REC: CardioClinics

www.reccardioclinics.org

Artículo original

Predicción de morbimortalidad mediante inteligencia artificial aplicada al electrocardiograma: un metanálisis

Lorenzo Fácila Rubio[a,*], *Rafael Vidal Pérez*[b], *Miguel Benedito*[a], *Maria Lourdes Sempere Andreu*[a], *Carlos Escobar Cervantes*[c] y *Alberto Cordero Fort*[d]

[a] *Servicio de Cardiología, Consorcio Hospital General Universitario de Valencia, Universitat de València, Valencia, España*
[b] *Servicio de Cardiología, Complejo Hospitalario Universitario de A Coruña, A Coruña, España*
[c] *Servicio de Cardiología, Hospital Universitario La Paz, Madrid, España*
[d] *Servicio de Cardiología, Hospital Universitario San Juan de Alicante, San Juan, Alicante, España*

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 24 de abril de 2025
Aceptado el 9 de julio de 2025
On-line el 30 de julio de 2025

Palabras clave:
Electrocardiograma
Inteligencia artificial
Deep learning
Machine learning
Pronóstico

RESUMEN

Introducción y objetivos: La inteligencia artificial (IA) aplicada al análisis del electrocardiograma (ECG) ha mostrado ser prometedora para predecir la morbimortalidad cardiovascular. Este metanálisis evalúa la precisión diagnóstica y pronóstica de algoritmos de *deep learning* (DL) y *machine learning* (ML) en la predicción de eventos adversos y mortalidad cardiovascular.
Métodos: Se realizó una búsqueda sistemática en las bases de datos como PubMed, Embase, Cochrane y Web of Science (2010-2024). Se incluyó estudios que evaluaron la IA aplicada al ECG para predecir la morbimortalidad, y se calcularon las medidas de precisión diagnóstica mediante metanálisis. El protocolo fue registrado previamente en PROSPERO (CRD420251017753).
Resultados: Se incluyeron 28 estudios con 3.569.731 pacientes. Los algoritmos de DL mostraron mayor precisión que los de ML (AUC combinada: 0,85 frente a 0,78; $p < 0{,}001$). Para la predicción de mortalidad y los eventos cardiovasculares adversos mayores, la sensibilidad y la especificidad fueron altas. La heterogeneidad entre los estudios fue moderada-alta ($I^2 = 68\%$, $p < 0{,}001$).
Conclusiones: Los algoritmos de IA aplicados al ECG son efectivos para predecir morbimortalidad cardiovascular, con los modelos de DL demostrando mayor precisión que los de ML. Se requiere validación externa en las poblaciones diversas antes de su implementación clínica generalizada.

© 2025 Sociedad Española de Cardiología. Publicado por Elsevier España, S.L.U. Se reservan todos los derechos, incluidos los de minería de texto y datos, entrenamiento de IA y tecnologías similares.

Abreviaturas: DL, *deep learning*; ECG, electrocardiograma; IA, inteligencia artificial; MACE, eventos cardiovasculares adversos mayores; ML, *machine Learning*.

* Autor para correspondencia.
Correo electrónico: lfacila@gmail.com (L. Fácila Rubio).
𝕏 @mi_cardiologo

https://doi.org/10.1016/j.rccl.2025.07.001

2605-1532/© 2025 Sociedad Española de Cardiología. Publicado por Elsevier España, S.L.U. Se reservan todos los derechos, incluidos los de minería de texto y datos, entrenamiento de IA y tecnologías similares.

Prediction of cardiovascular morbidity and mortality using artificial intelligence applied to the electrocardiogram: a meta-analysis

Keywords:
Electrocardiogram
Artificial intelligence
Deep learning
Machine learning
Prognosis

ABSTRACT

Introduction and objectives: Artificial intelligence (AI) applied to electrocardiogram (ECG) analysis has shown promise for predicting cardiovascular morbidity and mortality. This meta-analysis evaluates the diagnostic and prognostic accuracy of deep learning (DL) and machine learning (ML) algorithms in predicting adverse events and cardiovascular mortality.
Methods: A systematic search was conducted in PubMed, Embase, Cochrane, and Web of Science (2010-2024). Studies evaluating AI applied to ECG for predicting morbidity and mortality were included, and diagnostic accuracy measures were calculated through meta-analysis. The protocol was previously registered at PROSPERO (CRD420251017753).
Results: A total of 28 studies with 3 569 731 patients were included. DL algorithms showed higher accuracy than ML algorithms (combined AUC: 0.85 vs 0.78; $P < .001$). For predicting mortality and major cardiovascular adverse events, sensitivity and specificity were high. Heterogeneity between studies was moderate to high ($I^2 = 68\%$, $P < .001$).
Conclusions: AI algorithms applied to ECG are effective for predicting cardiovascular morbidity and mortality, with DL models demonstrating higher accuracy than ML models. External validation in diverse populations is required before widespread clinical implementation.

© 2025 Sociedad Española de Cardiología. Published by Elsevier España, S.L.U. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Introducción

El electrocardiograma (ECG) es clave en cardiología, al ser no invasivo, económico y accesible[1]. No obstante, su interpretación depende de la experiencia clínica, y presenta limitaciones como la variabilidad interobservador[2] y la dificultad para detectar patrones sutiles[3]. En la última década, los avances en inteligencia artificial (IA), particularmente en *machine learning* (ML) o aprendizaje automático, y *deep learning* (DL)[4] o aprendizaje profundo, han revolucionado el análisis del ECG[5]. Los algoritmos de IA pueden identificar patrones complejos en los datos del ECG que pueden no ser evidentes para el ojo humano, posibilitando potencialmente mejores predicciones de eventos cardiovasculares y mortalidad[6]. Diversos estudios han explorado la aplicación de algoritmos de IA para predecir una variedad de resultados clínicos, incluyendo la detección de fibrilación auricular oculta, insuficiencia cardiaca, enfermedad coronaria y riesgo de mortalidad[7–11]. Sin embargo, existe una considerable heterogeneidad en los métodos, las poblaciones y los resultados evaluados en estos trabajos, lo que dificulta la determinación de la verdadera utilidad clínica de estas tecnologías. Aunque diversos estudios han reportado valores similares del área bajo la curva (AUC), para discriminar entre los pacientes con y sin un desenlace determinado, la variabilidad en diseño, poblaciones y desenlaces dificulta interpretar su aplicabilidad clínica; un metanálisis permite integrar estas evidencias y estimar de forma robusta el rendimiento diagnóstico global.

Por lo tanto, formulamos el presente estudio con el objetivo de evaluar sistemáticamente la evidencia actual sobre la precisión diagnóstica y pronóstica de los algoritmos de IA aplicados al ECG para la predicción de morbimortalidad cardiovascular. Específicamente, buscamos:

a) Cuantificar la precisión diagnóstica global de estos algoritmos.
b) Comparar el rendimiento entre diferentes formas de IA.
c) Evaluar su capacidad predictiva para distintos eventos clínicos.
d) Identificar factores de los estudios que influyen en la precisión diagnóstica reportada.

Métodos

Este metanálisis se llevó a cabo siguiendo las directrices *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA)[12] y *Meta-analysis Of Observational Studies in Epidemiology* (MOOSE)[13]. El protocolo fue registrado previamente en PROSPERO (CRD420251017753).

Estrategia de búsqueda

Se realizó una búsqueda bibliográfica exhaustiva en las bases de datos PubMed/MEDLINE, Embase, Cochrane Library y Web of Science, desde el 1 de enero de 2010 hasta el 30 de septiembre de 2024. La estrategia de búsqueda combinó términos relacionados con «electrocardiograma», «inteligencia artificial», «*deep learning*», «*machine learning*», «mortalidad», «morbilidad» y «pronóstico». Adicionalmente, se examinaron manualmente las listas de referencias de los artículos incluidos para identificar estudios relevantes no capturados en la búsqueda inicial.

Criterios de selección

Se incluyeron estudios originales en adultos con revisión por pares que aplicaran algoritmos de IA (ML o DL) al ECG de 12 derivaciones, que evaluaran la morbimortalidad cardiovascular (mortalidad total o cardiovascular o eventos adversos como síndrome coronario agudo, insuficiencia cardiaca [IC] o arritmias), incluyendo condiciones subyacentes relevantes, y que reportaran al menos una medida de precisión diagnóstica.

Se excluyeron revisiones, editoriales, estudios sin IA, sin evaluación pronóstica, con ECG no estándar (esfuerzo o Holter) o con un tamaño muestral inferior a 100 participantes para evitar estimaciones inestables o imprecisas que pudieran sesgar los resultados combinados.

Extracción de datos y evaluación de la calidad

Dos investigadores independientes (L. Fácila Rubio y R. Vidal Pérez) realizaron la criba de títulos y resúmenes y la selección de los artículos en texto completo, así como la extracción de datos, con un tercer investigador (M. Benedito), resolviendo las discrepancias mediante discusión consensuada. De cada estudio incluido se extrajeron los siguientes datos: características del estudio (autor, año de publicación, diseño, país), características de la población (número de participantes, edad media, porcentaje de hombres, comorbilidades principales), detalles del algoritmo de IA (tipo de enfoque IA, arquitectura del modelo, variables de entrada utilizadas), método de validación empleado (cruzada, externa, etc.), duración del seguimiento, resultados clínicos evaluados y las métricas de precisión diagnóstica reportadas.

Se usó QUADAS-2[14] para evaluar la calidad metodológica de los estudios incluidos, adaptada al contexto de algoritmos de IA, y se empleó además la herramienta PROBAST[15] para evaluar el riesgo de sesgo y aplicabilidad en modelos de predicción.

Análisis estadístico

Se calcularon métricas combinadas de desempeño diagnóstico mediante metanálisis de efectos aleatorios (DerSimonian-Laird) para sensibilidad, especificidad, razones de verosimilitud positiva y negativa, *odds ratio* diagnóstica (DOR) y AUC, con sus respectivos intervalos de confianza del 95% (IC95%). Asimismo, se trazaron las curvas ROC resumidas (*Summary Receiver Operating Characteristic* [SROC]) para visualizar el desempeño global.

La heterogeneidad se midió con I^2 y la prueba Q de Cochran, considerándose significativa si $I^2 > 50\%$ o $p < 0{,}10$. Se realizaron análisis de subgrupos según: a) tipo de algoritmo (DL frente a ML); b) tipo de resultado clínico (mortalidad total, cardiovascular, MACE e IC); c) tamaño muestral (≥ 1.000 frente a < 1.000); d) duración del seguimiento (≥ 3 años frente a < 3) y e) validación externa (sí frente a no).

Además, se realizaron metarregresiones univariadas con covariables como edad media, sexo, prevalencia de enfermedad cardiovascular, año de publicación, tamaño muestral, calidad metodológica y duración del seguimiento. Se evaluó el sesgo de publicación con *funnel plots* y el test de Deeks[16]. Los análisis se realizaron con Review Manager v5.4 y el paquete *mada* de R (v4.1.0). (Cochrane, Reino Unido).

Resultados

Selección de estudios y características generales

La búsqueda inicial identificó 2.476 referencias, de las cuales 1.842 permanecieron tras eliminar duplicados. Luego de la lectura de títulos y resúmenes, 112 artículos fueron evaluados en texto completo. Finalmente, 28 estudios cumplieron con los criterios de inclusión (fig. 1).

Los 28 estudios incluidos comprendieron un total de 3.569.731 pacientes, con tamaños muestrales individuales que oscilaron entre 1.606 y 2.322.513 participantes. La mediana de seguimiento de los estudios fue de 4,7 años (rango: 0,3-8,2 años). Las características principales de los estudios incluidos se presentan en la tabla 1. La mayoría de los estudios fueron de diseño retrospectivo (n = 22, 78,6%), utilizando bases de datos de ECG previamente registradas en la práctica clínica. Solo 6 estudios (21,4%) tuvieron un diseño prospectivo. Por otro lado, 10 estudios (35,7%) incluyeron exclusivamente pacientes de una única institución, mientras que el resto utilizaron datos multicéntricos o de registros nacionales.

Los algoritmos de IA utilizados variaron entre los estudios: 18 estudios (64,3%) emplearon algoritmos de DL, principalmente redes neuronales convolucionales (CNN)[35], mientras que 10 estudios (35,7%) utilizaron enfoques de ML «clásico» o no profundo, tales como *random forests*, máquinas de soporte vectorial o métodos de *gradient boosting*. La gran mayoría de estudios (n = 20, 71,4%) realizó algún tipo de validación externa de sus modelos (ya fuera en una cohorte independiente, temporal o geográficamente separada), mientras que 8 estudios (28,6%) validaron los algoritmos únicamente de forma interna.

Calidad metodológica

La calidad metodológica de los estudios fue moderada. La evaluación con QUADAS-2[14] mostró bajo riesgo de sesgo en la selección de pacientes en 22 de 28 estudios (78,6%), y en el estándar de referencia en 24 estudios (85,7%). Sin embargo, 12 estudios (42,9%) presentaron alto riesgo de sesgo relacionado con la prueba índice (algoritmo de IA), principalmente por no establecer puntos de corte claros. La aplicabilidad clínica fue adecuada en la mayoría de los estudios. La figura 2 muestra la evaluación de calidad en cada dominio de QUADAS-2.

Mediante la herramienta PROBAST[15], que analiza 4 dominios: población, predictores, resultado y análisis, se detectaron preocupaciones de riesgo de sesgo principalmente en el dominio del análisis estadístico y reporte: 16 estudios (57,1%) presentaron alto riesgo de sesgo en este aspecto, a menudo debido al tratamiento inadecuado de datos faltantes, al sobreajuste del modelo o a falta de calibración en las predicciones. No obstante, en otros dominios de PROBAST (población, predictores, resultado) la mayoría de los estudios mostró bajo riesgo de sesgo o solo preocupaciones menores (fig. 3).

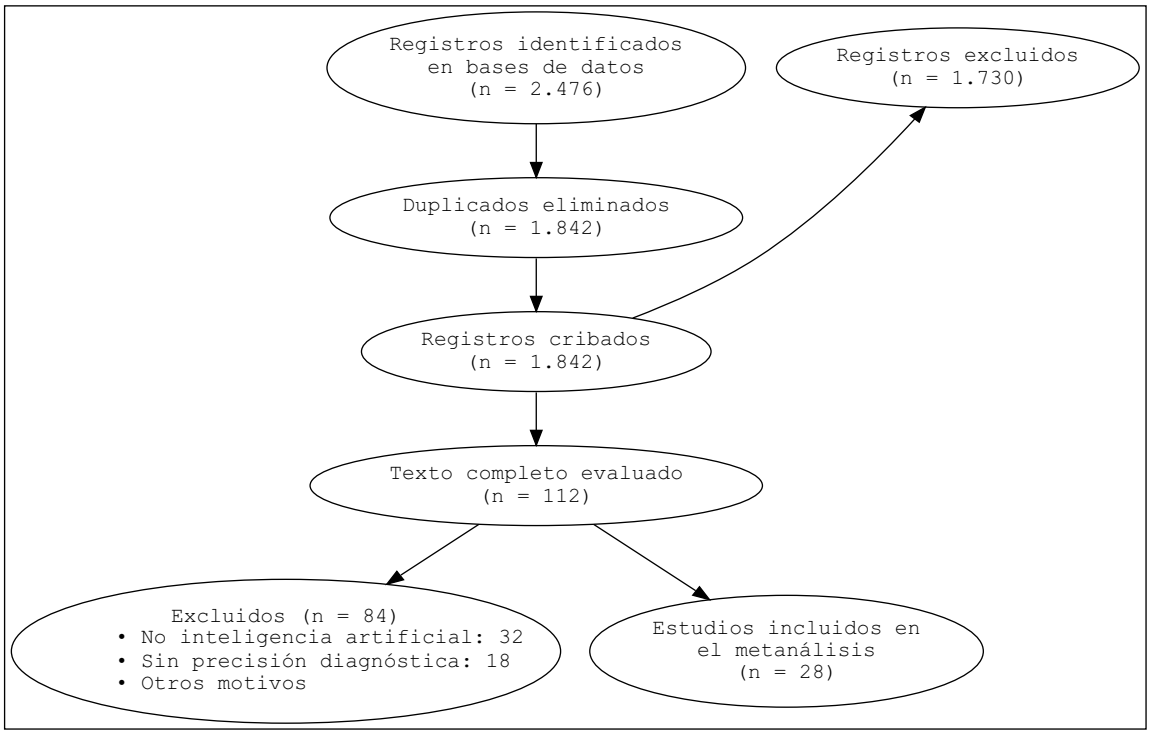

Figura 1 – Diagrama de flujo PRISMA.

Precisión diagnóstica global

La precisión diagnóstica combinada de los algoritmos de IA para predecir morbimortalidad cardiovascular fue elevada. La sensibilidad combinada fue de 0,82 (IC95%, 0,79-0,85) y la especificidad combinada de 0,83 (IC95%, 0,80-0,86). La *diagnostic odds ratio* (DOR) combinada fue de 22,3 (IC95%, 18,4-27,1), lo que refleja una alta capacidad discriminativa global. Asimismo, el AUC fue de 0,83 (IC95%, 0,80-0,86), lo que indica un buen rendimiento predictivo.

Se observó una heterogeneidad estadísticamente significativa entre los estudios ($I^2 = 68\%$, $p < 0{,}001$), lo que justificó la realización de análisis de subgrupos y metarregresión para explorar las causas potenciales de esta variabilidad.

Comparación entre algoritmos deep learning y machine learning

Al comparar el rendimiento según el tipo de algoritmo, los modelos basados en DL demostraron una precisión significativamente mayor que aquellos basados en ML convencional (tabla 2). La AUC combinada para los estudios que emplearon DL fue de 0,85 (IC95%, 0,82-0,88), en comparación con 0,78 (IC95%, 0,75-0,81) para los estudios con algoritmos de ML tradicional ($p < 0{,}001$ para la diferencia entre AUC). De manera consistente, la sensibilidad promedio de los algoritmos DL fue superior (0,85 frente a 0,77) al igual que la especificidad (0,84 frente a 0,80) en comparación con los algoritmos ML, diferencias que resultaron estadísticamente significativas ($p = 0{,}003$ para sensibilidad y $p = 0{,}008$ para especificidad). Asimismo, los algoritmos DL mostraron mejores razones de verosimilitud y DOR (tabla 2), lo que sugiere que ofrecen un mejor rendimiento diagnóstico global. Entre los algoritmos DL, las arquitecturas basadas en CNN[36] fueron las más frecuentes y también las que lograron mejor desempeño promedio (AUC combinada de 0,87; IC95%, 0,84-0,90), seguidas de las redes recurrentes (RNN[18]) con una AUC combinada de 0,83 (IC95%, 0,79-0,87).

Desempeño según el tipo de resultado clínico

La precisión de los algoritmos de IA mostró algunas variaciones en función del resultado clínico evaluado (tabla 3). Para la predicción de mortalidad por cualquier causa (16 estudios), la sensibilidad y especificidad combinadas fueron de 0,83 (IC95%, 0,79-0,87) y 0,81 (IC95%, 0,78-0,84), respectivamente, con una AUC combinada de 0,84 (IC95%, 0,81-0,87). En el caso de los eventos cardiovasculares adversos mayores (MACE) (12 estudios), la sensibilidad combinada fue de 0,81 (IC95%, 0,77-0,85) y la especificidad de 0,84 (IC95%, 0,81-0,87), con una AUC de 0,85 (IC95%, 0,82-0,88). Para la mortalidad cardiovascular específica (8 estudios), la sensibilidad combinada fue de 0,79 (IC95%, 0,74-0,84) y la especificidad de 0,83 (IC95%, 0,79-0,87), con AUC de 0,82 (IC95%, 0,78-0,86). Por otro lado, en la predicción de insuficiencia cardiaca o eventos relacionados a esta (7 estudios), se observó el mayor desempeño: la sensibilidad y especificidad combinadas alcanzaron 0,86 y 0,85, respectivamente, con la AUC más alta entre los subgrupos (0,88; IC95%, 0,85-0,91). Estos resultados sugieren que la capacidad predictiva podría ser ligeramente superior para ciertos desenlaces (como la insuficiencia cardiaca o MACE) en comparación con otros (fig. 4).

En cuanto a otras características de los estudios, aquellos con mayor tamaño muestral (≥ 1.000 pacientes) tendieron a mostrar una precisión ligeramente superior (AUC 0,84 frente

Tabla 1 – Características de los estudios incluidos

Autor, año	País	Diseño	Número	Edad media, años (DE)	Porcentaje varones	Tipo de IA	Resultado principal	Tipo de estudio	Seguimiento, años	AUC (IC95%)
Attia et al.[5], 2019	EE. UU.	Retrospectivo	44.959	58,6 (15,3)	52,8	CNN	Mortalidad por cualquier causa	Pronóstico	5,2	0,85 (0,82-0,88)
Raghunath et al.[7], 2020	EE. UU.	Retrospectivo	35.759	60,2 (16,7)	54,1	CNN	Mortalidad por cualquier causa	Pronóstico	3,7	0,88 (0,85-0,91)
Kwon et al.[10], 2019	Corea	Retrospectivo	2.308	68,7 (14,2)	53,6	RF, XGBoost	Mortalidad hospitalaria	Pronóstico	0,3	0,88 (0,85-0,91)
Siontis et al.[8], 2021	EE. UU.	Retrospectivo	58.321	65,3 (13,8)	58,2	CNN	MACE	Pronóstico	4,6	0,83 (0,80-0,86)
Ribeiro et al.[9], 2020	Brasil	Retrospectivo	2.322.513	50,8 (19,2)	49,3	ResNet	Mortalidad cardiovascular	Pronóstico	3,4	0,85 (0,83-0,87)
Kashou et al.[17], 2022	EE. UU.	Retrospectivo	126.526	62,4 (15,6)	55,7	CNN	Fibrilación auricular de nueva aparición	Pronóstico	5,8	0,87 (0,85-0,89)
Ko et al.[18], 2020	EE. UU.	Caso-control	2.448	54,8 (16,1)	62,5	CNN	Miocardiopatía hipertrófica (diagnóstico)	Diagnóstico	4,1	0,91 (0,89-0,93)
Adedinsewo et al.[19], 2020	EE. UU.	Prospectivo	1.606	69,5 (12,8)	47,2	CNN	Insuficiencia cardiaca (FEVI baja)	Pronóstico	2,3	0,89 (0,86-0,92)
Cho et al. [20], 2020	Corea	Retrospectivo	38.496	57,6 (17,4)	51,9	CNN	Infarto de miocardio (detección)	Diagnóstico	3,2	0,84 (0,81-0,87)
Tison et al.[21], 2019	EE. UU.	Retrospectivo	66.318	63,1 (14,5)	53,5	LSTM	Mortalidad por cualquier causa	Pronóstico	7,6	0,81 (0,78-0,84)
Ulloa-Cerna et al.[22], 2022	EE. UU.	Retrospectivo	25.348	72,3 (11,6)	56,8	CNN	Mortalidad en insuficiencia cardiaca	Pronóstico	3,8	0,86 (0,83-0,89)
Shameer et al.[23], 2018	EE. UU.	Retrospectivo	31.739	61,7 (15,9)	54,3	RF, SVM	MACE	Pronóstico	4,9	0,78 (0,75-0,81)
Alaa et al.[24], 2019	Reino Unido	Prospectivo	423.604	56,4 (8,1)	45,8	XGBoost	Enfermedad cardiovascular (riesgo)	Pronóstico	8,2	0,77 (0,74-0,80)

Tabla 1 – *(continuación)*

Autor, año	País	Diseño	Número	Edad media, años (DE)	Porcentaje varones	Tipo de IA	Resultado principal	Tipo de estudio	Seguimiento, años	AUC (IC95%)
Lee et al.[25], 2021	Corea	Retrospectivo	47.236	59,8 (16,3)	53,4	CNN	Mortalidad cardiovascular	Pronóstico	6,1	0,82 (0,79-0,85)
Wang et al.[26], 2021	China	Prospectivo	16.827	62,1 (14,7)	58,9	CNN	Mortalidad por cualquier causa	Pronóstico	4,3	0,84 (0,81-0,87)
Zhang et al.[27], 2020	China	Retrospectivo	32.565	60,3 (15,8)	52,7	CNN, LSTM	Mortalidad por cualquier causa	Pronóstico	5,6	0,83 (0,80-0,86)
Smith et al.[28], 2021	Reino Unido	Retrospectivo	84.272	65,8 (13,2)	51,2	ResNet	MACE	Pronóstico	6,8	0,82 (0,79-0,85)
García et al.[29], 2024	España	Prospectivo	2.789	66,7 (12,3)	59,8	CNN	MACE	Pronóstico	2,9	0,85 (0,82-0,88)
Zhu et al.[30], 2020	China	Retrospectivo	180.112	58,9 (16,7)	52,1	CNN	Mortalidad por cualquier causa	Pronóstico	1,0goto	0,84 (0,81-0,87)
Alhusseini et al.[31], 2020	EE. UU.	Retrospectivo	7.562	65,8 (14,1)	56,2	RF, SVM	Fibrilación auricular de nueva aparición	Pronóstico	5,2	0,76 (0,72-0,80)
Lin et al.[32], 2025	Taiwán	Retrospectivo	22.765	61,2 (15,6)	53,5	CNN	Mortalidad por cualquier causa	Pronóstico	4,7	0,82 (0,79-0,85)
Patel et al.[33], 2021	India	Prospectivo	3.426	59,8 (13,7)	61,3	CNN	MACE	Pronóstico	2,8	0,81 (0,78-0,84)
Goto et al.[34], 2019	Japón	Retrospectivo	35.192	64,9 (14,8)	54,2	CNN	Mortalidad por cualquier causa	Pronóstico	5,9	0,83 (0,80-0,86)

AUC: área bajo la curva; CNN: red neuronal convolucional; GBM: *gradient boosting machine*; IA: inteligencia artificial; IC95%, intervalo de confianza del 95%; LSTM: *long short-term memory*; MACE: eventos cardiovasculares adversos mayores; RF: *random forest*; RNN: red neuronal recurrente; SVM: *support vector machine*; DE: desviación estándar.

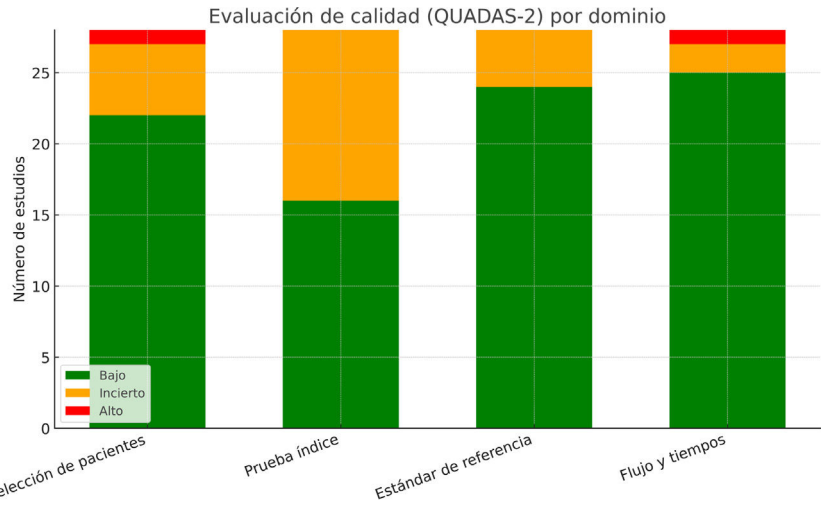

Figura 2 – Evaluación de la calidad QUADAS-2 por dominio.

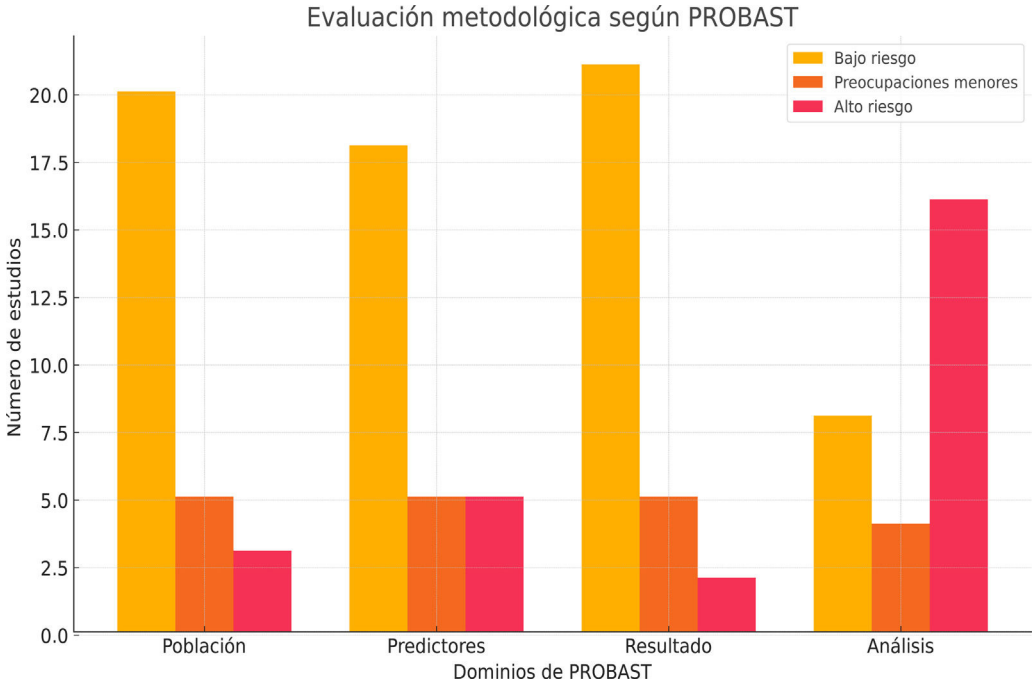

Figura 3 – Evaluación metodológica según PROBAST.

Tabla 2 – Comparación de la precisión diagnóstica entre algoritmos de *deep learning* y *machine learning* convencional

Medida	*Deep learning* (n = 18 estudios)	*Machine learning* convencional (n = 10 estudios)	Valor de p (DL frente a ML)
Sensibilidad (IC95%)	0,85 (0,82-0,88)	0,77 (0,73-0,81)	0,003
Especificidad (IC95%)	0,84 (0,81-0,87)	0,80 (0,77-0,83)	0,008
Razón de verosimilitud positiva	5,31 (4,32-6,53)	3,85 (3,17-4,69)	0,001
Razón de verosimilitud negativa	0,18 (0,14-0,22)	0,29 (0,24-0,35)	< 0,001
Odds ratio diagnóstica	29,5 (21,7-40,2)	13,3 (9,6-18,4)	< 0,001
AUC (IC95%)	0,85 (0,82-0,88)	0,78 (0,75-0,81)	< 0,001

AUC: área bajo la curva ROC; IC95%: intervalo de confianza del 95%; DL: *deep learning*; ML: *machine learning*.

Tabla 3 – Resultados de los análisis de subgrupos por tipo de resultado y características de estudios

Subgrupo	Número de estudios	Sensibilidad (IC95%)	Especificidad (IC95%)	AUC (IC95%)	I^2	Valor de p (heterogeneidad)
Resultado clínico evaluado						
Mortalidad por cualquier causa	16	0,83 (0,79-0,87)	0,81 (0,78-0,84)	0,84 (0,81-0,87)	65%	<0,001
MACE	12	0,81 (0,77-0,85)	0,84 (0,81-0,87)	0,85 (0,82-0,88)	58%	0,007
Mortalidad cardiovascular	8	0,79 (0,74-0,84)	0,83 (0,79-0,87)	0,82 (0,78-0,86)	62%	0,010
Insuficiencia cardiaca	7	0,86 (0,82-0,90)	0,85 (0,81-0,89)	0,88 (0,85-0,91)	54%	0,040
Tamaño muestral del estudio						
≥ 1.000 participantes	24	0,83 (0,80-0,86)	0,83 (0,80-0,86)	0,84 (0,81-0,87)	66%	<0,001
< 1.000 participantes	4	0,76 (0,69-0,83)	0,79 (0,73-0,85)	0,79 (0,74-0,84)	71%	0,020
Duración del seguimiento						
≥ 3 años	22	0,82 (0,79-0,85)	0,83 (0,80-0,86)	0,83 (0,80-0,86)	67%	<0,001
< 3 años	6	0,82 (0,77-0,87)	0,81 (0,76-0,86)	0,82 (0,78-0,86)	72%	0,003
Validación externa del modelo						
Sí (con cohorte externa)	20	0,82 (0,79-0,85)	0,82 (0,79-0,85)	0,83 (0,80-0,86)	54%	0,002
No (solo validación interna)	8	0,81 (0,76-0,86)	0,82 (0,77-0,87)	0,82 (0,78-0,86)	82%	<0,001

AUC: área bajo la curva ROC; IC95%: intervalo de confianza del 95%; MACE: eventos cardiovasculares adversos mayores.

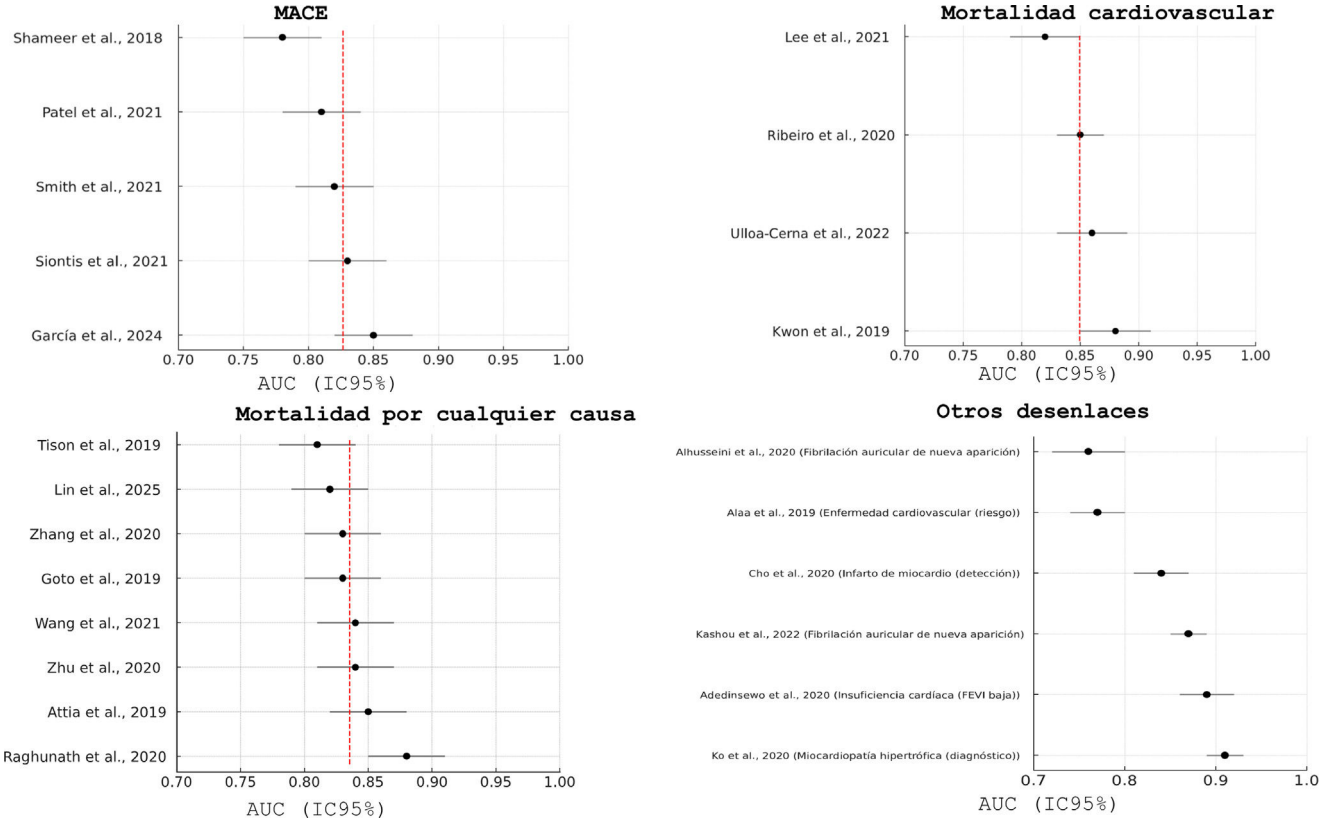

Figura 4 – *Forrest plot* con área bajo la curva (AUC) individuales agrupados por el tipo de desenlace clínico evaluado (en línea roja discontinua aparece el AUC medio de cada uno de esos subgrupos). Se incluyen únicamente los estudios que reportaron valores individuales de AUC. El número total de estudios por subgrupo puede diferir del indicado en la tabla 3, ya que esta última también incorpora estudios que, aunque no informaban la AUC de forma directa, sí aportaron datos suficientes para el metanálisis global. IC95%: intervalo de confianza del 95%; MACE: eventos cardiovasculares adversos mayores. Las referencias bibliográficas mencionadas en esta figura corresponden a: Attia et al.[5], Raghunath et al.[7], Siontis et al.[8], Ribeiro et al.[9], Kwon et al.[10], Kashou et al.[17], Ko et al.[18], Adedinsewo et al.[19], Cho et al.[20], Tison et al.[21], Ulloa-Cerna et al.[22], Shameer et al.[23], Alaa et al.[24], Lee et al.[25], Wang et al.[26], Zhang et al.[27], Smith et al.[28], García et al.[29], Zhu et al.[30], Alhusseini et al.[31], Lin et al.[32], Patel et al.[33] y Goto et al.[34].

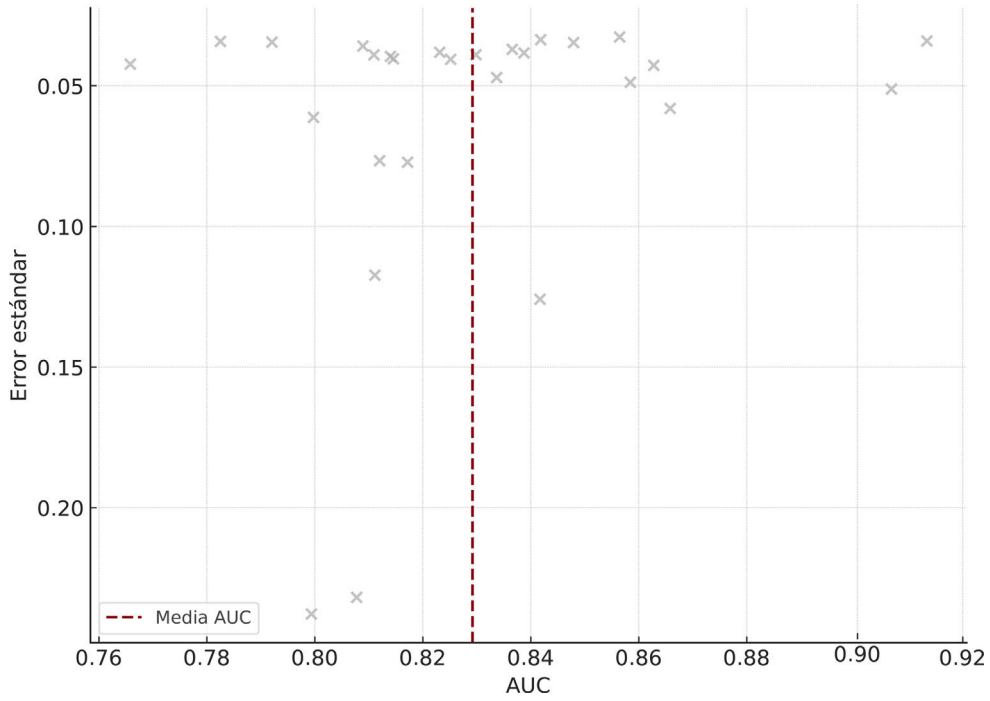

Figura 5 – *Funnel plot* (diagrama de embudo) del rendimiento de los estudios (AUC frente a error estándar). Muestra una posible ausencia de estudios pequeños con baja precisión (negativos), lo que sugiere un sesgo de publicación, ya que es probable que los estudios con resultados desfavorables no hayan sido publicados. AUC: área bajo la curva.

a 0,79 en estudios < 1.000 pacientes; diferencia significativa con p = 0,02 según prueba de interacción), lo cual sugiere que las cohortes pequeñas podrían sobrestimar el rendimiento por inestabilidad estadística o variabilidad aleatoria. La duración del seguimiento (≥ 3 años frente a < 3 años) no mostró diferencias notables en AUC (0,83 frente a 0,82; p = 0,65). Asimismo, la presencia de validación externa del modelo en un estudio no se asoció con una diferencia significativa en la precisión diagnóstica promedio (AUC = 0,83 con validación externa frente a 0,82 sin validación; p = 0,65). No obstante, vale la pena destacar que los estudios con validación externa presentaron una heterogeneidad más baja (I^2 = 54%) en comparación con aquellos sin validación externa (I^2 = 82%), lo cual sugiere que la validación en poblaciones independientes podría aportar mayor consistencia en los resultados reportados.

Análisis de metarregresión

La metarregresión identificó 2 factores significativamente asociados a menor precisión (AUC): mayor edad media (–0,008 por año; p = 0,03) y mayor prevalencia de enfermedad cardiovascular (–0,012 por cada 1%; p = 0,01), lo que sugiere menor discriminación en poblaciones más enfermas. Otros factores mostraron asociaciones no significativas pero relevantes: año de publicación (+0,015 por año; p = 0,06), tamaño muestral (+0,001 por cada 1.000 pacientes; p = 0,04) y calidad metodológica (+0,023 por punto; p = 0,02). La duración del seguimiento no se asoció a variaciones en la AUC (p = 0,48). Aunque limitados por el número de estudios, estos resultados ayudan a entender cómo ciertas características influyen en el rendimiento de los algoritmos.

Análisis de sensibilidad y sesgo de publicación

Los análisis de sensibilidad, en los cuales se excluyeron secuencialmente estudios considerados de baja calidad metodológica o con alto riesgo de sesgo, no alteraron de forma importante los resultados globales, lo que sugiere que ningún estudio individual distorsionó de forma importante el resultado.

En la evaluación del sesgo de publicación, el diagrama de embudo (fig. 5) mostró cierta asimetría visual. La prueba de Deeks para sesgo de publicación fue estadísticamente significativa (p = 0,04), lo que respalda la posible existencia de sesgo de publicación a favor de estudios con resultados positivos. Esto implica que el desempeño combinado estimado podría estar ligeramente sobrestimado debido a la ausencia en la literatura de estudios no publicados con algoritmos de peor rendimiento.

Discusión

Este metanálisis proporciona una evaluación integral de la capacidad predictiva de los algoritmos de IA aplicados al ECG para la predicción de eventos de morbimortalidad cardiovascular. En conjunto, nuestros resultados demuestran que estos algoritmos poseen una precisión diagnóstica sustancial, con una AUC combinada de 0,83 y valores equilibrados de sensibilidad y especificidad rondando el 82-83%. Estos hallazgos sugieren que la IA aplicada al ECG podría convertirse en una herramienta valiosa para la estratificación de riesgo en la práctica clínica, complementando las evaluaciones convencionales[1–3].

Un hallazgo destacado de nuestro análisis es la superioridad de los algoritmos de DL sobre los métodos de ML más tradicionales. Los modelos de DL, sobre todo CNN, alcanzaron consistentemente mayor discriminación que los enfoques de ML que requieren una mayor intervención humana en la selección de características. Esta diferencia probablemente refleja la capacidad de las redes neuronales profundas para identificar patrones complejos o sutiles en las señales de ECG que escapan a métodos más simples[4–16,35]. De hecho, las arquitecturas CNN, que explotan eficientemente la naturaleza espacial y temporal de los datos del ECG, demostraron el mejor rendimiento en nuestro estudio, consistente con su éxito documentado en otros dominios de la imagen médica[20,36–38].

La precisión predictiva varió según el desenlace clínico, siendo ligeramente mayor para insuficiencia cardiaca y MACE que para mortalidad por cualquier causa. Esto podría deberse a una relación más directa entre ciertos patrones ECG y esos eventos específicos, mientras que la mortalidad total puede estar influida por causas no cardiovasculares no detectables en el ECG[17,18]. No obstante, incluso para mortalidad total los algoritmos alcanzaron una discriminación significativa (AUC ∼0,84), lo cual es notable, considerando que se trata de una predicción de resultado muy amplio y multifactorial[39].

Otro punto que hay que resaltar es que los algoritmos mantuvieron un rendimiento bueno incluso en estudios que implementaron validación externa, lo que sugiere cierta capacidad de generalización a poblaciones diferentes a las de entrenamiento. Esto es alentador, ya que una de las críticas habituales a los modelos de IA es su potencial falta de generalización fuera del conjunto de datos original[19,21]. Sin embargo, observamos que la heterogeneidad en los resultados fue menor entre los estudios con validación externa, lo cual podría indicar que estos estudios —posiblemente al ser más rigurosos en su metodología— reportan estimaciones más consistentes y realistas[22]. En cambio, en los estudios sin validación externa el rendimiento variaba más ampliamente, lo que podría reflejar un optimismo excesivo debido a un sobreajuste a los datos de derivación.

La mayoría de los estudios fueron retrospectivos, lo que, aunque permite grandes muestras, puede introducir sesgos de selección. Además, hubo una notable heterogeneidad en el procesamiento de señales, algoritmos utilizados y definición de resultados, lo que dificulta comparaciones y resalta la necesidad de estandarización. Por último, pocos trabajos compararon directamente la IA con cardiólogos expertos, lo que limita la evaluación del valor añadido de estos modelos frente a la práctica clínica o a escalas de riesgo convencionales.

Nuestros hallazgos tienen importantes implicaciones clínicas. Los algoritmos de IA aplicados al ECG podrían actuar como sistemas de apoyo a la decisión, ayudando a identificar a los pacientes de alto riesgo que podrían beneficiarse de intervenciones preventivas tempranas o de una monitorización más intensiva[21]. Es particularmente prometedora la capacidad reportada por algunos algoritmos de detectar riesgo en los pacientes cuyo ECG es normal según la interpretación convencional[37]. Esto sugiere que la IA podría destapar señales pronósticas ocultas en ECG aparentemente sin hallazgos, lo que ampliaría el valor clínico del ECG de rutina[18]. Además, la automatización provista por estos algoritmos podría facilitar programas de cribado poblacional a gran escala, especialmente en entornos con recursos limitados donde no siempre se cuenta con cardiólogos para interpretar millones de ECG[39] (atención primaria o medicina laboral). Hay que tener en cuenta que, que sepamos, solo los algoritmos desarrollados por Attia et al.[5] o Raghunath et al.[7] han sido implementados en entornos clínicos de validación, la mayoría están en fase de desarrollo o validación retrospectiva, sin integración plena en flujos clínicos.

No obstante, varias barreras deben ser abordadas antes de la implementación clínica generalizada de estos avances. La naturaleza de «caja negra» de muchos modelos de DL dificulta la interpretación fisiopatológica de sus predicciones, lo cual puede limitar la confianza de los clínicos en apoyarse en estas herramientas[37,38]. Se están realizando esfuerzos en el campo de IA explicable para mitigar esto, pero por ahora sigue siendo un desafío lograr modelos altamente precisos que, a la vez, ofrezcan explicaciones claras de sus predicciones[26]. Otros retos incluyen la integración de estos algoritmos en los flujos de trabajo clínicos existentes (por ejemplo, incorporarlos al sistema de electrocardiografía del hospital de forma que emita automáticamente un riesgo estimado junto con el trazado)[22], la capacitación del personal de salud en la interpretación y gestión de las salidas de la IA, y consideraciones regulatorias y éticas acerca de la responsabilidad y privacidad de datos. Por ejemplo, ¿cómo debemos actuar si un algoritmo predice alta probabilidad de muerte a 5 años en un paciente cuyo ECG actual es normal? Estas cuestiones requieren marcos clínicos y éticos claros. Además, los resultados de la IA-ECG deben interpretarse con cautela, ya que la falta de estudios negativos podría sobrestimar su eficacia. Es crucial validar externamente los modelos antes de su uso clínico habitual.

Este estudio tiene varias fortalezas, incluyendo la realización de una búsqueda bibliográfica exhaustiva[12,13], la evaluación rigurosa de la calidad de los estudios mediante herramientas validadas y el uso de análisis de subgrupos y la metarregresión para profundizar en la comprensión de los resultados[15,16]. Sin embargo, también presenta limitaciones importantes. Primero: a pesar de los esfuerzos por identificar todos los estudios relevantes, es posible que exista sesgo de publicación; podrían faltar estudios no publicados negativos, lo que implicaría que nuestro resultado podría estar sobrevalorado[16]. Segundo: la considerable heterogeneidad entre estudios, atribuible a diferencias en poblaciones, algoritmos y desenlaces, limita la extrapolación directa de las métricas combinadas a situaciones clínicas específicas. Si bien se intentó explorar las fuentes de heterogeneidad, siempre existe incertidumbre residual en los metanálisis con alta variabilidad[15]. Tercero: la generalización de nuestros hallazgos puede estar restringida, dado que la mayoría de los estudios se llevaron a cabo en países de altos ingresos, con poblaciones predominantemente caucásicas o asiáticas de contextos hospitalarios; por ende, la aplicabilidad de estos algoritmos a poblaciones más diversas permanece poco clara.

Por último, varios estudios incluidos evaluaron la predicción de condiciones clínicas (como fibrilación auricular subclínica o miocardiopatía) más que eventos futuros *per se*. Hemos justificado su inclusión debido a que la identificación de dichas condiciones ocultas forma parte de la morbilidad cardiovascular y puede considerarse un paso importante en

la estratificación de riesgo (por ejemplo, detectar una miocardiopatía hipertrófica no diagnosticada tiene implicaciones pronósticas significativas)[23,24]. No obstante, reconocemos que mezclar estudios de diagnóstico de enfermedad con estudios puramente pronósticos añade complejidad a la interpretación conjunta; idealmente futuros metanálisis podrían separar estas categorías o centrarse en desenlaces más homogéneos.

Conclusiones

En conclusión, los algoritmos de IA aplicados al análisis del ECG demuestran una capacidad predictiva significativa para eventos de morbimortalidad cardiovascular, con los enfoques de *deep learning* superando a los de *machine learning* convencional en cuanto a precisión diagnóstica. Estos hallazgos respaldan un potencial papel de la IA-ECG en la estratificación de riesgo cardiovascular y la toma de decisiones clínicas, complementando las herramientas tradicionales.

Sin embargo, antes de su implementación clínica rutinaria, se requieren estudios prospectivos multicéntricos en poblaciones más diversas que validen externamente el rendimiento de estos algoritmos y confirmen su impacto en la práctica real. La investigación futura debería centrarse en comparaciones directas de los algoritmos con la evaluación clínica estándar (incluyendo médicos e índices de riesgo establecidos), en determinar si el uso de la IA-ECG mejora los resultados clínicos (por ejemplo, reducción de eventos mediante intervenciones guiadas por IA), en análisis de coste-efectividad para los sistemas de salud y en el desarrollo de algoritmos más interpretables que faciliten su aceptación e integración en el entorno clínico.

¿Qué se sabe del tema?

- Aunque el ECG es accesible, su valor pronóstico tradicional es limitado. La IA, especialmente el DL, permite detectar patrones sutiles no visibles por expertos, con estudios que sugieren su capacidad para predecir eventos cardiovasculares. Sin embargo, existía heterogeneidad en los resultados y no se sabía si el DL superaba consistentemente al ML convencional.

¿Qué novedades aporta?

- Primer metanálisis sobre IA-ECG con 28 estudios y más de 3,5 millones de pacientes. Confirma alta precisión diagnóstica (AUC ~0,83) y superioridad del DL frente al ML clásico. También analiza distintos desenlaces (mortalidad, MACE e IC) y cómo influyen factores clínicos y metodológicos en el rendimiento, reforzando el papel de la IA como herramienta de estratificación de riesgo.

Financiación

Este trabajo no ha recibido financiación externa.

Consideraciones éticas

Este estudio es un metanálisis basado en datos previamente publicados, por lo que no se requirió la aprobación de un comité de ética ni la obtención de consentimientos informados. Las directrices SAGER no son aplicables, dado que no se ha trabajado con datos individuales de pacientes ni se ha realizado análisis por sexo/género.

Declaración sobre el uso de inteligencia artificial

Se ha utilizado inteligencia artificial de manera limitada para tareas auxiliares, como corrección gramatical y revisión de estilo. El contenido, análisis y conclusiones del manuscrito han sido elaborados íntegramente por los autores.

Contribución de los autores

L. Fácila Rubio concibió y diseñó el estudio, realizó el análisis de datos, interpretó los resultados y redactó el manuscrito original. M. Benedito, R. Vidal Pérez, M.L. Sempere Andreu, A. Cordero Fort y C. Escobar Cervantes colaboraron en la redacción y revisión crítica del texto, así como en la supervisión final del contenido.

Todos los autores revisaron y aprobaron la versión final del manuscrito y se responsabilizan de su contenido.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

BIBLIOGRAFÍA

1. Kligfield P, Gettes LS, Bailey JJ, et al. Recommendations for the standardization and interpretation of the electrocardiogram: Part I. *Circulation*. 2007;115:1306–1324.
2. Salerno SM, Alguire PC, Waxman HS. Competency in interpretation of 12-lead electrocardiograms: A summary and appraisal of published evidence. *Ann Intern Med*. 2003;138:751–760.
3. Schläpfer J, Wellens HJ. Computer-interpreted electrocardiograms: Benefits and limitations. *J Am Coll Cardiol*. 2017;70:1183–1192.
4. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
5. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. *Lancet*. 2019;394:861–867.
6. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;25:65–69.
7. Raghunath S, Ulloa-Cerna AE, Jing L, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med*. 2020;26:886–891.

8. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence–enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol*. 2021;18:465–478.
9. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun*. 2020;11:1760.
10. Kwon JM, Kim KH, Jeon KH, et al. Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. *PLoS One*. 2019;14:e0219302.
11. Khurshid S, Friedman S, Reeder C, Di Achille P, Diamant N, Singh P, et al. Electrocardiogram-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation*. 2022;145:122–133.
12. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
13. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA*. 2000;283:2008–2012.
14. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155:529–536.
15. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med*. 2019;170:W1–W33.
16. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*. 2005;58:882–893.
17. Kashou AH, May AM, Noseworthy PA. Artificial Intelligence-Enabled ECG: A Modern Lens on an Old Technology. *Curr Cardiol Rep*. 2020;22:57.
18. Ko WY, Siontis KC, Attia ZI, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *J Am Coll Cardiol*. 2020;75:722–733.
19. Adedinsewo D, Carter RE, Attia ZI, et al. Artificial intelligence–enabled ECG algorithm to identify patients with left ventricular systolic dysfunction presenting to the emergency department with dyspnea. *Circ Heart Fail*. 2020;13:e007397.
20. Cho Y, Kwon JM, Kim KH, et al. Artificial intelligence algorithm for detecting myocardial infarction using six-lead electrocardiography. *Sci Rep*. 2020;10:20495.
21. Tison GH, Zhang J, Delling FN, Deo RC. Automated and interpretable patient ECG profiles for disease detection, tracking, and discovery. *Circ Cardiovasc Qual Outcomes*. 2019;12:e005289.
22. Ulloa-Cerna AE, Jing L, Pfeifer JM, et al. rECHOmmend: An ECG-based machine learning approach for identifying patients at increased risk of undiagnosed structural heart disease detectable by echocardiography. *Circulation*. 2022;146:36–47.
23. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: Are we there yet? *Heart*. 2018;104:1156–1164.
24. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS One*. 2019;14:e0213653.
25. Lee H, Shin M. Learning Explainable Time-Morphology Patterns for Automatic Arrhythmia Classification from Short Single-Lead ECGs. *Sensors (Basel)*. 2021;21:4331.
26. Wang J, Qiao X, Liu C, et al. Automated ECG classification using a non-local convolutional block attention module. *Comput Methods Programs Biomed*. 2021;203:106006.
27. Zhang H, Dong Z, Gao J, Lu P, Wang Z. Automatic screening method for atrial fibrillation based on lossy compression of the electrocardiogram signal. *Physiol Meas*. 2020;41:075005.
28. Smith SW, Walsh B, Grauer K, et al. A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. *J Electrocardiol*. 2019;52:88–95.
29. García CA, Bardají S, Pérez-Tirador P, Otero A. RHRVEasy:x Heart rate variability made easy. *PLoS One*. 2024;19:e0309055.
30. Zhu H, Cheng C, Yin H. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: A cohort study. *Lancet Digit Health*. 2020;2:e348–e357. Erratum in: Lancet Digit Health. 2020;2:e455.
31. Alhusseini MI, Abuzaid F, Rogers AJ, et al. Machine Learning to Classify Intracardiac Electrical Patterns During Atrial Fibrillation: Machine Learning of Atrial Fibrillation. *Circ Arrhythm Electrophysiol*. 2020;13:e008160.
32. Lin CH, Liu ZY, Chu PH, et al. A multitask deep learning model utilizing electrocardiograms for major cardiovascular adverse events prediction. *NPJ Digit Med*. 2025;8:1.
33. Patel HB, Yanamala N, Patel B, et al. Electrocardiogram-Based Machine Learning Emulator Model for Predicting Novel Echocardiography-Derived Phenogroups for Cardiac Risk-Stratification: A Prospective Multicenter Cohort Study. *J Patient Cent Res Rev*. 2022;9:98–107.
34. Goto S, Goto S. Application of Neural Networks to 12-Lead Electrocardiography - Current Status and Future Directions. *Circ Rep*. 2019;1:481–486.
35. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88.
36. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nat Med*. 2019;25:70–74.
37. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1:206–215.
38. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nat Med*. 2019;25:44–56.
39. Yao X, McCoy RG, Friedman PA, et al. ECG AI–Guided Screening for Low Ejection Fraction (EAGLE): Rationale and design of a pragmatic cluster randomized trial. *Am Heart J*. 2020;219:31–36.