

# Comparación pronóstica entre *scores* de riesgo y la aplicación de redes neuronales para la predicción de la mortalidad a corto y mediano plazo en pacientes con insuficiencia cardíaca

## Prognostic Comparison between Risk Scores and Neural Networks to Predict Short- and Mid-Term Mortality in Acute Heart Failure

MARIA JIMENA GAMBARTE , CLAUDIO HIGA , FEDOR NOVO , GRACIANA MARIA CIAMBRONE ,  
OMAR DARIO TUPAYACHI VILLAGOMEZ , AGUSTINA GINESI , MARIA SOL DONATO , IGNACIO NOGUES, RAÚL ALFREDO BORRACCI\*

### RESUMEN

**Introducción:** En el contexto de la insuficiencia cardíaca (IC) existen *scores* de riesgo para evaluar la mortalidad por cualquier causa durante el primer año, con áreas bajo la curva ROC que oscilan entre 0,59 y 0,80.

**Objetivo:** Desarrollar y validar un modelo basado en algoritmos de redes neuronales (RN) destinado a mejorar el rendimiento de los modelos tradicionales para predecir mortalidad a corto y mediano plazo de pacientes con IC aguda.

**Material y métodos:** Se analizó una base de datos con 181 variables de 483 pacientes con IC aguda en un hospital de comunidad de la Ciudad de Buenos Aires (junio de 2005-junio de 2019). Se utilizaron 25 variables para calcular 5 modelos de riesgo validados para predecir la mortalidad a 30 días, 6 meses y un año: EFFECT, ADHERE, GWTG-HF, 3C-HF y ACUTE-HF.

**Resultados:** La edad media fue  $78 \pm 11,1$  años, 58% eran varones, el 35% de las IC eran de etiología isquémico necrótica, y la fracción de eyección media fue 52% (35-60). En término de discriminación a 30 días, fueron mejores el *score* EFFECT (ROC: 0,68) y el 3C-HF (ROC: 0,67) que el ACUTE-HF (ROC: 0,54). A los 6 meses y al año, el *score* EFFECT (ROC: 0,69 y 0,69) superó al ADHERE (ROC: 0,53 y 0,56) ( $p=0,011$  y  $p=0,003$ , respectivamente), y los *scores* EFFECT GWRG-HF (ROC: 0,68 y 0,66) y 3C-HF (ROC: 0,67 y 0,67) superaron al *score* ACUTE-HF (ROC: 0,53 y 0,56). De los algoritmos de RN los mejores resultados se obtuvieron con un perceptrón multicapa (PMC) con dos capas ocultas. Se usó una RN de arquitectura de capas 24-9-7-2 con los siguientes resultados: ROC: 0,82, valor predictivo negativo (VPN) 93,2% y valor predictivo positivo (VPP) 66,7% para mortalidad a 30 días; ROC: 0,87, VPN: 89,1% y VPP: 78,6% para mortalidad a 6 meses; y ROC: 0,85, VPN: 85,6% y VPP: 78,9% para mortalidad al año. En términos de discriminación, los algoritmos de RN superaron a los *scores* tradicionales ( $p < 0,001$ ). Los factores que obtuvieron  $\geq 50\%$  de importancia estandarizada para predecir la mortalidad a los 30 días fueron en orden descendente la creatinina sérica, la hemoglobina, la frecuencia respiratoria, la urea, el sodio, la edad y la presión arterial sistólica. Agregaron capacidad pronóstica la clase III-IV NYHA y la demencia para mortalidad a 6 meses, y la frecuencia cardíaca y la disfunción renal crónica para mortalidad al año.

**Conclusiones:** Los modelos con algoritmos de RN fueron significativamente superiores a los *scores* de riesgo tradicionales en nuestros pacientes con IC. Estos hallazgos constituyen una hipótesis de trabajo a validar con una mayor muestra de casos y en forma multicéntrica.

**Palabras claves:** Insuficiencia cardíaca, Pronóstico, Mortalidad, *Score* de riesgo, Redes neuronales, Inteligencia artificial, Mortalidad.

### ABSTRACT

**Background:** Heart failure (HF) risk scores to assess all-cause mortality during the first year have areas under the ROC curve (AUC) ranging between 0.59 and 0.80.

**Objective:** To develop and validate a neural network (NN) algorithm-based model to improve traditional scores' performance for predicting short- and mid-term mortality of patients with acute HF.

**Methods:** A prospective clinical database was analyzed including 483 patients admitted with diagnosis of acute HF in a coronary care unit community hospital of Buenos Aires, between June 2005 and June 2019. Among 181 demographic, laboratory, treatment and follow-up variables, only 25 were selected to calculate five acute heart failure risk scores aimed to predict 30-day, 6-month and 1-year mortality: EFFECT, ADHERE, GWTG-HF, 3C-HF, and ACUTE-HF.

**Results:** Mean age was  $78 \pm 11.1$  years, 58% were men, 35% had ischemic necrotic HF and median left ventricular ejection

REV ARGENT CARDIOL 2021;89:435-446. <http://dx.doi.org/10.7775/rac.es.v89.i5.20434>

VER ARTÍCULO RELACIONADO: Rev Argent Cardiol 2021;89:385-387. <http://dx.doi.org/10.7775/rac.es.v89.i5.20449>

Recibido: 18/02/2021 - Aceptado: 06/09/2021

**Financiamiento:** Los autores no recibieron financiamiento alguno para la realización de este estudio

**Este trabajo obtuvo el Premio Congreso de Cardiología 2020**

Hospital Alemán, Buenos Aires, Argentina

**Autor correspondiente:** María Jimena Gambarte

fraction was 52% (35-60). At 30 days, the EFFECT score (AUC:0.68) and the 3C-HF score (AUC: 0.68) showed better performance than the ACUTE-HF score (AUC: 0.54). At 6-month and 1-year follow-up, the EFFECT score (ROC: 0.69 and 0.69) outperformed the ADHERE score (AUC: 0.53 and 0.56), and EFFECT (AUC: 0.69 and 0.69), GWRG-HF (AUC = 0.68 and 0.66), and 3C-HF (AUC:0.67 and 0.67) scores outperformed the ACUTE-HF score (AUC:0.53 and 0.56). The best results with NN algorithms were obtained with a two-hidden layer multilayer perceptron. A 24-9-7-2-layer architecture NN was used with the following results: AUC: 0.82, negative predictive value (NPV) 93.2% and positive predictive value (PPV) 66.7% for 30-day mortality; AUC: 0.87, NPV: 89.1% and PPV: 78,6% for 6-month mortality; and AUC: 0.85, NPV: 85.6% and PPV: 78.9% for 1-year mortality. In terms of discrimination, NN algorithms outperformed all the traditional scores ( $p < 0.001$ ). For this algorithm, the most influential factors in descending order that scored  $\geq 50\%$  normalized importance to predict 30-day mortality were serum creatinine, hemoglobin, respiratory rate, blood urea nitrogen, serum sodium, age and systolic blood pressure. Also, NYHA functional class III-IV and dementia added prognostic capacity to 6-month mortality, and heart rate and chronic kidney disease to 1-year mortality.

**Conclusions:** The models with NN algorithms were significantly superior to traditional risk scores in our population of patients with HF. These findings constitute a working hypothesis to be validated with a larger and multicenter sample of cases.

**Key words:** Heart failure - Prognosis - Mortality - Risk score - Deep learning - Artificial intelligence

## Abreviaturas

<b>ACCF</b>	American College of Cardiology Foundation	<b>IC</b>	Insuficiencia cardíaca
<b>ADHERE</b>	Acute Decompensated Heart Failure National Registry	<b>OPTIMIZE-HF</b>	Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure
<b>AHA</b>	American Heart Association	<b>PMC</b>	Perceptrón multicapa
<b>EFFECT</b>	Enhanced Feedback for Effective Cardiac Treatment	<b>RN</b>	Redes neuronales
<b>ESCAPE</b>	Evaluation Study of Congestive Heart Failure and Pulmonary Artery Catheterization Effectiveness	<b>RIC</b>	Rango intercuartílico
<b>FBR</b>	Red de función de base radial	<b>ROC</b>	Receiver Operating Characteristic
<b>GWTG-HF</b>	Get With the Guidelines-Heart Failure	<b>VPN</b>	Valor predictivo negativo
<b>HANBAH</b>	Hemoglobin, age, sodium, blood urea nitrogen, atrial fibrillation, and high-density lipoprotein	<b>VPP</b>	Valor predictivo positivo

## INTRODUCCIÓN

La insuficiencia cardíaca (IC) es un enfermedad con alta prevalencia a nivel mundial asociada a una morbilidad considerable, altos costos y mal pronóstico a mediano plazo. Debido a esto, la evaluación del riesgo de la IC es importante para la toma de decisiones clínicas. Las guías de IC de la *American College of Cardiology Foundation / American Heart Association* (ACCF/AHA) mencionan la utilidad de los *scores* validados para estimar el riesgo de mortalidad en pacientes hospitalizados con IC (clase IIa, nivel de evidencia B). (1) No obstante, una reciente evaluación crítica sobre la utilidad de estos *scores* concluyó que su aplicación en la práctica clínica sigue siendo limitada. (2)

En la IC aguda, existen *scores* de riesgo con múltiples variables de uso clínico, con áreas bajo la curva ROC (*Receiver Operating Characteristic*) que oscilan entre 0,59 y 0,80 al evaluar la mortalidad por cualquier causa durante el primer año. (3-14) La evaluación de estos modelos, basada en cohortes de validación externa, mostró áreas ROC de 0,69-0,81 para el GWTG-HF (*Get With the Guidelines-Heart Failure*), (15-18) 0,69-0,70 para el EFFECT (*Enhanced Feedback for Effective Cardiac Treatment*), (19-20) 0,64-0,68 para el ADHERE (*Acute Decompensated Heart Failure National Regis-*

*try*) (18-19) y 0,74 para el OPTIMIZE-HF (*Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure*), (21) con evidencia en la mayoría de ellos de una calibración adecuada.

Una revisión sistemática sobre modelos de predicción en IC identificó 117 *scores* que incluyen 249 variables diferentes, siendo los niveles de urea y el sodio en sangre los predictores más importantes. (22) La predicción de la mortalidad fue mejor en los registros prospectivos, los cuales utilizaron una mayor cantidad de variables clínicas. En esta revisión, el área ROC promedio de todos los modelos en conjunto fue 0,66, con 0,71, 0,68 y 0,63 para aquellos que predecían mortalidad, hospitalización por IC o ambos, respectivamente.

Dado que la predicción de la mortalidad en pacientes con IC sigue siendo sólo moderadamente exitosa, se han propuesto algoritmos de inteligencia artificial para evaluar el riesgo en situaciones agudas. Recientemente, un modelo basado en aprendizaje profundo (*deep learning*) alcanzó un área ROC de 0,88 y 0,79 al predecir la mortalidad intrahospitalaria y a 12 meses por IC aguda (16). A pesar de los resultados anteriores, otro estudio contemporáneo demostró que si bien los algoritmos de aprendizaje automático superaron a la regresión logística para predecir los reingresos a 30 días por IC descompensada, las mejoras fueron solo margi-

nales, con áreas ROC entre 0,61 y 0,78, obteniendo el método de aprendizaje profundo y un algoritmo Naïve Bayes los mejores resultados. (23-26)

El objetivo del trabajo fue desarrollar y validar un modelo basado en algoritmos de redes neuronales (RN) destinado a mejorar el rendimiento de los modelos tradicionales para predecir la mortalidad a corto y mediano plazo (30 días, 6 meses y al año) de pacientes con IC aguda.

## MATERIAL Y MÉTODOS

Se analizó una base de datos con 483 pacientes ingresados en forma prospectiva con diagnóstico de IC aguda en la Unidad Coronaria de un hospital de comunidad de la Ciudad de Buenos Aires, durante el período comprendido entre junio 2005-junio 2019. La base de datos completa incluía 181 variables; entre ellas, datos demográficos, de laboratorio, de

imágenes, de tratamiento y de seguimiento. De estas, solo se utilizaron 25 variables para calcular cinco modelos de riesgo de IC aguda destinados a predecir la mortalidad a 30 días, 6 meses y al año.

Los modelos de predicción usados en el análisis fueron: EFFECT (4), ADHERE (3), GWTG-HF (5), 3C-HF (7) y ACUTE-HF (8). En la Tabla 1 se muestran las variables incluidas en cada cálculo. Dos modelos, el ESCAPE (*Evaluation Study of Congestive Heart Failure and Pulmonary Artery Catheterization Effectiveness*) (6) y el OPTIMIZE-HF (10,13,21) fueron excluidos del estudio debido a la ausencia de datos completos en nuestra base, como los valores de BNP al alta utilizados en el puntaje de riesgo ESCAPE, y en el caso del OPTIMIZE-HF, debido a que en dicho estudio los pacientes tenían que estar tratados con milrinona por inestabilidad hemodinámica durante 48-72 hs. El punto final primario fue la mortalidad por todas las causas a 30 días (intra-hospitalaria), 6 meses y al año.

De los 25 predictores incluidos en los cinco modelos tradicionales se excluyó la variable "raza negra" y los otros 24 se

**Tabla 1.** Variables incluidas en los scores de riesgo en insuficiencia cardiaca para predicción de mortalidad a 30 días, 6 meses y 1 año.

Variables	EFFECT	GWTG-HF	ADHERE	3C-HF	ACUTE-HF
Edad	X	X		X	X
Frecuencia cardíaca		X			
Frecuencia respiratoria	X				
Clase III-IV (NYHA)				X	
Presión arterial sistólica	X	X	X		
Hipertensión				X	
Urea	X	X	X		
Creatinina			X		X
Sodio	X	X			
Enf. Cerebrovascular*	X				X
Demencia	X				
EPOC	X	X			
Cirrosis Hepática	X				
Cáncer	X				
Hemoglobina/anemia	X			X	
Raza Negra		X			
DBT (daño órgano blanco)				X	
Insuficiencia renal crónica				X	
Fibrilación auricular				X	
No beta bloqueante				X	
No IECA				X	
FEVI baja				X	X
Enf. valvular Grave**				X	X
Ventilación no invasiva					X
Hospitalización previa					X

\*Incluye accidente cerebrovascular y ataque isquémico transitorio.

\*\*Incluye insuficiencia mitral moderada.

Enf: Enfermedad EPOC: enfermedad pulmonar obstructiva crónica. DBT: diabetes. IECA: Inhibidores del sistema renina angiotensina. FEVI: fracción de eyección ventricular izquierda.

utilizaron para testear en tres modelos basados en algoritmos RN: perceptrón multicapa (PMC) de una y dos capas ocultas, y una red de función de base radial (FBR). Para lograr dicho objetivo, la base de datos se dividió en 2 grupos: 70% para el testeo de los algoritmos RN y 30% para la validación. Un algoritmo RN es un tipo especial de regresión no lineal que presenta múltiples valores de mínimos locales; por lo tanto, cada vez que se ejecute el algoritmo de entrenamiento (*training*), este convergerá en un modelo diferente. Para elegir el mejor modelo, el entrenamiento se repitió 50 veces para cada modelo RN. Simultáneamente, cada vez que se probaron los modelos en la cohorte de validación, se registraron la precisión, el área ROC, el valor predictivo negativo (VPN) y el valor predictivo positivo (VPP). Solo se seleccionaron los modelos RN que mostraron el mejor poder de discriminación a fin de compararlos con las predicciones de los cinco modelos tradicionales.

Todos los modelos basados en algoritmos RN tipo PMC fueron implementados con el método de estandarización de las covariables en la capa de entrada, funciones de activación de tangente hiperbólica en capas ocultas, una función de activación *softmax* en la capa de salida y una función de error de entropía cruzada (*cross-entropy error function*). El modelo de FBR fue implementado también con el método de estandarización de las covariables en la capa de entrada, una función de activación de *Gaussiana* o *softmax* en la capa oculta, una función de activación "identidad" en la capa de salida y una función de error de suma de cuadrados (*sum of squares error function*).

### Análisis estadístico

Las variables categóricas se expresaron como frecuencias absolutas y porcentajes, mientras que las variables continuas como media, desviación estándar o mediana y rango intercuartílico (RIC). Se utilizó la prueba de bondad de ajuste de Kolmogorov-Smirnov para analizar las distribuciones normales. La comparación estadística de las áreas ROC con sus respectivos intervalos de confianza del 95% se realizó con la prueba de De Long. La calibración se evaluó con el método de  $\chi^2$  de Hosmer-Lemeshow. Todos los modelos fueron también comparados con respecto a sus variables predictoras mediante la realización de un análisis de conglomerados jerárquicos, a fin de identificar subgrupos que compartían los mismos predictores. El análisis estadístico y el modelado de RN se realizaron con el *software* IBM SPSS 23.0 Statistics (IBM Corporation, Armonk, NY). Para la comparación de Áreas bajo la Curva ROC se utilizó el test de De Long, utilizando el programa MEDCALC versión 20.015. Se consideró estadísticamente significativo un valor de  $p \leq 0,05$ .

### Consideraciones éticas

El estudio fue aprobado por el Comité de Ética Institucional, y debido al carácter observacional se prescindió de consentimiento informado.

## RESULTADOS

Las características basales de la población del estudio utilizadas para los cálculos de los modelos tradicionales y para los algoritmos RN (testeo y validación) se muestran en la Tabla 2.

La figura 1 muestra las curvas ROC y la tabla 3 resume el desempeño de los modelos de predicción de mortalidad a 30 días, 6 meses y un año para pacientes con insuficiencia cardíaca. En término de discriminación a 30 días, fueron mejores el *score* EFFECT que

el ACUTE-HF (De Long  $p = 0,041$ ) y el *score* 3C-HF que el ACUTE-HF ( $p = 0,047$ ). A los 6 meses y al año de seguimiento, el *score* EFFECT superó al ADHERE ( $p = 0,011$  y  $p = 0,003$ , respectivamente), mientras que el *score* EFFECT ( $p < 0,001$  y  $p < 0,001$ ), el GWTG-HF ( $p = 0,001$  y  $p = 0,006$ ) y el 3C-HF ( $p = 0,001$  y  $p = 0,002$ ) superaron al *score* ACUTE-HF, respectivamente en los mismos períodos de tiempo.

Con respecto a los algoritmos de RN, los mejores resultados se obtuvieron con el PMC de dos capas ocultas (arquitectura de capas 24-9-7-2). Las características de este modelo de RN se muestran en la Figura 2, y el rendimiento de este algoritmo se resume en la Tabla 4.

En términos de discriminación, los algoritmos RN superaron a todos los modelos tradicionales (De Long  $p < 0,001$ ) para mortalidad de todas las causas a 30 días, 6 meses y al año, respectivamente. Con respecto al resto de los algoritmos RN, las áreas ROC variaron entre 0,81 y 0,82, y 0,75 y 0,78 para el PMC de una capa oculta y el modelo basado en FBR, respectivamente para los mismos períodos de tiempo. (Figura 3)

La "importancia estandarizada independiente" (*independent normalized importance*) de las variables para el algoritmo RN se muestra en la Figura 4. Para este algoritmo, los factores más influyentes en orden descendente que obtuvieron  $\geq 50\%$  de "importancia estandarizada" para predecir la mortalidad a 30 días fueron: creatinina y urea sérica, hemoglobina, frecuencia respiratoria, concentración de sodio, edad y presión arterial sistólica. Además de estas variables, la clase III-IV de la NYHA y la demencia se asociaron con mayor mortalidad a 6 meses, y la frecuencia cardíaca y la disfunción renal crónica con la mortalidad al año.

Finalmente, se realizó un análisis de conglomerados jerárquico sobre las variables de predicción de todos los *scores*, a fin de identificar subgrupos de modelos con características similares. Los subgrupos creados se presentan en el dendrograma de la Figura 5 que muestra el vínculo promedio entre los *scores*. Basado en este dendrograma, el algoritmo RN mostró más similitudes con los *scores* ADHERE, EFFECT y GWTG-HF. Cuando sólo se consideraron en el análisis los factores más influyentes que obtuvieron una puntuación  $\geq 50\%$  para "importancia estandarizada", el algoritmo RN compartió el 100% de los predictores con el ADHERE, el 64% de los predictores de EFFECT y el 57% de GWTG-HF.

## DISCUSIÓN

En el presente estudio, el modelo basado en el algoritmo RN superó a los modelos tradicionales en la predicción de la mortalidad a corto y mediano plazo en pacientes con IC aguda. El modelo de PMC de dos capas ocultas no solo mejoró estadísticamente la discriminación global, sino que también mantuvo un buen desempeño con respecto a los VPN y VPP hasta el año de seguimiento. Este punto es fundamental, ya que la mayoría de los algoritmos de RN tienden a mejorar sus resultados basándose principalmente en un aumento del VPN

**Tabla 2.** Características basales de la población (n = 483)

Variables	n (%)
Edad en años (media $\pm$ DE)	78 $\pm$ 11,1
Sexo masculino	279 (57,8)
Disnea clase funcional III-IV (NYHA)	90 (18,6)
Hipertensión	438 (90,7)
Diabetes	102 (21,1)
Etiología	
Isquémica	169 (35,0)
Hipertensiva	123 (25,5)
Enfermedad valvular	96 (19,9)
Otras etiologías	95 (19,7)
Insuficiencia renal crónica	78 (16,1)
Accidente cerebrovascular	58 (12,0)
Enfermedad pulmonar crónica	77 (15,9)
Anemia	77 (15,9)
Fibrilación auricular	239 (49,5)
Insuficiencia mitral moderada/grave	75 (15,5)
Ventilación no invasiva	286 (59,2)
Demencia	36 (4,6)
Cáncer	71 (14,7)
Tratamiento beta bloqueante	239 (49,5)
Tratamiento con IECA	149 (30,8)
Urea en sangre en mg% (mediana y RIC)	51 (38-71)
Hemoglobina en g% (media $\pm$ DE)	12,9 $\pm$ 3,9
Creatinina sérica en mg% (media $\pm$ DE)	1,3 $\pm$ 0,96
Sodio plasmático en mEq/L (media $\pm$ DE)	136 $\pm$ 10,1
Presión sistólica al ingreso en mmHg (mediana y RIC)	142 (130-160)
Frecuencia cardíaca (lpm) al ingreso (media $\pm$ DE)	91 $\pm$ 25
Frecuencia respiratoria por min (media $\pm$ DE)	19 $\pm$ 2
Fracción de eyección de ventriculo izquierdo % (mediana y RIC)	52 (35-60)

DE: desviación estándar. RIC: rango intercuartil. IECA: Inhibidores de la enzima convertidora de angiotensina

en lugar del VPP. Si bien solo se utilizaron los mismos 24 predictores incluidos en alguno de los 5 modelos tradicionales para el entrenamiento y la validación los algoritmos RN, este nuevo enfoque metodológico fue suficiente para mejorar en forma significativa la predicción de los resultados.

Entre las 24 variables, las más influyentes para predecir mortalidad a corto y mediano plazo en los algoritmos RN fueron la creatinina y urea sérica, la hemoglobina, la frecuencia respiratoria, la concentración de sodio, la edad y la presión arterial sistólica al ingreso. Además de estas variables, la clase funcional III-IV de la NYHA y la demencia se asociaron con mayor mortalidad a 6 meses, y la frecuencia cardíaca y la disfunción renal crónica con la mortalidad al año.

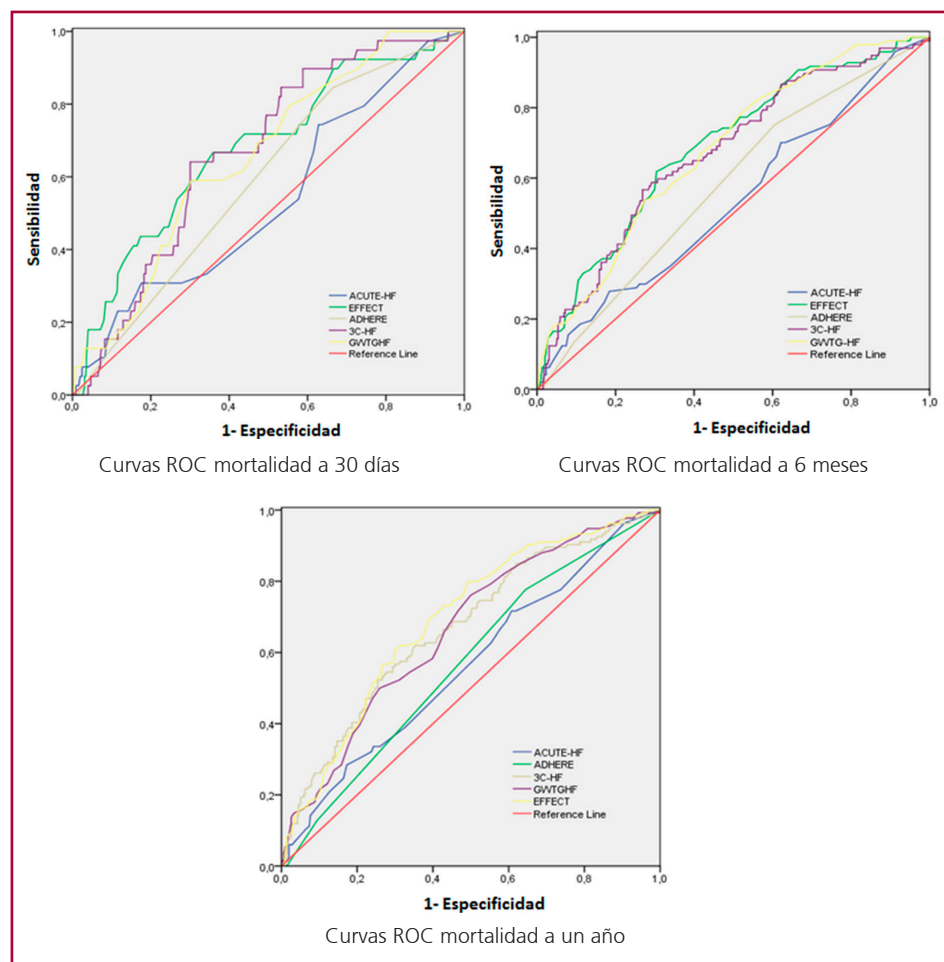
Los algoritmos RN han sido criticados en múltiples oportunidades por ser considerados como una “caja

negra” con capacidad limitada para identificar las posibles relaciones causales. En el presente estudio identificamos los factores más influyentes a través de los valores de importancia estandarizada. El análisis de importancia de las variables predictoras independientes calcula los pesos de cada una de ellas en el algoritmo RN a través de un análisis de sensibilidad sobre la base del testeo de la muestra. Además, como resultado del análisis de conglomerados jerárquico se determinaron las similitudes y relaciones entre el algoritmo RN y el resto de los *scores* de predicción.

Una mejora en la exactitud de la predicción sería útil para los pacientes con IC, y principalmente para aquellos con un peor pronóstico, que podrían beneficiarse con un tratamiento más dirigido, agresivo y con seguimiento más cercano. Además, estos *scores* mejorados podrían ayudar al diseño de los ensayos

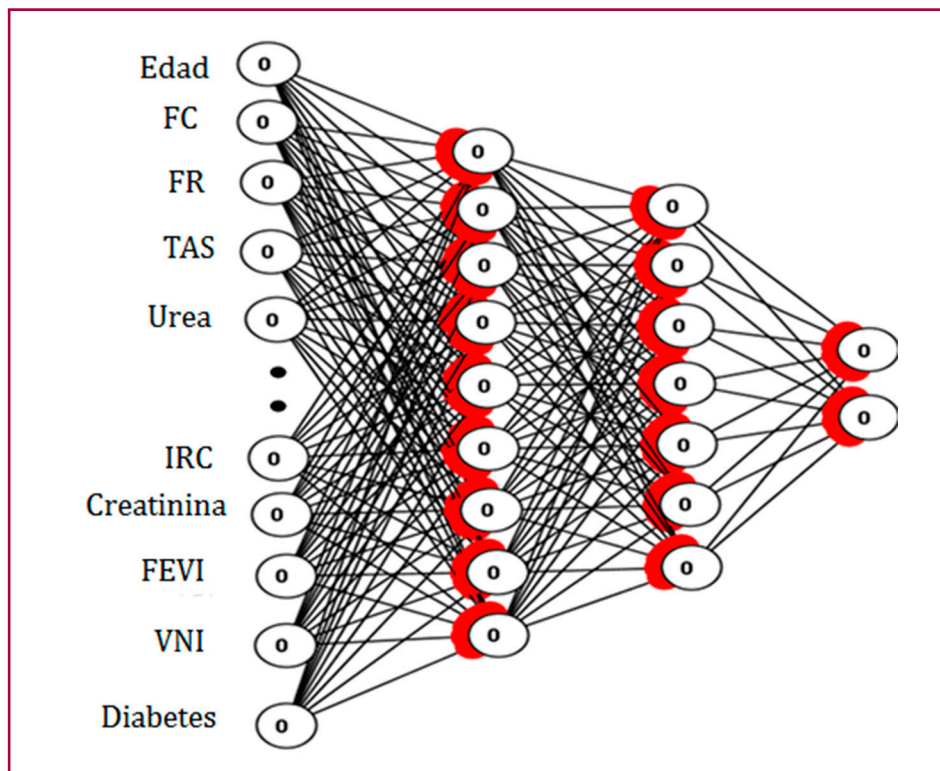
**Tabla 3.** Modelos de predicción de mortalidad a 30 días, 6 meses y al año en pacientes con insuficiencia cardíaca

	EFFECT	GWTG-HF	ADHERE	3C-HF	ACUTE-HF
<b>Mortalidad a 30 días:</b>					
Area bajo curva ROC	<b>0,68</b>	<b>0,66</b>	<b>0,59</b>	<b>0,67</b>	<b>0,54</b>
IC 95%	0,59-0,77	0,58-0,74	0,51-0,68	0,59-0,75	0,44-0,63
Hosmer-Lemeshow $\chi^2$	9,68	10,0	3,10	12,3	16,7
grados de libertad	8	8	2	8	6
Valor p	0,289	0,262	0,212	0,138	0,011
<b>Mortalidad a 6 meses:</b>					
Area bajo curva ROC	<b>0,69</b>	<b>0,68</b>	<b>0,58</b>	<b>0,67</b>	<b>0,53</b>
IC 95%	0,63-0,75	0,62-0,74	0,52-0,64	0,61-0,73	0,47-0,60
Hosmer-Lemeshow $\chi^2$	6,05	6,55	2,93	5,59	9,84
grados de libertad	8	8	2	8	6
Valor p	0,641	0,586	0,231	0,693	0,132
<b>Mortalidad al 1 año:</b>					
Area bajo curva ROC	<b>0,69</b>	<b>0,66</b>	<b>0,57</b>	<b>0,67</b>	<b>0,56</b>
IC 95%	0,64-0,74	0,61-0,72	0,51-0,63	0,62-0,72	0,51-0,62
Hosmer-Lemeshow $\chi^2$	4,65	11,8	3,03	2,98	6,68
grados de libertad	8	8	2	8	6
Valor p	0,794	0,163	0,220	0,936	0,352



**Fig. 1.** Curvas ROC de los diferentes modelos predictivos

**Fig. 2.** Arquitectura de la red neuronal del tipo perceptrón multicapa con dos capas ocultas.



FC, frecuencia cardíaca. FR, frecuencia respiratoria. TAS, tensión arterial sistólica. IRC, insuficiencia renal crónica. FEVI, fracción de eyección del ventrículo izquierdo. VNI, ventilación no invasiva

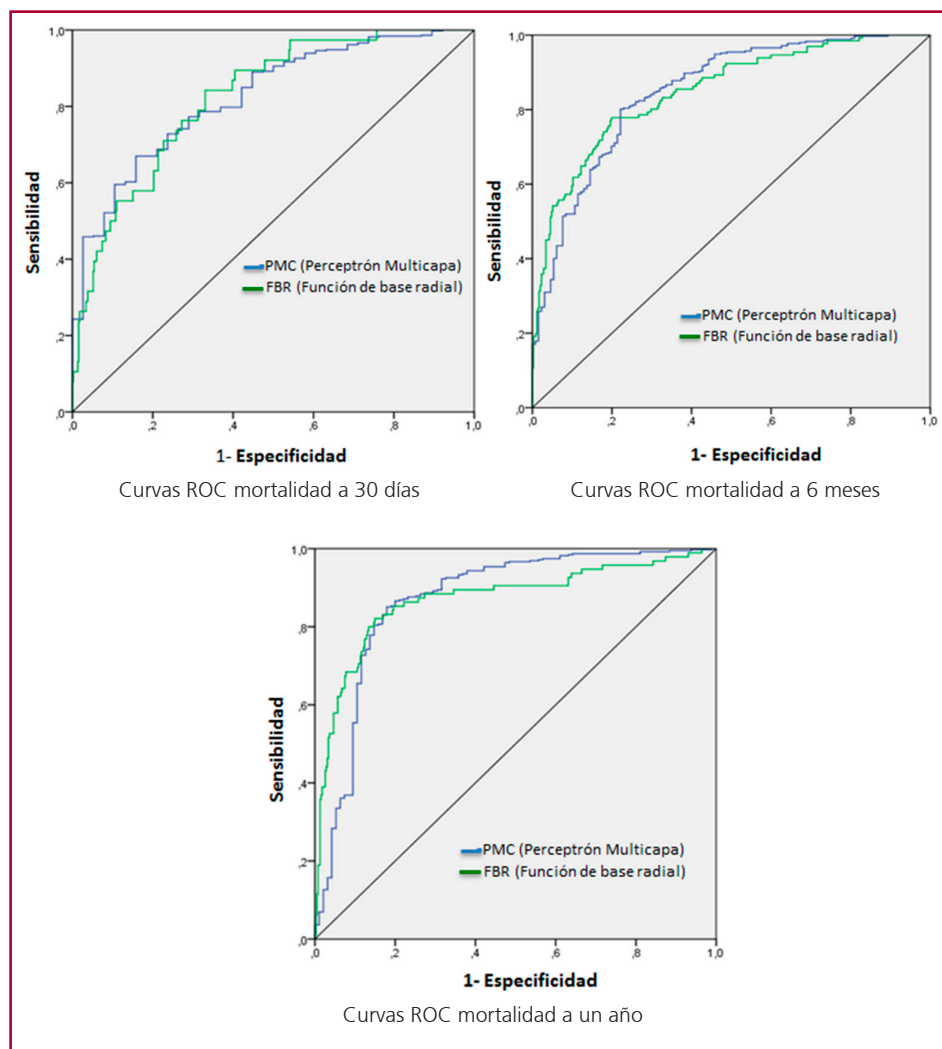
**Tabla 4.** Desempeño del modelo basado en redes neuronales para la predicción de la mortalidad por insuficiencia cardíaca a 30 días, 6 meses y un año

	Exactitud (IC95%)	Área ROC (IC95%)	VPN (IC95%)	VPP (IC95%)
<b>Mortalidad a 30 días:</b>				
PMC con dos capas ocultas arquitectura: 24-9-7-2	92,9% (90,5-95,3%)	0,82 (0,79-0,85)	93,2% (90,9-95,6%)	66,7% (28,9-100%)
<b>Mortalidad a 6 meses:</b>				
PMC con dos capas ocultas arquitectura: 24-9-7-2	87,7% (84,7-90,8%)	0,87 (0,84-0,90)	89,1% (85,9-92,2%)	78,6% (67,8-89,3%)
<b>Mortalidad a 1 año:</b>				
PMC con dos capas ocultas arquitectura: 24-9-7-2	84,4% (81,0-87,8%)	0,85 (0,81-0,88)	85,6% (81,9-89,2%)	78,9% (69,8-88,1%)

IC95%: intervalo de confianza del 95%. ROC: receiver operating characteristic. PMC: perceptrón multicapa. VPN: valor predictivo negativo. VPP: valor predictivo positivo.

clínicos facilitando la elección de aquella población con una mayor tasa potencial de eventos. Hasta ahora, la mayoría de los estudios demostraron que la predicción de la mortalidad y, en particular la de hospitalización en pacientes con IC sigue teniendo un éxito limitado, sin diferencias significativas en el valor de discriminación entre pacientes con IC crónica o aguda. Una revisión sistemática que hizo referencia al poder de discriminación de los *scores* de riesgo en IC, demostró que 69 de

117 modelos no presentaban validación externa. Estos modelos probablemente sobrestimaron la capacidad de predicción al utilizar una validación interna basada en el método *bootstrap*. (22) Es esperable que estos últimos modelos reporten áreas ROC más altas que los estudios con modelos validados en una población de pacientes diferente. De manera similar, los estudios de cohortes y prospectivos generalmente generan áreas ROC más altas que los modelos basados en análisis



**Fig. 3.** Curvas ROC para los modelos PMC (Perceptrón Multicapa) y FBR (Función de base radial)

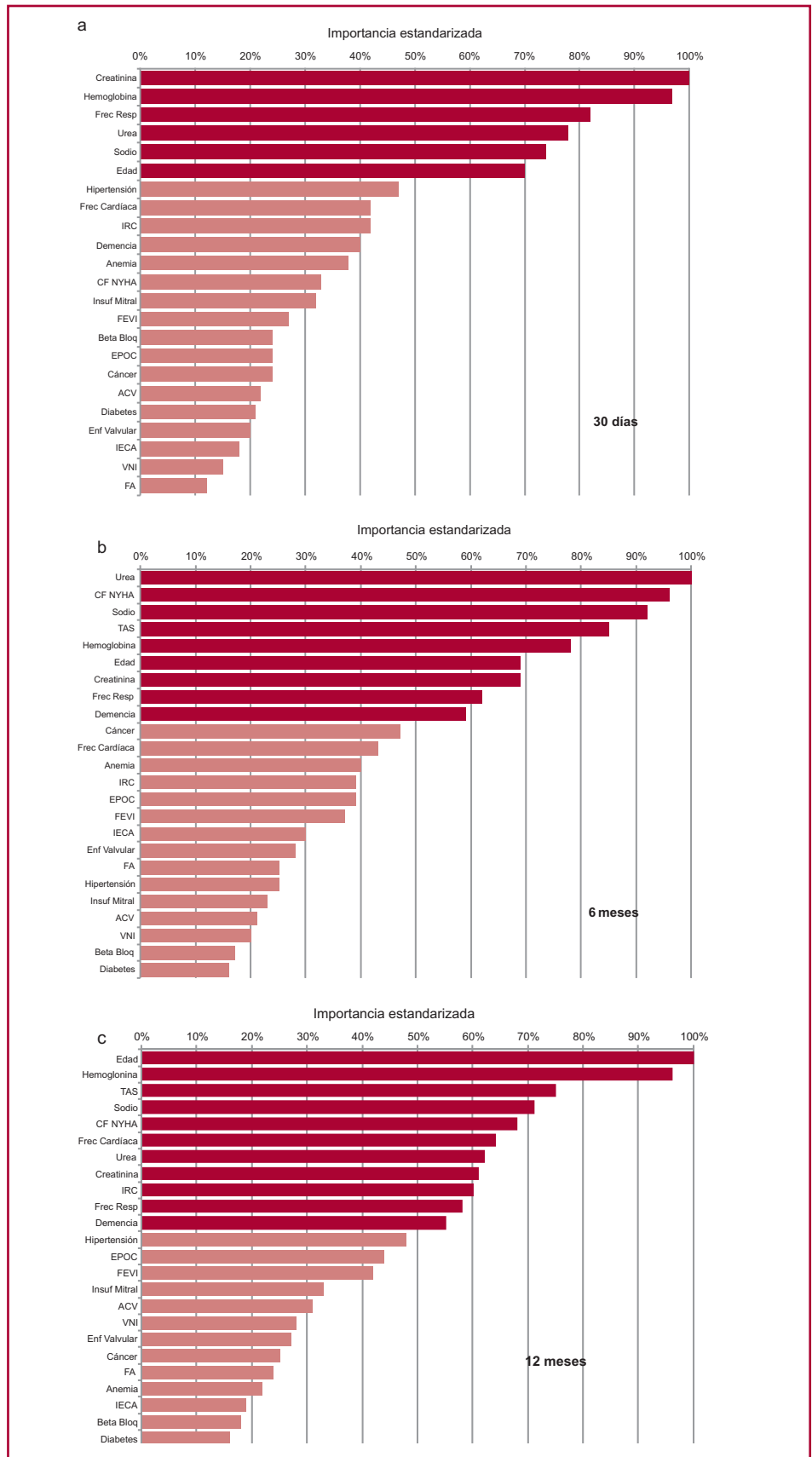
retrospectivos. Las áreas ROC de nuestro modelo de predicción fueron 0,82 y 0,85 para la mortalidad a 30 días y al año respectivamente, resultados similares a los obtenidos por Kwon et al. (16) quienes utilizaron un algoritmo basado en aprendizaje profundo (ROC: 0,88 y 0,79 para el mismo período de seguimiento, respectivamente). Además, el algoritmo RN de Kwon superó a otros métodos de aprendizaje automático como la regresión logística, *random forest*, *support vector machine* (máquinas de vector soporte) y las redes bayesianas. Hasta donde sabemos, sólo existen otros cuatro estudios que utilizaron algoritmos de RN o métodos tradicionales de aprendizaje automático en pacientes con IC, pero en estos casos se evaluaron las readmisiones a 30 días en lugar de la mortalidad por todas las causas. (23-26) Recientemente, una técnica de aprendizaje automático que utiliza *Support Vector Machines* (SVM) con núcleo *Gaussiano* se usó para validar un nuevo y simple modelo capaz de predecir la mortalidad a corto y largo plazo en pacientes con IC aguda. Sin embargo, este *score* de seis factores de-

nominado HANBAH (acrónimo de *hemoglobin, age, sodium, blood urea nitrogen, atrial fibrillation, and high-density lipoprotein*) solo alcanzó un área ROC de 0,75 en su mejor desempeño. (27) Teniendo en cuenta que uno de nuestros algoritmos RN del tipo FBR fue equivalente a una SVM con núcleo *Gaussiano*, nuestros valores de área ROC entre 0,75 y 0,78 fueron similares a los obtenidos con la puntuación HANBAH.

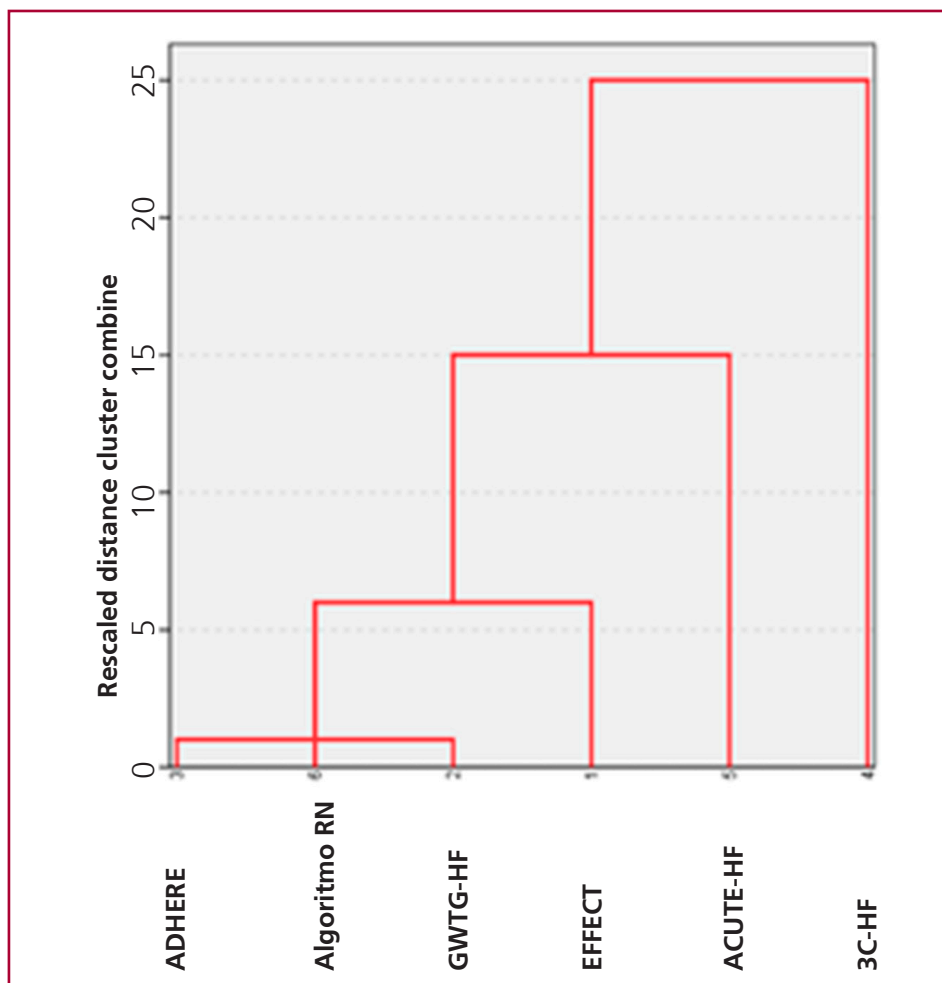
Como se mencionó previamente, en general en la literatura médica, los métodos de inteligencia artificial han demostrado que, aunque la mayoría de sus modelos se desempeñan con mejor nivel de discriminación y exactitud con respecto a los *scores* tradicionales, también suelen mostrar niveles altos de VPN y bajos de VPP. Una mayor exactitud en la predicción de la mortalidad hospitalaria y a mediano plazo luego de una admisión por IC aguda, es importante para identificar a las personas de alto riesgo que necesitarían recibir tratamiento y atención intensificada. Pero el riesgo individual podría calcularse mejor mediante *scores* suficientemente precisos que muestren valores



**Fig. 4.** Importancia estandarizada de las variables del perceptrón multicapa con 2 capas ocultas para predecir la mortalidad a (a) 30 días, (b) 6 meses y (c) al año en pacientes con insuficiencia cardíaca aguda



TAS: tensión arterial sistólica; CF: clase funcional; IRC: insuficiencia renal crónica EPOC: enfermedad pulmonar obstructiva crónica; FEVI: fracción de eyección ventricular izquierda; ACV: accidente cerebrovascular; VNI: ventilación no invasiva; IECA: inhibidores de la enzima convertidora de angiotensina; FA: fibrilación auricular



**Fig. 5.** Dendrograma del análisis de conglomerados jerárquicos en base a la vinculación promedio (average linkage) entre los grupos. Los modelos se ordenaron de acuerdo a sus similitudes como resultado del análisis jerárquico. Para el caso del algoritmo RN solo se tuvieron en cuenta los factores más influyentes con una "importancia estandarizada" mayor o igual a 50%.

de VPP más altos, en vez de valores altos de VPN. En aquellas condiciones médicas de baja incidencia de resultados adversos o eventos, la precisión general de los modelos de predicción de riesgo puede exagerarse con un VPN alto, aunque existan valores bajos de VPP. En consecuencia, los mejores modelos deberían basarse especialmente en el VPP y en la sensibilidad. En el presente estudio, el algoritmo de RN mostró altos niveles de precisión con un VPP aceptable entre 67% y 79%, aproximadamente.

A diferencia del enfoque estadístico convencional, el algoritmo RN no requiere de la preselección de variables significativas, ya que los factores menos importantes se ignoran automáticamente en el proceso de ajuste del modelo. Además, este modelo no limita el número de predictores de entrada, y puede utilizar toda la información disponible de una base de datos sin pérdida potencial. Sin embargo, en el presente estudio guiamos la selección de variables basados solo en los predictores incluidos en *scores* anteriores. Si bien se utilizaron las mismas variables de entrada, la mejora del rendimiento puede explicarse porque los algoritmos RN pueden detectar relaciones no lineales entre

variables independientes y dependientes más allá del alcance de la regresión logística.

Existen algunas limitaciones del presente estudio. Primero, aunque se pudieron identificar las variables más importantes en las que se basó la predicción del algoritmo RN, y las relaciones entre este nuevo modelo y los *scores* tradicionales por medio del dendrograma, el algoritmo RN sigue siendo de alguna forma una "caja negra", ya que no podemos interpretar el enfoque usado para clasificar el riesgo de los pacientes en forma individual. En segundo lugar, este modelo de predicción se desarrolló con un número limitado de variables extraídas de una base de datos de un solo centro. En tercer lugar, las variables candidatas para los modelos predictivos de insuficiencia cardíaca se seleccionaron originalmente a partir de variables clínicas extraídas de estudios publicados anteriormente, en los que se realizaron múltiples análisis de regresión logística para eliminar aquellos factores no asociados con la mortalidad a corto y mediano plazo. En consecuencia, se debe considerar que las variables de entrada utilizadas en el actual algoritmo RN han sido guiadas al menos por un análisis previo de regresión logística. Esta paradoja podría generar algún sesgo al

comparar los rendimientos entre el algoritmo RN y los *scores* tradicionales basados en regresiones logísticas.

Finalmente, se debe considerar que la falta de un set de prueba mas allá de la parcelación de entrenamiento y validación, puede conllevar a una sobreestimación de los resultados.

## CONCLUSIONES

El presente trabajo estudió la utilidad de las RN para predecir la mortalidad hospitalaria y a mediano plazo en pacientes con IC aguda. Mediante el uso de los predictores individuales incluidos en los 5 *scores* tradicionales (EFFECT, ADHERE, GWTG-HF, 3C-HF y ACUTE-HF) como variables de entrada de las RN, se demostró que el algoritmo basado inteligencia artificial fue más preciso y tuvo un mejor poder de discriminación que los *scores* mencionados. El análisis basado en RN constituye un modelo alternativo que mejora los resultados de los enfoques tradicionales.

## Declaración de Conflictos de interés

Los autores declaran no tener conflictos de interés en relación al presente artículo.

## BIBLIOGRAFÍA

1. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE Jr, Drazner MH, et al. 2013 ACCF/AHA guideline for the management of heart failure: executive summary: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol* 2013;62:1495-539. <https://doi.org/10.1016/j.jacc.2013.05.020>
2. Ferrero P, Iacovoni A, D'Elia E, Vaduganathan M, Gavazzi A, Senni M. Prognostic scores in heart failure - Critical appraisal and practical use. *Int J Cardiol* 2015;188:1-9. <https://doi.org/10.1016/j.ijcard.2015.03.154>
3. Fonarow GC, Adams KF Jr, Abraham WT, Yancy CW, Boscardin WJ, ADHERE Scientific Advisory Committee, Study Group, and Investigators. et al. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. *JAMA* 2005;293:572-80. <https://doi.org/10.1001/jama.293.5.572>
4. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV, et al. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *JAMA* 2003;290:2581-7. <https://doi.org/10.1001/jama.290.19.2581>
5. Peterson PN, Rumsfeld JS, Liang L, Hernandez AF, Peterson ED, Fonarow GC, et al. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. *Circ Cardiovasc Qual Outcomes* 2010;3:25-32. <https://doi.org/10.1161/CIRCOUTCOMES.109.854877>
6. O'Connor CM, Hasselblad V, Mehta RH, Tasissa G, Califf RM, Fiuzat M, et al. Triage after hospitalization with advanced heart failure: the ESCAPE (Evaluation Study of Congestive Heart Failure and Pulmonary Artery Catheterization Effectiveness) risk model and discharge score. *J Am Coll Cardiol* 2010;55:872-8. <https://doi.org/10.1016/j.jacc.2009.08.083>
7. Senni M, Parella P, De Maria R, Cottini C, Böhm M, Ponikowski P, et al. Predicting heart failure outcome from cardiac and comorbid conditions: The 3C-HF score. *Int J Cardiol* 2013;163:206-11. <https://doi.org/10.1016/j.ijcard.2011.10.071>
8. Cameli M, Pastore MC, De Carli G, Henein MY, Mandoli GE, Lisi E, et al. ACUTE HF score, a multiparametric prognostic tool for acute heart failure: A real-life study. *Int J Cardiol* 2019;296:103-8. <https://doi.org/10.1016/j.ijcard.2019.07.015>
9. Lee DS, Stitt A, Austin PC, Stukel TA, Stukel TA, Schull MJ, Chong A, et al. Prediction of heart failure mortality in emergent care: a cohort study. *Ann Intern Med* 2012;156:767-75. <https://doi.org/10.7326/0003-4819-156-11-201206050-00003>
10. Felker GM, Leimberger JD, Califf RM, Cuffe MS, Massie BM, Adams KF Jr, et al. Risk stratification after hospitalization for decompensated heart failure. *J Card Fail* 2004;10:460-6. <https://doi.org/10.1016/j.cardfail.2004.02.011>
11. Hsieh M, Auble TE, Yealy DM. Validation of the Acute Heart Failure Index. *Ann Emerg Med* 2008;51:37-44. <https://doi.org/10.1007/s12035-008-8015-2>
12. O'Connor CM, Mentz RJ, Cotter G, Metra M, Cleland JG, Davison BA, et al. The PROTECT in-hospital risk model: 7-day outcome in patients hospitalized with acute heart failure and renal dysfunction. *Eur J Heart Fail* 2012;14:605-12. <https://doi.org/10.1093/eurjhf/hfs029>
13. Abraham WT, Fonarow GC, Albert NM, Stough WG, Gheorghide M, Greenberg BH, et al. OPTIMIZE-HF Investigators and Coordinators. Predictors of in-hospital mortality in patients hospitalized for heart failure: insights from the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF). *J Am Coll Cardiol* 2008; 29:347-56. <https://doi.org/10.1016/j.jacc.2008.04.028>
14. Okazaki H, Shirakabe A, Hata N, Yamamoto M, Kobayashi N, Shinada T, et al. New scoring system (APACHE-HF) for predicting adverse outcomes in patients with acute heart failure: evaluation of the APACHE II and Modified APACHE II scoring systems. *J Cardiol* 2014;64:441-9. <https://doi.org/10.1016/j.jcc.2014.03.002>
15. Shiraishi Y, Kohsaka S, Abe T, Mizuno A, Goda A, Izumi Y, et al; West Tokyo Heart Failure Registry Investigators. Validation of the Get With The Guideline-Heart Failure risk score in Japanese patients and the potential improvement of its discrimination ability by the inclusion of B-type natriuretic peptide level. *Am Heart J* 2016;171:33-9. <https://doi.org/10.1016/j.ahj.2015.10.008>
16. Kwon JM, Kim KH, Jeon KH, Lee SE, Lee HY, Cho HJ, et al. Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. *PLoS One* 2019;14: e0219302. <https://doi.org/10.1371/journal.pone.0219302>
17. Yagy T, Kumada M, Nakagawa T. Novel risk stratification with time course assessment of in-hospital mortality in patients with acute heart failure. *PLoS One* 2017;12:e0187410. <https://doi.org/10.1371/journal.pone.0187410>
18. Win S, Hussain I, Hebl VB, Dunlay SM, Redfield MM. Inpatient Mortality Risk Scores and Postdischarge Events in Hospitalized Heart Failure Patients: A Community-Based Study. *Circ Heart Fail* 2017;10:e003926. <https://doi.org/10.1161/CIRCHEARTFAILURE.117.003926>
19. Lagu T, Pekow PS, Shieh MS, Stefan M, Pack QR, Kashef MA, et al. Validation and Comparison of Seven Mortality Prediction Models for Hospitalized Patients With Acute Decompensated Heart Failure. *Circ Heart Fail* 2016;9:e002912. <https://doi.org/10.1161/CIRCHEARTFAILURE.115.002912>
20. Martín-Sánchez FJ, Gil V, Llorens P, Herrero P, Jacob J, Fernández C, et al. Acute Heart Failure Working Group of the Spanish Society of Emergency Medicine Investigation Group. Barthel Index-Enhanced Feedback for Effective Cardiac Treatment (BI-EFFECT) Study: contribution of the Barthel Index to the Heart Failure Risk Scoring System model in elderly adults with acute heart failure in the emergency department. *J Am Geriatr Soc* 2012;60:493-8. <https://doi.org/10.1111/j.1532-5415.2011.03845.x>
21. Yap J, Lim FY, Chia SY, Allen JC Jr, Jaufeerally FR, Macdonald MR, et al. Prediction of Survival in Asian Patients Hospitalized With Heart Failure: Validation of the OPTIMIZE-HF Risk Score. *J Card Fail* 2019;25:571-5. <https://doi.org/10.1016/j.cardfail.2019.02.016>

22. Ouwkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. *JACC Heart Fail* 2014;2:429-36. <https://doi.org/10.1016/j.jchf.2014.04.006>
23. Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li SX, et al. Analysis of Machine Learning Techniques for Heart Failure Readmissions. *Circ Cardiovasc Qual Outcomes* 2016;9:629-40. <https://doi.org/10.1161/CIRCOUTCOMES.116.003039>
24. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure. *JAMA Cardiol* 2017;2:204. <https://doi.org/10.1001/jamacardio.2016.3956>
25. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC Med Inform Decis Mak* 2018;18:44. <https://doi.org/10.1186/s12911-018-0620-z>
26. Shameer K, Johnson KW, Yahi A, Miotto R, Li LI, Ricks D, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai Heart Failure Cohort. *Pac Symp Biocomput* 2017;22:276-87. [https://doi.org/10.1142/9789813207813\\_0027](https://doi.org/10.1142/9789813207813_0027)
27. Guo CY, Chan CH, Chou YC, Sung SH, Cheng HM. A Statistical Predictive Model Consistent Within a 5-Year Follow-up Period for Patients with Acute Heart Failure. *J Chin Med Assoc* 2020;83:1008-13. <https://doi.org/10.1097/JCMA.0000000000000403>