Articulo original

Análisis de una serie temporal en un centro de atención terciaria del linfoma no Hodgkin

Analysis of a Time Series in a Tertiary Care Center for Non-Hodgkin Lymphoma

Kali Cepero Llauger[1*] https://orcid.org/0000-0001-8159-8953
Ibis Karina Pardo Ramírez[1] https://orcid.org/0000-0002-5711-5971
Roberto León Castellón[1] https://orcid.org/0000-0002-6085-8565

[1]Hospital Clínico Quirúrgico Docente Hermanos Ameijeiras. La Habana, Cuba.

[*]Autor para la correspondencia**:** kalicep@infomed.sld.cu

RESUMEN

Introducción: La incidencia del linfoma no Hodgkin se incrementa entre 1 y 2 % anualmente. Ordenar cronológicamente su incidencia mensual permite analizar características inherentes a una serie temporal.

Objetivo: Determinar las características que distinguen una serie temporal de 11 años de linfoma no Hodgkin.

Métodos: Se realizó un estudio descriptivo en el Hospital Clínico Quirúrgico Hermanos Ameijeiras en el período de enero de 2011 a diciembre de 2021, la muestra estuvo conformada por 132 pacientes diagnosticados con la enfermedad en el Servicio de Hematología. Se realizó un análisis espectral y de las autocorrelaciones. Se utilizaron modelos tradicionales y ARIMA para los datos. Se realizaron pronósticos con los modelos de *machine learning*.

Resultados: En el gráfico de la secuencia no se observó tendencia y la repetición de los picos sugiere estacionalidad, en el análisis espectral muestra que ocurre a los 12 meses y en el análisis aditivo del componente estacional. El ruido de la serie tiene una distribución normal. Los algoritmos de *machine learning* disminuyen el error de los pronósticos.

Conclusiones: Se determinó las características que distinguen una serie temporal en una serie de tiempo no aleatoria de carácter estacionario en media, con alguna variabilidad de varianzas, con estacionalidad, donde el error en la extrapolación pronóstico disminuye utilizando algoritmos de *machine learning*.

Palabras claves: linfoma no Hodgkin; series temporales; aprendizaje de máquina.

ABSTRACT

Introduction: The incidence of non-Hodgkin lymphoma increases between 1 and 2% each year. Ordering its monthly incidence chronologically allows it to analyze characteristics inherent to a time series.

Objective: To determine the characteristics that distinguish an 11-year series of non-Hodgkin lymphoma.

Methods: A descriptive study was carried out at Hermanos Ameijeiras Clinical Surgical Hospital from January 2011 to December 2021. The sample consisted of 132 patients diagnosed with the disease in the Hematology Service. A spectral and autocorrelation analysis was performed. Traditional and ARIMA models were used for the data. Forecasts were made with machine learning models.
Results: No trend was observed in the sequence graph and the repetition of the peaks suggests seasonality, which is demonstrated in the spectral analysis that occurs at 12 months. Through the additive analysis of the seasonal component, it was found that the highest positive seasonal factors belong to August (1.223) and December (0.969). The noise of the series has a normal distribution and machine learning algorithms reduce the forecast error.
Conclusions: The characteristics that distinguish a time series from a non-random time series of a stationary nature on average, with some variability of variances, with seasonality, where the error in forecast extrapolation decreases using machine learning algorithms, were determined.
Keywords: non-Hodgkin lymphoma; temporal series; machine learning.

Recibido: 27/08/2023
Aceptado: 03/10/2023

Introducción

En el diagnóstico de las enfermedades malignas como el cáncer de mama, la tiroides, la próstata, el melanoma y las hemopatías, existen en ellas variaciones estacionales que pueden responder a fenómenos biológicos y/o administrativos según el sitio de estudio.[1]
El comportamiento estacional en el diagnóstico de las leucemias y el linfoma de Hodgkin han sido estudiado por alrededor de 25 años, la mayoría de estos estudios han encontrado una incidencia en los meses de primavera en contraste con los estudios en el linfoma no Hodgkin (LNH) que han sido limitados.[2,3]
A partir del reducido número de estudios de series temporales de LNH, modelamos su incidencia mensual en un período de 11 años en el Hospital Clínico Quirúrgico Docente Hermanos Ameijeiras, con la finalidad de determinar las características que la distinguen y su pronóstico.

Métodos

Se realizó un estudio descriptivo en el Hospital Clínico Quirúrgico Hermanos Ameijeiras en el período comprendido entre enero de 2011 a diciembre de 2021. Se conformó una muestra con todos los pacientes diagnosticados con el linfoma no Hodgkin en el Servicio de Hematología para un total de 132 pacientes.
Se definieron las fechas correspondientes al diagnóstico y se representaron en un gráfico de secuencia para observar cómo fue la evolución de la serie a lo largo del tiempo para tener una

apreciación inicial de la tendencia, la estacionalidad, la heterocedasticidad, para posteriormente determinar si es posible o no realizar los pronósticos.

Se realizó también un análisis espectral y de las autocorrelaciones para determinar con más certeza si existe o no estacionalidad, para posteriormente realizar su análisis por el método correspondiente, a partir de la información visual del gráfico de secuencia generado (fluctuaciones estacionales y tendencia) y el análisis de dispersión por nivel. Para los pronósticos se tuvo en cuenta los modelos tradicionales y ARIMA a fin, para seleccionar el que mejor ajuste a la data; además, se realizaron pronósticos con modelos de *machine learning*, los cuales son autoadaptativos y se basan en pocas suposiciones *a priori* acerca del problema en estudio. Se analizó el ruido de la serie original y el generado con el modelo pronóstico convencional.

Resultados

Se organizó en una serie cronológica con 132 observaciones. En la secuencia de la serie temporal se representó la evolución de la incidencia mensual de LNH durante 11 años; las líneas discontinuas marcan el comienzo de cada año y las de color verde representan la media (fig.1).

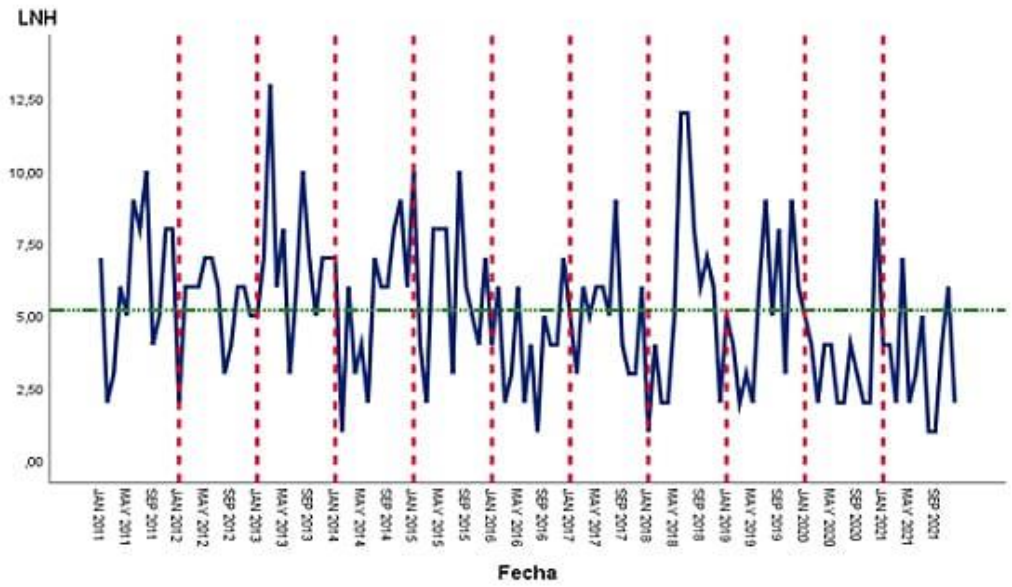


Fig. 1- Gráfico de secuencia para serie de tiempo linfoma no Hodgkin.

El análisis preliminar de esta representación sugiere que la incidencia mensual de LNH en el tiempo y lugar de estudio tienen un carácter estacionario en media, con alguna variabilidad de varianzas, no se observa tendencia y la repetición de los picos sugiere estacionalidad. Se trata de una serie de tiempo no aleatoria con una prueba de rachas donde el estadístico de contraste es igual a 50,000 con un valor p = 0,006. No existe una clara tendencia creciente o decreciente de la media ni de la mediana en función del tiempo.

En la función de la autocorrelación se puedo observar rezagos significativos, lo cual indica que los datos no son aleatorios y el estadístico de Box-Ljungpara referente a las autocorrelaciones tiene un valor $p < 0,05$. No existe un decaimiento lento y progresivo de los rezagos, lo cual hace más fuerte su carácter estacionario. En cuanto a las autocorrelaciones parciales se observan dos que superaron los límites de confianza (la primera y la doceava), lo cual sugiere con mayor fuerza la existencia de estacionalidad. (fig. 2 ab)

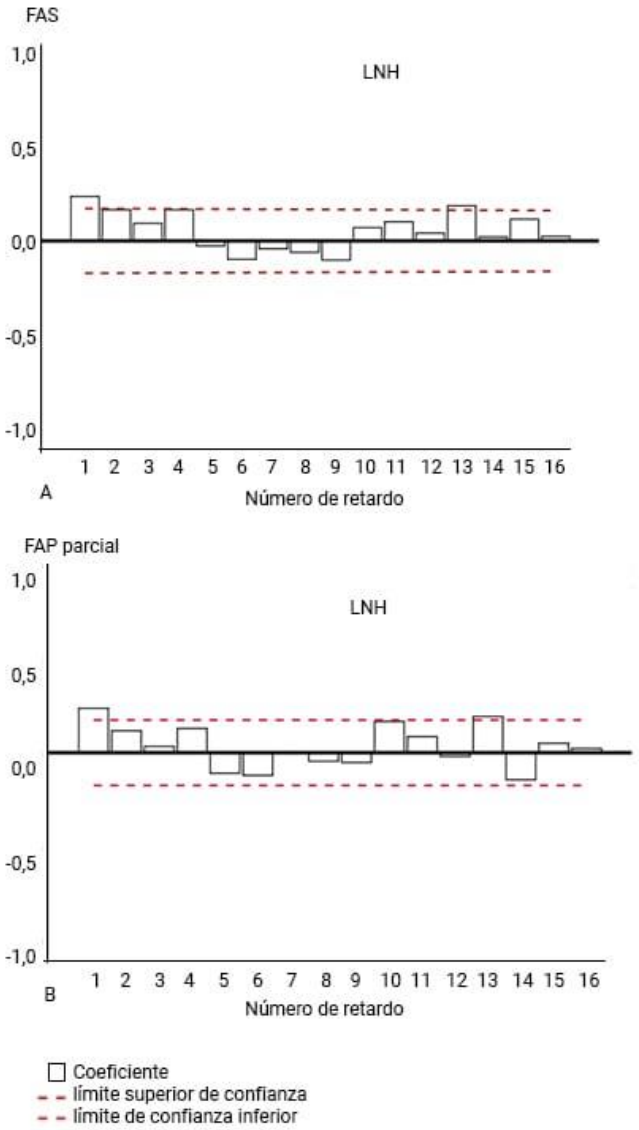
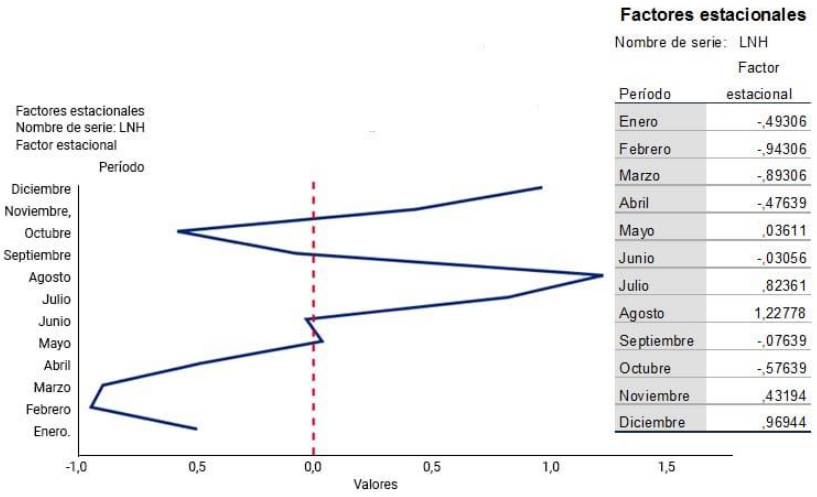


Fig. 2- a). Autocorrelaciones b). Autocorrelaciones parciales.

En el análisis espectral existe un primer pico indicativo de ciclicidad igual a 50 meses y un segundo pico que indica estacionalidad a los 12 meses.

Se realizó un análisis de dispersión por nivel, donde no se apreció dependencia entre variabilidad y nivel, lo cual implica que los componentes de esta serie de tiempo se combinan de forma aditiva.

Se observa la representación del componente estacional de la serie (líneas verdes) y el análisis gráfico y cuantitativo de los factores estacionales por el método aditivo. Existe en esta serie variabilidad estacional. El valor 0 indica que no existe estacionalidad; cuando es positivo, el valor de la variable toma valores superiores a los de la media en ese período; si es negativo, el valor de la variable toma valores inferiores a los de la media en ese período. En esta serie existe estacionalidad, con un incremento de los ingresos por LNH por sobre la media del período, más marcado en los meses de agosto y diciembre (fig. 3).



Factores estacionales

Nombre de serie: LNH

Período	Factor estacional
Enero	-,49306
Febrero	-,94306
Marzo	-,89306
Abril	-,47639
Mayo	,03611
Junio	-,03056
Julio	,82361
Agosto	1,22778
Septiembre	-,07639
Octubre	-,57639
Noviembre	,43194
Diciembre	,96944

Fig. 3- Componente estacional serie linfoma no Hodgkin.

El componente residual o ruido de la serie tiene una distribución normal (prueba de Kolmogorov-Smirnov = 0,046; p = 0,200).

Se originaron las predicciones hasta diciembre de 2025 mediante un modelo estacional simple, que fue el que mostró mejor ajuste, donde la mayor incidencia de casos de LNH continúa ocurriendo en los meses de agosto y diciembre, con una incidencia aproximada entre 4 y 5 respectivamente, lo que puede llegar en el mes agosto como máximo a 10 ingresos y 9 en el mes de diciembre (fig. 4).

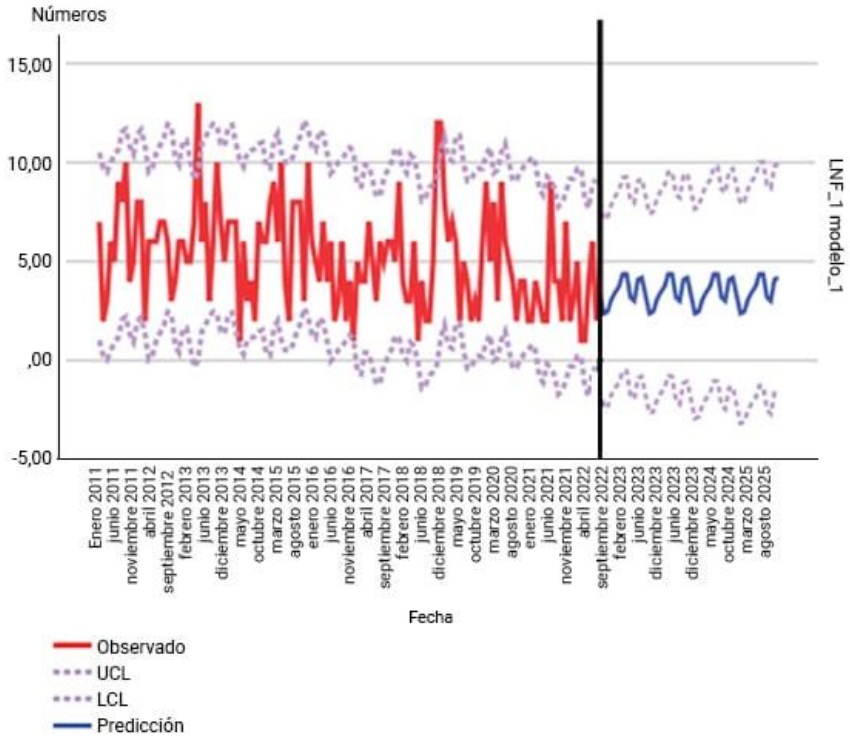


Fig.4- Pronóstico de ingresos por linfoma no Hodgkin hasta 2025.

El ruido de este modelo es aleatorio, con prueba de rachas (p = 0,861) y de Ljung-Box (p = 0,057) y sigue una distribución normal (prueba de Kolmogorov-Smirnov = 0,069; p = 0,200). (Ruido blanco gaussiano).

Al emplear algoritmos de *machine learning* en la modelación y pronóstico de esta serie temporal se observa que el error del modelo disminuye en un algoritmo de regresión lineal (MAE = 1,1601) en la figura 5A, y en el de *Support Vector Machine* (MAE = 0,9449) en la figura 5C, y con relación al error del modelo estacional simple (MAE = 1,382) que aparece anteriormente en la figura 4, los pronósticos fueron muy similares en estos tres modelos; por lo que se mantiene la estacionalidad en el pronóstico.

En el caso del perceptrón multicapa el MAE aumenta a medida que se realizan los pronósticos en el tiempo, llegando a ser de 1,556 para el último pronóstico; además de no conservar el patrón estacional de la serie original; lo cual no logra corregirse agregando capas, ni modifica la arquitectura de este; lo que pronostica instancias negativas (fig. 5 abc).

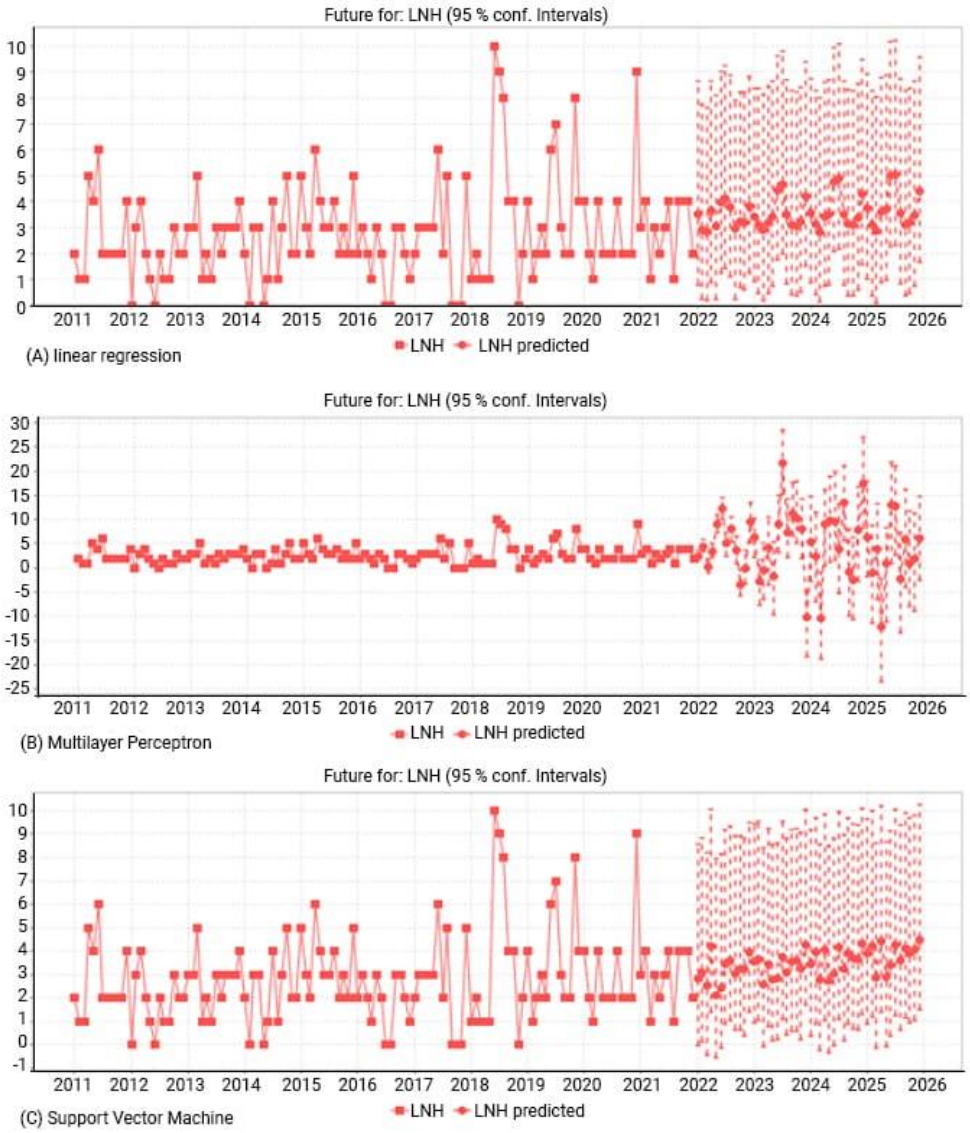


Fig. 5- Modelación y pronósticos con algoritmos de *machine learning.*

Discusión

Para la mayoría de las enfermedades oncológicas existe una distribución uniforme en su detección anual, pero en algunos casos hay un incremento del diagnóstico en algunas épocas del año.[1]

Existen pocos reportes de análisis de serie temporal de linfoma no Hodgkin. Koutros y otros[2,3] encontraron un pico de incidencia en los meses de primavera (marzo y abril) que lo atribuyen a una etiología infecciosa viral por una probable activación inmune frecuente en esta época del año, que puede influir en la linfomagénesis, lo que difiere de lo encontrado en nuestro estudio donde el mayor número de casos se diagnosticó en los meses de agosto y diciembre.

La deficiencia de vitamina D durante el invierno por la no exposición al sol puede explicar una rápida expansión de un clon para el desarrollos del linfoma no Hodgkin que pudiera

diagnosticarse en los meses que siguen. Se cree que la vitamina D inhibe la proliferación del linfoma *in vitro*.[2]

En el presente estudio se incluyeron todos los pacientes con diagnóstico de linfoma no Hodgkin realizados en la institución, pero hay que señalar que en los pocos estudios realizados al respecto hay una asociación entre LNH de células B y enfermedades virales en los meses de primavera, no existiendo una variación estacional para el LNH de células T/NK.[3]

El análisis de series temporales se utiliza frecuentemente para la predicción, con la finalidad de tomar decisiones adecuadas en diferentes áreas de la salud. Por ejemplo, el uso de técnicas de series temporales para la predicción de tasas de fertilidad, morbilidad y mortalidad, es importante para los administradores de atención médica, ya que estos datos sirven como indicadores de salud de una sociedad.[4]

Los modelos de aprendizaje automático se han establecido en la última década como serios competidores de los modelos estadísticos clásicos en el área de pronóstico.[5]

Los modelos de redes neuronales artificiales (RNA) se ha estado aplicando en las últimas décadas como métodos alternativos a los problemas de predicción y clasificación, ya que son métodos autoadaptativos, basados en pocas suposiciones *a priori* acerca del problema en estudio.

En el análisis de series de tiempo la aplicación de las RNA es considerada como técnicas de regresión no paramétrica y no lineal, con la adaptabilidad, la generalización, el aprendizaje y la posibilidad de representar relaciones no lineales en contraposición de los modelos de Box-Jenkins o ARIMA, que suponen que las series de tiempo son generadas a partir de procesos lineales.[6]

Singhy y otros[7] utilizaron el algoritmo de soporte de vectores de máquina en la predicción de la pandemia de la COVID-19 para explorar el impacto en la identificación, los fallecidos y la recuperación, basado en datos de series de tiempo.

Batista y otros[8] realizaron el análisis de la tendencia de una serie de tiempo local y otra nacional de LNH en el período comprendido entre los años 1980 y 2014; por tanto, concluyeron que la tendencia al aumento de la incidencia de muerte en el LNH se observa en ambas series (Chapecó y Brasil), pero en Chapecó es mucho mayor. Este estudio no analiza el resto de las componentes de las series temporales mencionadas.

No se encontraron referencias acerca del empleo de algoritmos de *machine learning* para la modelación y el pronóstico de series temporales de LNH, tampoco referentes al análisis clásico.

Se concluye que la incidencia mensual de LNH desde enero de 2011 hasta diciembre de 2021 en el hospital en estudio describe una serie temporal no aleatoria, de carácter estacionario en media, con alguna variabilidad de varianzas, con estacionalidad y donde el error en la extrapolación pronóstico disminuye utilizando los algoritmos de *machine learning*.

Referencias bibliográficas

1. Lambe M, Blomqvist P, Bellocco R. Seasonal variation in the diagnosis of cancer: a study based on national cancer registration in Sweden. Br J Cancer. 2003;88(9):1358-60. DOI: https://dx.doi.org/10.1038/sj.bjc.6600901.

2. Chang ET, Clarke CA, Glaser SL. Making sense of seasonal fluctuations in lymphoma diagnosis. Leuk Lymphoma. 2007;48(2):223-4. DOI: https://dx.doi.org/10.1080/10428190601158662.

3. Koutros S, Holford TR, Hahn T, Lantos PM, McCarthy PL Jr, Risch HA, *et al*. Excess diagnosis of non-Hodgkin's lymphoma during spring in the USA. Leuk Lymphoma. 2007;48(2):357-66. DOI: https://dx.doi.org/10.1080/10428190601076799

4. Mehrmolaei S, Keyvanpour MR. TsP-SA: usage of time series techniques on healthcare data. Int J Electron Healthc. 2018;10(3):190. DOI: https://dx.doi.org/10.1504/ijeh.2018.10015340.

5. Ahmed NK, Atiya AF, Gayar NE, El-Shishiny H. An empirical comparison of machine learning models for time series forecasting. Econom Rev. 2010;29(5-6):594-621. DOI: https://dx.doi.org/10.1080/07474938.2010.481556

6. Menacho Chiok CH. Comparación de los métodos de series de tiempo y redes neuronales. An Cient. 2014;75(2):245. DOI: https://dx.doi.org/10.21704/ac.v75i2.960.

7. Singh V, Poonia RC, Kumar S, Dass P, Agarwal P, Bhatnagar V, *et al*. Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine. J Discrete Math Sci Cryptogr. 2020;23(8):1583-97. DOI: http://dx.doi.org/10.1080/09720529.2020.1784535.

8. Batista JDAL, Oliveira A, Barbato PR. Tendência de mortalidade por linfoma não Hodgkin emuma área de exposiçãoao glifosato: comparativo entre Chapecó-SC e o cenário nacional.| Mortalitytrendsby Non-Hodgkin Lymphoma in a glyphosateexposurearea: comparisonbetween Chapecó-SC. Revista Brasileira de Pesquisa em Ciências da Saúde. 2021;8(15):14-9. Disponible en: https://revistas.icesp.br/index.php/RBPeCS/article/view/1166/1153.

Contribuciones de los autores

Conceptualización: Kali Cepero Llauger, Ibis Karina Pardo Ramírez, Roberto León Castellón.
Curación de datos: Kali Cepero Llauger, Ibis Karina Pardo Ramírez.
Análisis formal: Roberto León Castellón.
Investigación: Kali Cepero Llauger, Ibis Karina Pardo Ramírez.
Metodología: Kali Cepero Llauger, Ibis Karina Pardo Ramírez.
Administración del proyecto: Kali Cepero Llauger.
Software: Kali Cepero Llauger, Ibis Karina Pardo Ramírez.
Supervisión: Kali Cepero Llauger.
Validación: Kali Cepero Llauger, Ibis Karina Pardo Ramírez.
Visualización: Kali Cepero Llauger.
Redacción del borrador original: Kali Cepero Llauger, Ibis Karina Pardo Ramírez, Roberto León Castellón.
Redacción, revisión y edición: Kali Cepero Llauger, Ibis Karina Pardo Ramírez, Roberto León Castellón.

Conflicto de intereses

Los autores declaran que no existe conflicto de intereses.