)
IV	649 (52 %)
ECOG	
0	355 (29 %)
1	531 (44 %)
2	268 (22 %)
3	63 (5,2 %)
4	1 (<0,1 %)
Desconocido	27
Estado del Paciente	
Fallecido	1066 (86 %)
Vivo	179 (14 %)

Tabla 1. Características de la población de estudio (N = 1245) [*Characteristics of the study population (N = 1245)*].

El gráfico de densidad del tiempo de supervivencia calculado desde la inclusión a los ensayos clínicos hasta el fallecimiento o el fin del estudio se presenta en la Figura 3. En este gráfico se puede observar la presencia de un primer pico de densidad alto seguido de otro pico más pequeño alrededor del mes 24. Esto pudiera indicar la presencia de dos subpoblaciones en este conjunto de datos.

En el análisis de Kaplan-Meier se determinó que la mediana de supervivencia global del conjunto de pacientes con cáncer de pulmón avanzado es de 11 meses. Se puede observar en la Figura 4 la meseta larga y estable al final de la curva de supervivencia del conjunto total de datos analizados que sugiere, consistentemente con lo observado en la curva de densidad, la presencia de pacientes con larga supervivencia.

La prueba de multimodalidad mostró la existencia de dos modas, dado que $k = 2$ es el primer valor para el cual el $p-$valor resulta por encima del nivel de significación 0,05 (Figura 5).

De esta manera quedó demostrada la bimodalidad del conjunto de datos analizado y se utilizó un modelo de mezcla

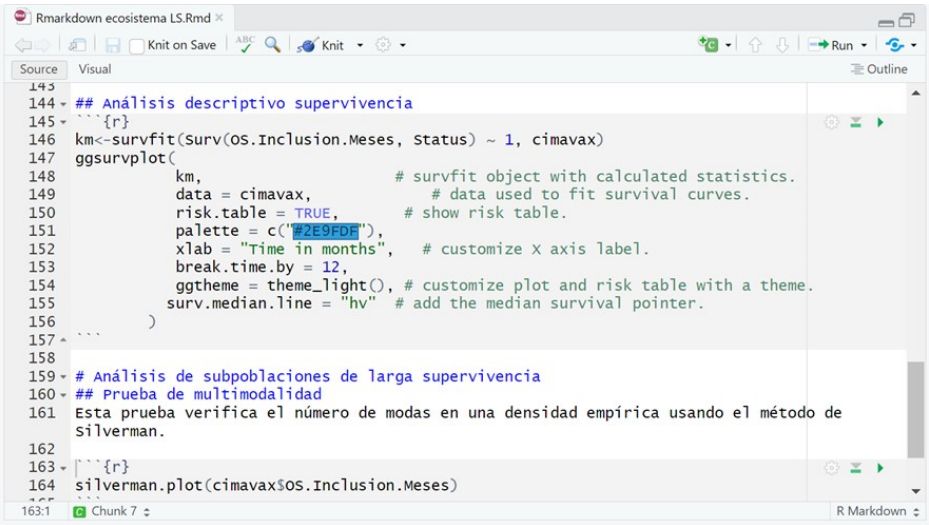

Figura 2. Vista de la implementación en RMarkdown para el análisis de subgrupos de larga supervivencia [*View of the implementation in RMarkdown for the analysis of long-term survival subgroups*]

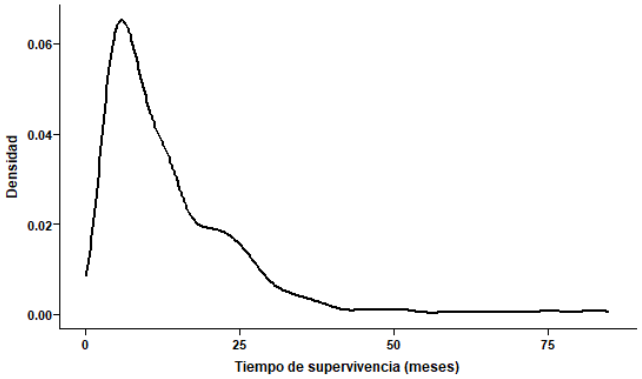

Figura 3. Gráfico de densidad respecto al tiempo de sobrevida [*Density graph versus survival time*].

de 2 componentes.

Mediante la función `weibullRMM_SEM`, se obtuvieron los parámetros de la distribución Weibull de las subpoblaciones y se calcularon las medianas de supervivencia y proporciones de pertenenciac como se describe en las secciones 1.1.5 y 1.1.6. En la Tabla 2 se muestran estos resultados.

Se estimó una mediana de supervivencia de 8,7 meses para la subpoblación de corta supervivencia, lo que representa el 73 % del grupo total, mientras que para la subpoblación de larga supervivencia se estimó una mediana de supervivencia de 20,9 meses, lo que representa el 27 %.

3. Discusión

El presente estudio implementó un ecosistema de ciencia de datos basado en R para el análisis de supervivencia en pacientes con cáncer de pulmón avanzado, identificando subpoblaciones con corta y larga supervivencia. Los resultados demostraron la existencia de dos subpoblaciones claramente diferenciadas: una con corta supervivencia (73 %, mediana de 8,7 meses) y otra con larga supervivencia (27 %, mediana de 20,9 meses). Estos hallazgos confirman la utilidad del enfoque propuesto para caracterizar la heterogeneidad en la respuesta al tratamiento, particularmente en pacientes tratados con la vacuna CIMAvaxEGF.

La identificación de subpoblaciones con diferentes patrones de supervivencia es crucial para el desarrollo de terapias personalizadas. En este sentido, el ecosistema implementado ofrece una herramienta versátil y reproducible que permite realizar un análisis completo, desde la preparación de los datos hasta la visualización de los resultados.

La metodología propuesta, basada en pruebas de multimodalidad y modelos paramétricos de mezcla de supervivencia, supera las limitaciones de las herramientas comerciales tradicionales, como IBM SPSS, que no incluyen funcionalidades para este tipo de análisis avanzado. Además, el uso de RMarkdown facilita la reproducibilidad y transparencia del proceso analítico.

La detección de una subpoblación de larga supervivencia es consistente con estudios previos [16], que han identificado subgrupos de pacientes con respuestas excepcionales a inmunoterapias, lo que resalta la importancia de continuar la investigación sobre los factores biológicos y clínicos que contribuyen a esta variabilidad en la supervivencia.

Una de las principales fortalezas de este trabajo es la integración de múltiples herramientas de código abierto en un único ecosistema, lo que permite un análisis robusto y flexible.

Sin embargo, es importante reconocer algunas limitaciones. En primer lugar, el estudio se centró en un conjunto de datos específico de pacientes con cáncer de pulmón avanzado tratados con CIMAvaxEGF, por lo que los resultados podrían no ser directamente extrapolables a otros tipos de cáncer o tratamientos.

Futuras investigaciones deberían explorar la aplicabili-

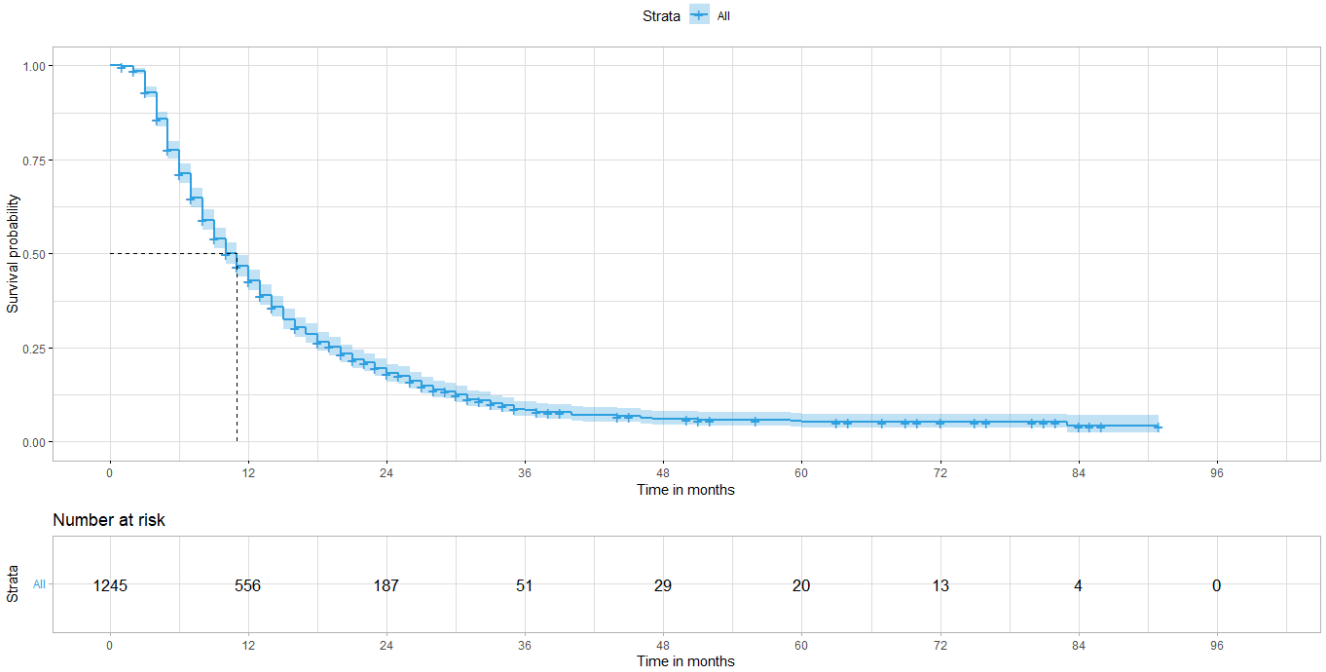

Figura 4. Curva de supervivencia del conjunto total de datos [*Survival curve of the total data set*].

Subpoblación de corta supervivencia					Subpoblación de larga supervivencia				
N_1	λ_1	β_1	Mediana	Pertenencia %	N_2	λ_2	β_2	Mediana	Pertenencia %
907	0,58	1,88	8,7	73	338	0,41	1,18	20,9	27

Tabla 2. Parámetros de las subpoblaciones de supervivencia [*Parameters of survival subpopulations*].

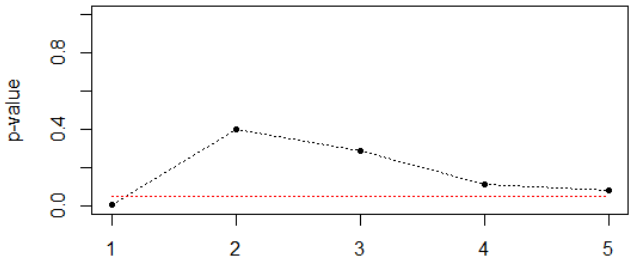

Figura 5. Resultado de las pruebas de multimodalidad de Silverman para distintos posibles k valores de modas [*Result of Silverman's multimodality tests for different values of k modes*].

dad de este enfoque en otros contextos oncológicos. Además, aunque el modelo de mezcla de Weibull utilizado es adecuado para capturar la heterogeneidad en los datos, podría ser beneficioso incorporar otros modelos paramétricos o no paramétricos para mejorar la precisión de las estimaciones.

4. Conclusiones

El ecosistema de ciencia de datos desarrollado en este estudio representa una herramienta valiosa para el análisis de supervivencia en cáncer, lo que permite la identificación de subpoblaciones con diferentes patrones de respuesta al tratamiento. Este enfoque no solo contribuye a una mejor comprensión de la heterogeneidad en la supervivencia, sino que también sienta las bases para el desarrollo de terapias más personalizadas y efectivas. Futuros trabajos podrían ampliar la aplicación de este ecosistema a otros tipos de cáncer e incorporar nuevas metodologías para una caracterización más detallada de las subpoblaciones identificadas.

Suplementos

Este artículo no contiene información suplementaria.

Conflictos de interés

Se declara que no existen conflictos de interés. No existen subvenciones involucradas en este trabajo.

Contribución de autoría

Conceptualización J.L.P.H.

Curación de datos J.L.P.H.

Análisis formal J.L.P.H.

Investigación P.L.L.A., J.L.P.H.

Metodología P.L.L.A., L.S.V., J.L.P.H.

Administración de proyecto P.L.L.A., L.S.V.

Software J.L.P.H.

Supervisión P.L.L.A., L.S.V.

Validación P.L.L.A., J.L.P.H.

Visualización J.L.P.H.

Redacción: preparación del borrador original J.L.P.H.

Redacción: revisión y edición P.L.L.A., L.S.V., J.L.P.H.

Referencias

[1] Allaire, J.J.: *rmarkdown: Dynamic Documents for R*, 2024. https://cran.r-project.org/web/packages/rmarkdown/index.html.

[2] Bansal, A. and S. Srivastava: *Tools used in data analysis: A comparative study*. International Journal of Recent Research, 5(1):15–18, 2018. https://www.ijrra.net/Vol5issue1/IJRRA-05-01-04.pdf.

[3] Benaglia, T., D. Chauveau, D.R. Hunter, and D.S. Young: *mixtools: an R package for analyzing mixture models*. Journal of Statistical Software, 32:1–29, 2010. https://www.jstatsoft.org/article/view/v032i06.

[4] CECMED: *CIMAvax(R)-EGF (Conjugado químico de Factor de Crecimiento Epidérmico humano recombinante acoplado a la proteína recombinante rP64K)*. https://www.cecmed.cu/registro/rcp/biologicos/cimavaxr-egf-conjugado-quimico-factor-crecimiento-epidermico-humano.

[5] Dudley, W.N., R. Wickham, and N. Coombs: *An introduction to survival statistics: Kaplan-Meier analysis*. Journal of the Advanced Practitioner in Oncology, 7(1):91, 2016.

[6] Farewell, V.T.: *The use of mixture models for the analysis of survival data with long-term survivors*. Biometrics, pages 1041–1046, 1982. https://pubmed.ncbi.nlm.nih.gov/7168793/.

[7] Foundation, R: *The R Project for Statistical Computing*, 2025. https://www.r-project.org.

[8] Hall, P. and M. York: *On the calibration of Silverman's test for multimodality*. Statistica Sinica, pages 515–536, 2001. https://www.jstor.org/stable/24306875.

[9] IBM: *Ibm spss advanced statistics*, 2024. https://www.ibm.com/docs/es/spss-statistics/27.0.0?topic=edition-advanced-statistics.

[10] IBM: *Ibm spss statistics*, 2024. https://www.ibm.com/products/www.ibm.com/products/spss-statistics.

[11] McLachlan, G.J. and D. Peel: *Finite mixture models*. John Wiley and Sons, 2000. https://onlinelibrary.wiley.com/doi/book/10.1002/0471721182.

[12] Mudholkar, G.S., D.K. Srivastava, and G.D. Kollia: *A generalization of the Weibull distribution with application to the analysis of survival data*. Journal of the American Statistical Association, 91(436):1575–1583, 1996.

[13] Peikert, A. and A.M. Brandmaier: *A reproducible data analysis workflow with R Markdown, Git, Make, and Docker*. Quantitative and Computational Methods in Behavioral Sciences, pages 1–27, 2021. https://qcmb.psychopen.eu/index.php/qcmb/article/view/3763.

[14] Preubner, J.: *silvermantest: Investigate the Number of Modes using Kernel Density Estimates*. https://rdrr.io/github/jenzopr/silvermantest/.

[15] Sanchez, L., P. Lorenzo-Luaces, C. Fonte, and A. Lage: *Mixture survival models methodology: an application to cancer immunotherapy assessment in clinical trials*. arXiv preprint, 2019. https://arxiv.org/abs/1911.09765.

[16] Sanchez, L., L. Muchene, P. Lorenzo-Luaces, C. Viada, P.C. Rodriguez, S. Alfonso, T. Crombet, E. Neninger, Z. Shkedy, and A. Lage: *Differential effects of two therapeutic cancer vaccines on short- and long-term survival populations among patients with advanced lung cancer*. Seminars in Oncology, 45(1):52–57, 2018, ISSN 0093-7754. https://doi.org/10.1053/j.seminoncol.2018.04.005.

[17] Silverman, B.W.: *Using kernel density estimates to investigate multimodality*. Journal of the Royal Statistical Society: Series B (Methodological), 43(1):97–99, 1981. https://www.jstor.org/stable/2985156.

[18] Sjoberg, D.D., K. Whiting, M. Curry, J.A. Lavery, and J. Larmarange: *Reproducible Summary Tables with the gtsummary Package*. The R Journal, 13:570–580, 2021. https://doi.org/10.32614/RJ-2021-053.

[19] Therneau, T.M. and P.M. Grambsch: *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000, ISBN 0-387-98784-3. https://link.springer.com/book/10.1007/978-1-4757-3294-8.

[20] Wickham, H.: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016,

ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.

[21] Wickham, H., M. Averick, J. Bryan, W. Chang, L.D'A. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, and J. Hester: *Welcome to the Tidyverse*. Journal of Open Source Software, 4(43):1686, 2019. https://joss.theoj.org/papers/10.21105/joss.01686.

[22] Xie, Y., J.J. Allaire, and G. Grolemund: *R markdown: The definitive guide*. Chapman and Hall/CRC, 2018. https://bookdown.org/yihui/rmarkdown/.