

Artículo Especial

Inteligencia artificial multimodal en cardiología: síntesis crítica del rendimiento diagnóstico y la aplicabilidad clínica.

Multimodal AI in Cardiology: a critical review of diagnostic performance and clinical applicability

Fernando Ramos-Zaga.

Universidad Privada del Norte. Lima, Perú.

INFORMACIÓN DEL ARTÍCULO

Recibido el 17 de Agosto de 2026

Aceptado después de revisión

el 19 de Agosto de 2026

www.revistafac.org.ar

El autor declara no tener
conflicto de intereses.

Palabras clave:

Imágenes cardiovasculares,
señales biomédicas,
validación externa,
aprendizaje profundo,
transferencia clínica.

Keywords:

Cardiovascular imaging,
biomedical signals,
external validation,
deep learning,
clinical translation.

RESUMEN

Introducción: las enfermedades cardiovasculares mantienen una elevada carga asistencial y requieren estrategias diagnósticas capaces de integrar imágenes, señales biomédicas y datos clínicos. En este contexto, la inteligencia artificial ha mostrado utilidad potencial para apoyar la segmentación cardíaca, la caracterización tisular, la detección de arritmias, isquemia y valvulopatías, aunque su desempeño debe interpretarse según las condiciones de validación.

Objetivo: analizar críticamente el desempeño diagnóstico reportado de sistemas de inteligencia artificial en el análisis de imágenes cardiovasculares y señales biomédicas, considerando las condiciones de validación y las limitaciones metodológicas que restringen su transferencia al entorno asistencial.

Material y métodos: se realizó una revisión narrativa con orientación sistemática siguiendo PRISMA 2020. La búsqueda se efectuó en Scopus, Web of Science y PubMed, incluyendo artículos empíricos publicados entre 2021 y 2024. De 872 registros identificados, 17 estudios cumplieron los criterios de inclusión.

Resultados: los estudios basados en imágenes reportaron desempeños elevados en segmentación ecocardiográfica, con coeficientes Dice superiores a 0,90 en algunas tareas, aunque con limitaciones ante baja calidad de imagen, muestras reducidas o validación externa insuficiente. Los estudios basados en señales mostraron AUROC superiores a 0,90 en varias aplicaciones, pero con especificidades variables y caídas de rendimiento en cohortes externas.

Conclusión: la evidencia sugiere un potencial diagnóstico relevante, condicionado por heterogeneidad muestral, sobreajuste, sesgo de selección, métricas incompletas y debilidades regulatorias. Su transferencia clínica requiere validación prospectiva, externa y vigilancia continua.

Multimodal AI in Cardiology: a critical review of diagnostic performance and clinical applicability

ABSTRACT

Introduction: cardiovascular diseases continue to impose a substantial clinical burden and call for diagnostic strategies capable of integrating imaging, biomedical signals, and clinical data. In this context, artificial intelligence has shown promise in supporting cardiac segmentation, tissue characterization, and the detection of arrhythmia, ischemia, and valvular disease; however, its performance must be interpreted considering validation conditions.

Objective: to critically assess the reported diagnostic performance of artificial intelligence systems applied to cardiovascular imaging and biomedical signal analysis, with particular attention to validation settings and methodological constraints that may hinder clinical translation.

Methods: a narrative review with a systematic orientation was conducted in accordance with PRISMA 2020 guidelines. Searches were performed in Scopus, Web of Science, and PubMed, including empirical studies published between 2021 and 2024. Of 872 records identified, 17 studies met the inclusion criteria.

Results: imaging-based studies reported high performance in echocardiographic segmentation, with Dice coefficients exceeding 0.90 in selected tasks, although limitations emerged in the presence of low image quality, small sample sizes, or limited external validation. Signal-based studies achieved AUROC values above 0.90 in several applications, yet showed variable specificity and notable performance drops in external cohorts.

Conclusion: the available evidence points to meaningful diagnostic potential, shaped by sample heterogeneity, overfitting, selection bias, reliance on limited metrics, and regulatory shortcomings. Clinical translation is likely to depend on robust prospective validation, independent external testing, and ongoing monitoring.

INTRODUCCIÓN

Las enfermedades cardiovasculares continúan representando una de las principales causas de morbilidad global, con impacto particular en pacientes expuestos a barreras de acceso diagnóstico, seguimiento irregular o disponibilidad limitada de tecnologías especializadas^{1,2}. En este escenario, la población clínicamente relevante incluye sujetos con sospecha o diagnóstico de enfermedad cardiovascular evaluados mediante imágenes cardíacas, electrocardiografía, fonocardiografía u otras señales biomédicas. La carga asistencial creciente y la necesidad de decisiones oportunas han favorecido el interés por herramientas capaces de apoyar la interpretación diagnóstica en entornos con distintos niveles de complejidad³.

La inteligencia artificial aplicada a la cardiología puede definirse como el conjunto de sistemas computacionales orientados a identificar patrones diagnósticos en datos clínicos, imágenes cardiovasculares y señales biomédicas. Dentro de este campo, el aprendizaje automático y el aprendizaje profundo permiten modelar relaciones no lineales, procesar información no estructurada e integrar fuentes heterogéneas mediante enfoques multimodales^{4,5}. Dicha capacidad ha ampliado el análisis automatizado de fenómenos cardiovasculares, aunque también ha introducido desafíos relacionados con interpretabilidad, validación y trazabilidad de las decisiones algorítmicas.

En la práctica cardiológica, estos sistemas se han aplicado al análisis de ecocardiografía, resonancia magnética cardíaca, tomografía computarizada, electrocardiografía y fonocardiografía, con objetivos como segmentar cavidades, estimar función ventricular, detectar fibrosis, identificar arritmias o apoyar el cribado de valvulopatías. Algunos estudios han comunicado niveles elevados de discriminación, sensibilidad o concordancia técnica en tareas específicas, especialmente bajo condiciones controladas de evaluación⁶. Sin embargo, la utilidad clínica de estos resultados depende de su estabilidad frente a variaciones en población, calidad de adquisición, dispositivos, estándares de referencia y contexto asistencial⁷.

A pesar del crecimiento de la evidencia, persisten áreas no resueltas que dificultan una traslación directa al entorno asistencial. La heterogeneidad muestral, el predominio de diseños retrospectivos, la validación externa limitada, el riesgo de sobreajuste y el uso aislado de métricas técnicas pueden generar una apreciación incompleta del desempeño real^{8,9}. Además, la literatura tiende a favorecer la publicación de resultados positivos, lo que puede reforzar interpretaciones optimistas sobre sistemas cuya aplicabilidad aún requiere comprobación en escenarios clínicos diversos¹⁰.

Por ello, el presente artículo de revisión tiene como objetivo analizar críticamente el desempeño diagnóstico reportado de sistemas de inteligencia artificial en el análisis de imágenes cardiovasculares y señales biomédicas, considerando las condiciones de validación y las limitaciones metodológicas que restringen su transferencia al entorno asistencial. De ese modo, se busca situar la evidencia disponible en un marco clínico que permita valorar con mayor precisión el alcance real de estas herramientas.

MATERIAL Y MÉTODOS

La presente revisión narrativa se desarrolló siguiendo las directrices de la declaración PRISMA 2020, con el objetivo de asegurar un proceso transparente, reproducible y con control de sesgos metodológicos. La estrategia de búsqueda se diseñó para identificar evidencia sobre el uso de inteligencia artificial en el diagnóstico cardiovascular a partir de imágenes médicas y señales biomédicas. Las búsquedas se realizaron en las bases de datos Scopus, Web of Science y PubMed, seleccionadas por su cobertura en ciencias biomédicas y tecnológicas.

La ecuación booleana fue la siguiente: TITLE-ABS-KEY (("artificial intelligence" OR "machine learning" OR "deep learning") AND (cardiolog* OR "cardiovascular disease" OR cardiac) AND (echocardiograph* OR "cardiac MRI" OR "cardiac magnetic resonance" OR "computed tomography" OR electrocardiograph* OR ECG OR phonocardiograph*) AND (diagnos* OR detect* OR segment* OR "risk prediction") AND ("external validation" OR "clinical validation" OR generaliz* OR "real world" OR "prospective study")) AND (LIMIT-TO (OA, "all")) AND PUBYEAR > 2020 AND PUBYEAR < 2025.

Se establecieron criterios de inclusión para la selección de estudios. Se incluyeron estudios empíricos originales que abordaran el uso de inteligencia artificial en diagnóstico cardiovascular mediante imágenes o señales biomédicas, publicados en los últimos cinco años, en formato de artículo científico y con acceso completo al texto. Se requirió que los estudios presentaran una descripción de la población evaluada, los métodos de adquisición de datos y los instrumentos utilizados, así como validación interna o externa del modelo. Se consideraron estudios que incluyeran validación clínica o análisis de generalización.

Los criterios de exclusión incluyeron estudios no empíricos, tales como revisiones narrativas, editoriales, cartas al editor o resúmenes de congresos, así como aquellos que no se centraran en el diagnóstico cardiovascular o que abordaran aplicaciones no clínicas de la inteligencia artificial. También se excluyeron publicaciones fuera del rango temporal definido, estudios con datos incompletos o sin acceso al texto completo y aquellos con limitaciones metodológicas, como ausencia de grupo de comparación, falta de validación o diseño experimental débil. Asimismo, se descartaron estudios en los que no fue posible identificar la población evaluada o que no proporcionaban información sobre los instrumentos de recolección de datos.

El proceso de selección inició con 872 artículos, de los cuales se eliminaron 192 duplicados, quedando 680 registros únicos. En la revisión de títulos y resúmenes se excluyeron 548 por no ser pertinentes, permaneciendo 132. Luego, se eliminaron 54 documentos por tipo de publicación y 13 por diseño no elegible, reduciendo la muestra a 65 estudios para recuperación.

Durante la recuperación, 7 artículos no estuvieron disponibles, quedando 58 para evaluación en texto completo. En esta fase se excluyeron 18 por no cumplir criterios de inclusión, 12 por resultados no relevantes y 11 por limitaciones metodológicas, lo que condujo a la inclusión final de 17 estudios.

Limitaciones metodológicas para la transferencia asistencial

Si bien las métricas presentan limitaciones intrínsecas en su capacidad de representar la fidelidad anatómica, la fragilidad del diagnóstico por inteligencia artificial se profundiza al considerar las deficiencias en el diseño de los estudios de validación. La heterogeneidad en el tamaño muestral constituye el primer obstáculo para la interpretación de la evidencia disponible. En la literatura cardiovascular conviven investigaciones basadas en millones de registros con estudios piloto de menos de 100 sujetos¹¹. Esta disparidad impide realizar comparaciones directas, dado que la estabilidad de los resultados y la amplitud de los intervalos de confianza dependen estrechamente del volumen de datos analizado, lo que impide establecer comparaciones directas de eficacia entre modelos entrenados bajo escalas muestrales tan dispares¹².

El sobreajuste o overfitting representa el riesgo metodológico más crítico en modelos complejos entrenados con muestras pequeñas. Este fenómeno ocurre cuando el algoritmo memoriza las particularidades accidentales o el ruido del conjunto de entrenamiento en lugar de identificar patrones clínicos aplicables a otros pacientes¹³. En el aprendizaje profundo, la gran cantidad de parámetros ajustables permite que el modelo genere un rendimiento interno sobresaliente que se desvanece al enfrentar datos nuevos¹⁴. Aunque existen técnicas de regularización para mitigar este efecto, su presencia en muestras reducidas cuestiona la fiabilidad de las predicciones en la práctica diaria¹⁵.

La disponibilidad de cohortes masivas tampoco garantiza la validez de los resultados si estas carecen de representatividad poblacional. Un modelo puede mostrar una precisión estadística elevada en millones de casos procedentes de un único sistema sanitario, pero dichos resultados reflejan únicamente los protocolos y perfiles demográficos de ese entorno específico¹⁶. La precisión numérica no debe confundirse con la capacidad de reflejar la diversidad clínica, especialmente en patologías cuya presentación varía significativamente según el sexo, la edad y las comorbilidades, sugiriendo que la precisión numérica *in silico* no siempre es un indicador fiable de utilidad diagnóstica universal¹¹.

La transferencia clínica de estas herramientas se ve limitada por el predominio de diseños retrospectivos unicéntricos. Estos estudios utilizan datos históricos de una sola institución, lo que facilita el desarrollo pero restringe la generalización de los hallazgos¹⁷. Cada centro hospitalario posee una huella institucional única conformada por sus equipos médicos, protocolos de adquisición y prácticas de registro¹⁸. Al desplazar un algoritmo a un hospital con diferentes plataformas tecnológicas o poblaciones distintas, el rendimiento suele disminuir entre un 5% y un 20%¹⁹.

Esta brecha de validez externa se acentúa debido a la variabilidad técnica de los dispositivos médicos. En la imagen cardíaca, el fabricante del equipo y el software de reconstrucción introducen variaciones visuales que pueden confundir al algoritmo²⁰. Un modelo optimizado para un entorno con imágenes de alta calidad puede fallar al enfrentarse a la realidad asistencial, donde son frecuentes las señales incompletas o los registros con ruido técnico²¹. La utilidad real de la herramienta depende de su capacidad para sostener un desempeño aceptable bajo estas condiciones variables.

El sesgo de selección en las fases de desarrollo refuerza la distancia entre la investigación y la clínica. Las cohortes de entrenamiento suelen excluir a pacientes con presentaciones atípicas o datos incompletos, centrándose en casos con diagnósticos confirmados y seguimiento óptimo²². Este filtrado produce modelos diseñados para una versión idealizada de la medicina que no coincide con el paciente complejo de la práctica cotidiana²³. Asimismo, la concentración de datos en regiones geográficas específicas limita la aplicabilidad global de estas tecnologías²⁴.

El reporte selectivo de resultados positivos influye en la percepción de eficacia de la inteligencia artificial. La falta de transparencia en la comunicación de todas las configuraciones probadas y el sesgo de publicación hacia desempeños elevados distorsionan el estado del arte²⁵. La adopción sistemática de estándares como TRIPOD-AI resulta indispensable para garantizar que el clínico reciba información veraz sobre la capacidad diagnóstica real⁹. La transición hacia el uso asistencial exige, por tanto, una validación externa independiente y prospectiva que demuestre beneficios tangibles en los desenlaces de los pacientes²⁶.

El análisis de las limitaciones de validez clínica evidencia que el rendimiento de los modelos no puede interpretarse de manera aislada, sino en función de las condiciones metodológicas que lo sustentan. En consecuencia, resulta necesario examinar cómo dichas restricciones se manifiestan en aplicaciones concretas. A partir de esta base conceptual, la siguiente sección aborda el desempeño de la inteligencia artificial en el procesamiento de imágenes y señales biomédicas cardiovasculares, con el fin de evaluar la correspondencia entre los resultados reportados y su aplicabilidad clínica.

DESEMPEÑO CLÍNICO DE LA INTELIGENCIA ARTIFICIAL EN IMÁGENES Y SEÑALES CARDIOVASCULARES

Imágenes

El análisis del rendimiento operativo en la caracterización tisular y la segmentación anatómica revela una dependencia crítica entre la arquitectura computacional empleada y la naturaleza de la muestra. La variabilidad observada en los indicadores de precisión técnica responde a la heterogeneidad de los diseños experimentales y a los estándares de referencia utilizados para la validación, cuyas características fundamentales y hallazgos métricos se sintetizan en la *tabla 1*.

La arquitectura tecnológica de los estudios analizados se fundamenta principalmente en modelos de Deep Learning, destacando el uso de redes U-Net para segmentación multiescala y Redes Adversarias Convolucionales (GAN) para la delimitación multitiqueta de las cuatro cavidades cardíacas^{27,28}. De forma similar, se integran módulos de control de calidad automatizado para optimizar la selección de cuadros en ecocardiografía 2D²⁹. En el ámbito de la caracterización tisular, los enfoques varían desde algoritmos computacionales para tomografía (TC) nativa hasta redes neuronales convolucionales (CNN) aplicadas al análisis de textura en secuencias de cine-RM^{30,31}. Existe una discrepancia técnica notable en la detección de fibrosis, donde el uso de Máquinas de Vectores de Soporte (SVM) ha demostrado ser más eficaz que las CNN con datos limitados, alcanzando una precisión del 71%³².

TABLA 1.

Sistemas de inteligencia artificial aplicados al análisis de imágenes cardiovasculares.

Est.	Arquitectura IA (Índice)	Metodología (n + Gold Standard)	Desempeño (Métricas clave)
27	<i>Deep Learning</i> (U-Net) para segmentación multiescala.	15,000 estudios eco; validación externa en CAMUS vs. reportes clínicos.	Precisión vistas: 93%. Dice VI: 0.86–0.90 (vistas 2, 3, 4C) y 0.91 (externo).
28	Deep Convolutional Adversarial Network (GAN).	450 pacientes (<i>dataset</i> propio + CAMUS) vs. segmentación manual.	Dice: VI 92.1%, VD 86.3%, AI 89.6%, AD 91.4%. Corr. Vol. VI: 0.94.
29	<i>Deep Learning</i> con módulo de Control de Calidad (QC).	91 participantes (entrenamiento en CAMUS) vs. segmentación manual.	Especificidad: 96%. Mejora de <i>frames</i> utilizables: 76% a 96%.
30	Algoritmo computacional para TC nativa.	8 pacientes (Cáncer mama HER2+) vs. RM Cardíaca (LGE, T1, ECV).	ICC: 1.0 (IC 95%: 0.99–1). Correlación (r): 0.88 con LGE.
31	CNN - Análisis de textura en Cine-RM.	75 sujetos (estudio retrospectivo) vs. RM con Realce Tardío (LGE).	Precisión (Acc): 89% en validación y 70% en set de prueba independiente.
32	SVM (Support Vector Machine) vs. CNN.	642 estudios RM (50% fibrosis) vs. RM post-contraste (LGE).	SVM: Precisión 71%, Sensibilidad 72% (Superior a CNN en este contexto).

CNN: redes neuronales convolucionales; GAN: redes adversarias convolucionales; ICC: coeficiente de correlación intraclase; LGE: realce tardío de gadolinio; QC: control de calidad; RM: resonancia magnética; SVM: máquinas de vectores de soporte; VI: ventrículo izquierdo.

En cuanto a la robustez metodológica, se observa una marcada heterogeneidad en el tamaño de las muestras, que oscilan desde bases de datos masivas con 15,000 estudios ecocardiográficos hasta estudios piloto con solo 8 pacientes^{27,30}. Varios estudios coinciden en fortalecer su generalización mediante el uso del *dataset* externo CAMUS para la validación de sus algoritmos de segmentación^{27,28,29}. No obstante, otros diseños presentan limitaciones por sesgos de población específica, como cohortes exclusivas de mujeres con cáncer de mama HER2+, o muestran una caída en la robustez al enfrentar conjuntos de prueba independientes, evidenciando problemas de overfitting en modelos de cuantificación de daño tisular^{30,31}.

El rigor diagnóstico de las investigaciones se sustenta en la comparación directa con estándares de referencia clínicos consolidados. En la evaluación de la segmentación y función ventricular, se emplean reportes clínicos y la segmentación manual realizada por expertos como base de comparación^{28,29}. Esto coincide con el estándar de oro utilizado para la detección de fibrosis miocárdica, donde la Resonancia Magnética (RM) con Realce Tardío de Gadolinio (LGE), el mapeo T1 y el volumen extracelular (ECV) actúan como el estándar frente al cual se validan las nuevas técnicas de IA^{30,31,32}. Estos marcos de referencia aseguran que las métricas de precisión obtenidas tengan una correspondencia directa con la realidad patológica del paciente.

El desempeño y la capacidad de generalización presentan variaciones críticas según la modalidad de imagen y la tarea específica. En ecocardiografía, se reportan coeficientes Dice de 0.921 para el ventrículo izquierdo y una correlación de volumen de 0.94, además de una especificidad del 96% tras aplicar controles de calidad^{28,29}. En contraste, la detección de fibrosis muestra un rendimiento más irregular; mientras que un algoritmo para TC nativa alcanzó un coeficiente de correlación intraclase (ICC) de 1.0 en una muestra pequeña, otros modelos sufrieron una reducción de la precisión del 89% en validación al 70% en el set de prueba^{30,31}. Asimismo, se identifica una menor precisión diagnóstica en condiciones de baja calidad de imagen o presencia de fibrilación auricular, lo que subraya la necesidad de optimizar los modelos para

entornos clínicos complejos, donde un coeficiente de Dice elevado puede no ser suficiente para garantizar la precisión anatómica requerida en decisiones quirúrgicas o hemodinámicas críticas²⁷.

Señales

La capacidad predictiva de los sistemas de procesamiento de señales biomédicas se fundamenta en la extracción de patrones temporales y espaciales que, en entornos controlados, presentan niveles de sensibilidad superiores al diagnóstico clínico convencional. No obstante, la robustez de estas herramientas debe contrastarse con el volumen de las cohortes de entrenamiento y la rigurosidad de la validación externa independiente. En la *tabla 2* se presenta una relación detallada de la evidencia analizada, categorizada por arquitectura, metodología y métricas clave de desempeño diagnóstico.

La implementación de diversas arquitecturas tecnológicas define el estado actual del diagnóstico cardiológico digital, destacando el uso predominante de redes neuronales convolucionales y modelos de *deep learning* para el procesamiento de electrocardiogramas^{33,34,36}. De forma similar, se observa una tendencia hacia el *ensemble learning*, específicamente mediante métodos como *Stacking* y *Random Forest*, para mejorar la precisión en la detección de arritmias y soplos pediátricos^{35,41}. En contraste con los enfoques de señal única, los desarrollos más recientes integran arquitecturas multimodales que fusionan señales de fonocardiografía y electrocardiografía sincronizada mediante parches portátiles o estetoscopios digitales con algoritmos integrados^{40,42}. Estas tecnologías se complementan con el uso de máquinas de vectores de soporte (SVM) para el análisis de características temporales y espaciales en señales de vectorcardiografía³⁷.

En cuanto a la robustez metodológica, existe una disparidad significativa en las poblaciones analizadas, con estudios que emplean bases de datos masivas de hasta 552,372 electrocardiogramas para el entrenamiento de modelos predictivos y 135,476 registros en entornos de centros terciarios^{33,34}. Lo antes señalado coincide con el rigor de investigaciones multicéntricas que validan algoritmos con 595 pacientes para isquemia o 1,960 pacientes hospitalizados

TABLA 2.

Sistemas de inteligencia artificial aplicados al análisis de señales biomédicas cardiovasculares.

Est.	Arquitectura IA (Índice)	Metodología (n + Gold Standard)	Desempeño (Métricas clave)
33	Comparativa: IA Clásica vs. Deep Learning (DL).	135,476 ECGs (Retrospectivo) vs. Informe de cardiólogo.	Sensibilidad: 90% (ambos). Especificidad: 62% (Clásica) vs. 69% (DL).
34	Deep Learning explicable (predicción de PAF).	552,372 ECGs (318,321 pacientes) vs. Ritmo Sinusal (NSR).	AUROC: 0.905 ± 0.007 para predicción a un mes.
35	<i>Ensemble Learning</i> (Stacking, Bagging, AdaBoost).	100 registros (Physionet) vs. Parámetros morfológicos onda P.	Precisión: 92%. Sensibilidad: 88%. Especificidad: 96%. AUROC: 0.911.
36	CNN con Grad-CAM (Explicable).	57,273 segmentos Holter vs. Anotación de cardiólogo.	AUROC: 0.988. Sensibilidad: 97.1%. Precisión: 95.3%.
37	SVM (Características tempo-espaciales ECG/VCG).	429 sujetos (377 isquemia/52 control) vs. Validación externa.	AUC: 0.904 (Interno) y 0.814 (Externo). Precisión: 90.3%.
38	Machine Learning basado en Vectorcardiografía (VCG).	595 pacientes vs. Angiografía coronaria.	Precisión global: 82.5% (Mujeres) y 90.7% (Hombres).
39	Deep Learning (Sonidos cardíacos crudos).	499 pacientes (Multicéntrico) vs. Ecocardiografía.	Precisión: 84.1%–100%. Sensibilidad/Especificidad en estenosis mitral: 100%.
40	IA integrada en Estetoscopio Digital.	369 pacientes (Atención primaria) vs. Ecocardiografía + Expertos.	Sensibilidad: 94.1% (IA) vs. 41.2% (Médico). Especificidad: 84.5%.
41	Clasificadores ML (RF, SVM, Naive Bayes).	116 niños (estudio prospectivo) vs. Ecocardiografía.	Precisión: >90% en detección de cardiopatías congénitas (CHD).
42	Algoritmo sincronizado (ECG + Fonocardiograma).	1,960 pacientes hospitalizados vs. Ecocardiografía.	AUROC (LVEF < 40%): 0.91. Sensibilidad: 92.4%. Especificidad.: 76.8%.
43	Multimodal AI (Auscultación + ECG).	1,052 pacientes + cohorte externa (103) vs. Ecocardiografía.	AUC (LVEF $\leq$ 40%): 0.75. Mayor eficiencia que sensores individuales.

AUC/AUROC: área bajo la curva característica operativa del receptor; DL: aprendizaje profundo (deep learning); ECG: electrocardiograma; PCG: fonocardiograma; SVM: máquinas de vectores de soporte; VCG: vectorcardiografía.

para disfunción sistólica^{38,42}. Sin embargo, otros análisis presentan muestras más acotadas, como el uso de 100 registros de bases de datos públicas o cohortes de 116 pacientes pediátricos, recurriendo predominantemente a la validación cruzada para mitigar el sesgo por tamaño muestral^{35,41}. La validación externa independiente es un factor diferenciador presente solo en algunos diseños, lo que permite evaluar la capacidad de transferencia de los modelos a entornos clínicos reales^{37,43}.

El rigor diagnóstico de estas investigaciones se fundamenta en la comparación con estándares de oro clínicos consolidados. En la detección de valvulopatías y disfunción ventricular, la ecocardiografía actúa como el referente principal para validar la precisión de la inteligencia artificial^{39,40,41,42,43}. Por otro lado, la identificación de isquemia y arritmias se valida frente al informe de cardiólogos expertos o mediante angiografía coronaria en casos de estenosis asintomática^{33,36,38}. El alineamiento con criterios clínicos tradicionales asegura que las métricas de desempeño de la IA tengan una correlación directa con la práctica médica estándar, aunque la mayoría de los autores enfatizan la necesidad de validaciones prospectivas adicionales antes de una implementación generalizada.

El desempeño y la generalización de los modelos muestran niveles de eficacia elevados, aunque con variaciones según la patología. Se han reportado valores de AUROC superiores a 0.90 en la predicción de fibrilación auricular y detección de isquemia^{34,35,37}. Específicamente, el uso de inteligencia artificial en estetoscopios digitales demostró una sensibilidad del 94.1% frente al 41.2% del

personal de atención primaria en la detección de valvulopatías⁴⁰. No obstante, se observan discrepancias en la generalización; por ejemplo, se registró una caída del AUC de 0.904 en validación interna a 0.814 en validación externa al detectar isquemia miocárdica³⁷. De forma similar, mientras que algunos modelos alcanzan una especificidad del 100% para estenosis mitral, otros sistemas de detección de arritmias muestran una especificidad moderada de entre 62% y 69%, lo que sugiere un riesgo persistente de falsos positivos en el cribado poblacional^{33,39}.

La integración de la evidencia derivada de imágenes y señales biomédicas permite delimitar con mayor precisión el alcance real del desempeño de los modelos, evidenciando que los resultados técnicos dependen de condiciones metodológicas, poblacionales y operativas específicas. Sobre esta base, en la siguiente sección se abordan las deficiencias estructurales en la supervisión regulatoria, la ética algorítmica y los estándares de publicación científica, como determinantes del modo en que dicho desempeño es validado, interpretado y trasladado a la práctica clínica.

Deficiencias estructurales en la supervisión regulatoria, la ética algorítmica y los estándares de publicación científica

Las limitaciones técnicas y metodológicas analizadas en las secciones previas adquieren su pleno significado cuando se las considera en el contexto del ecosistema sistémico que regula, evalúa y disemina el conocimiento sobre inteligencia artificial médica, es decir, el conjunto de normas, instituciones y procesos que determinan cómo se desarrollan, validan y utilizan estas

tecnologías. Este ecosistema presenta deficiencias estructurales que configuran un entorno donde herramientas con limitaciones documentadas pueden alcanzar el mercado y la práctica clínica con una vigilancia insuficiente de sus efectos reales³⁵.

En este marco, la aprobación regulatoria de dispositivos médicos basados en inteligencia artificial ha seguido en la mayoría de las jurisdicciones, particularmente bajo el sistema estadounidense, una vía denominada 510(k), que es un proceso de autorización que permite aprobar un dispositivo si demuestra ser similar a otro previamente aprobado, originalmente diseñado para dispositivos de bajo riesgo⁴⁴. Esta vía no exige típicamente estudios clínicos prospectivos para la autorización, es decir, estudios realizados en tiempo real con pacientes nuevos, lo que ha permitido la entrada al mercado de algoritmos diagnósticos cuya evidencia de eficacia clínica resulta limitada.

Como consecuencia directa de este esquema, la auditoría sistemática de los dispositivos aprobados ha revelado puntuaciones de transparencia notablemente bajas, con escasa divulgación de los parámetros de entrenamiento, es decir, las condiciones bajo las cuales el modelo fue desarrollado, las características demográficas de las cohortes de validación y los enlaces a publicaciones científicas que respalden los desempeños declarados⁴⁴.

Desde una perspectiva conceptual, la doctrina de la equivalencia sustancial introduce una distorsión epistémica adicional cuando se aplica a tecnologías de inteligencia artificial. Los algoritmos modernos pueden ser considerados sustancialmente equivalentes a predicados aprobados décadas atrás, es decir, dispositivos anteriores que sirven como referencia regulatoria, generando linajes regulatorios que cuestionan la pertinencia de la evidencia original frente a las arquitecturas computacionales actuales⁴⁵. La acumulación de aprobaciones por equivalencia sin reevaluación sustantiva configura un sistema donde la novedad técnica no se acompaña de una evaluación clínica igualmente actualizada, situación que la literatura especializada ha denominado deuda regulatoria acumulada⁴⁶.

En términos operativos, la vigilancia tras la puesta en el mercado, conocida como *post-market surveillance*, que corresponde al seguimiento del desempeño y seguridad del dispositivo una vez que ya está en uso clínico, presenta debilidades estructurales que resultan particularmente problemáticas en el caso de algoritmos cuyo comportamiento puede alterarse mediante actualizaciones, reentrenamientos o cambios en las poblaciones a las que se aplican⁴⁷.

Dicha limitación se entiende mejor si se considera que los marcos regulatorios tradicionales fueron diseñados para dispositivos físicos con propiedades estables a lo largo del tiempo, mientras que los sistemas de aprendizaje automático presentan una dinamicidad inherente, es decir, pueden cambiar su comportamiento con nuevos datos, lo que demanda paradigmas de supervisión continua todavía no plenamente desarrollados⁴⁸. La iniciativa Software as a Medical Device, impulsada por la International Medical Device Regulators Forum, ha intentado adaptar los marcos a esta realidad, aunque su implementación efectiva enfrenta resistencias institucionales significativas⁴⁹.

Bajo esta misma problemática, los dilemas éticos derivados de la opacidad algorítmica se intersectan directamente con las deficiencias regulatorias descritas. La naturaleza de caja negra de

los modelos de aprendizaje profundo implica que sus decisiones no son fácilmente interpretables, lo que compromete principios bioéticos fundamentales como la transparencia, la rendición de cuentas y el consentimiento informado, ya que el clínico difícilmente puede explicar al paciente las bases sobre las que un algoritmo ha emitido una recomendación diagnóstica determinada⁵⁰. Esta opacidad no es meramente un problema técnico, sino un desafío ético sustantivo que afecta la relación clínica y la atribución de responsabilidad cuando el algoritmo produce errores con consecuencias para el paciente⁵¹.

En paralelo, el sesgo algorítmico constituye otra dimensión ética que las deficiencias regulatorias permiten que persista sin escrutinio adecuado. Este tipo de sesgo ocurre cuando el modelo produce resultados sistemáticamente diferentes para ciertos grupos de pacientes. La evidencia paradigmática proviene del análisis de un algoritmo comercial ampliamente utilizado en el sistema sanitario estadounidense, que demostró perpetuar inequidades raciales no por uso explícito de la variable raza, sino por la elección del costo sanitario como variable subrogada de la necesidad médica, es decir, utilizando el gasto como indicador indirecto de enfermedad⁵².

Cuando el dispositivo regulatorio no exige auditorías de equidad ni la divulgación de las características demográficas de las cohortes de desarrollo, los sesgos algorítmicos permanecen invisibilizados durante las etapas iniciales de implementación. Esta falta de transparencia dificulta la identificación de posibles diferencias en el desempeño entre subgrupos poblacionales, lo que limita la capacidad de anticipar efectos desiguales en contextos clínicos reales⁵³. En consecuencia, los sesgos solo se hacen evidentes cuando generan daños documentables en la práctica asistencial, momento en el cual su corrección implica intervenciones técnicas y organizativas más complejas, además de costos elevados y desafíos políticos vinculados a la rendición de cuentas.

Aun con estos avances, las directrices internacionales han buscado abordar estas deficiencias mediante principios para una inteligencia artificial confiable en salud, aunque su carácter no vinculante limita su impacto⁵⁴. En Europa, el Reglamento de Inteligencia Artificial ha introducido categorías de riesgo y obligaciones para sistemas con impacto significativo en salud, incluidos los médicos, aunque su aplicación aún enfrenta desafíos interpretativos⁵⁵. La falta de armonización entre marcos regulatorios sigue dificultando la implementación y circulación responsable de estas tecnologías⁴⁶.

Desde otra arista, la distorsión del conocimiento médico asociada a las prácticas de publicación constituye una dimensión adicional del problema. El sesgo de publicación, por el cual los estudios con resultados positivos se publican con mayor frecuencia que aquellos con resultados negativos, es especialmente marcado en inteligencia artificial médica, donde los modelos con alto rendimiento reciben mayor visibilidad²⁵. Como consecuencia, el estado del arte percibido se inclina hacia resultados favorables, mientras que los fallos de generalización o pérdida de desempeño en validaciones externas permanecen subrepresentados⁵⁶.

Como respuesta a esta limitación, los registros prospectivos de modelos predictivos han sido propuestos como mecanismo para mitigar el sesgo de publicación, exigiendo el registro previo

de los protocolos de desarrollo y validación antes de la generación de resultados⁵⁷. La adopción de esta práctica, sin embargo, dista de ser universal y enfrenta resistencias culturales en una comunidad científica acostumbrada a la flexibilidad metodológica del campo computacional. Las iniciativas de revisión por pares especializadas en inteligencia artificial médica, así como las listas de verificación específicas para reporte de modelos predictivos, constituyen esfuerzos importantes pero todavía insuficientes para garantizar la calidad sistemática de la literatura publicada⁵⁸.

Bajo este conjunto de tensiones, la intersección entre deficiencias regulatorias, dilemas éticos y distorsiones editoriales configura un ecosistema donde el desarrollo tecnológico supera las capacidades de evaluación crítica. En este contexto, el clínico debe adoptar una postura prudente frente a las afirmaciones de utilidad, exigiendo validación externa rigurosa, transparencia metodológica y monitorización tras la implementación antes de incorporar estas herramientas a la práctica⁵⁹. En cardiología, dada la elevada carga de enfermedad y el impacto potencial de los errores diagnósticos, este nivel de exigencia resulta especialmente necesario²⁴.

CONCLUSIONES

La evaluación del desempeño diagnóstico de la inteligencia artificial en cardiología puede entenderse no solo como un ejercicio de cuantificación técnica, sino también como un problema de naturaleza metodológica y epistémica. En este marco, el uso predominante de indicadores matemáticos tiende a asociarse con interpretaciones optimistas del rendimiento, las cuales no siempre se mantienen en contextos asistenciales reales. En consecuencia, el desempeño reportado de estas arquitecturas parece depender en gran medida de las condiciones bajo las cuales fueron validadas, lo que sugiere la necesidad de una lectura crítica y contextualizada de sus resultados.

En esta misma línea, la evidencia analizada sugiere la existencia de una divergencia entre la precisión estadística y la fidelidad anatómica observada en la práctica clínica. Diversas métricas ampliamente utilizadas presentan limitaciones en su sensibilidad espacial, lo que puede traducirse en una subestimación de errores en estructuras de menor tamaño o con contornos complejos. Asimismo, la noción de deuda regulatoria acumulada ofrece un marco interpretativo útil para examinar cómo ciertos mecanismos de aprobación podrían no requerir evaluaciones clínicas plenamente alineadas con la complejidad actual de los modelos, particularmente en escenarios de limitada transparencia algorítmica.

A partir de lo anterior, desde una perspectiva asistencial, la transferencia de estos sistemas parece estar condicionada por factores metodológicos recurrentes, entre ellos el predominio de diseños retrospectivos unicéntricos. La literatura disponible muestra que el rendimiento puede experimentar reducciones al aplicarse en entornos distintos a los de desarrollo, lo que refuerza la importancia de validaciones externas independientes. En este contexto, la práctica clínica podría beneficiarse de una interpretación prudente de los indicadores de desempeño, priorizando aquellos sistemas que demuestren estabilidad bajo condiciones heterogéneas, tanto a nivel poblacional como técnico.

Las líneas futuras de investigación podrían orientarse hacia el

desarrollo de estudios prospectivos y multicéntricos que permitan mitigar sesgos asociados a conjuntos de datos seleccionados. Del mismo modo, resulta pertinente avanzar en esquemas de vigilancia posterior a la implementación que consideren la naturaleza dinámica de los sistemas de aprendizaje automático. En paralelo, la consolidación de estándares de evaluación que integren métricas complementarias podría contribuir a una caracterización más precisa del desempeño, facilitando comparaciones más consistentes entre estudios.

Por consiguiente, la integración de sistemas algorítmicos en cardiología introduce cuestiones que trascienden la dimensión técnica, al involucrar aspectos relacionados con la práctica clínica, la responsabilidad profesional y la equidad en salud. En este sentido, el desarrollo futuro del campo parece depender no solo de avances computacionales, sino también de la capacidad de articular dichos avances con criterios clínicos verificables. Bajo esta perspectiva, el valor de la automatización diagnóstica podría entenderse en función de su desempeño sostenido en contextos reales, así como de su integración transparente dentro del proceso clínico.

Declaración de uso de IA

Se utilizó el modelo extenso de lenguaje GPT-5.3 de OpenAI con el propósito de identificar y corregir errores gramaticales y ortográficos. La instrucción aplicada fue: "detecta y corrige todos los errores de redacción y ortografía". Una vez obtenida la respuesta del sistema, se realizó una revisión del texto resultante para asegurar que conservara el tono e intención del manuscrito original.

BIBLIOGRAFIA

1. Mensah GA, Fuster V, Murray CJL, et al. Global burden of cardiovascular diseases and risks, 1990-2022. *J Am Coll Cardiol* 2023; 82: 2350 - 2473.
2. Roth GA, Mensah GA, Johnson CO, et al. Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 Study. *J Am Coll Cardiol* 2020; 76: 2982 - 3021.
3. Chong B, Jayabaskaran J, Jauhari SM, et al. Global burden of cardiovascular diseases: projections from 2025 to 2050. *Eur J Prev Cardiol* 2024; 32: 1001 -1015.
4. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521: 436 - 444.
5. Soenksen LR, Ma Y, Zeng C, et al. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digit Med* 2022; 5: 149.
6. Schilling M, Unterberg-Buchwald C, Lotz J, et al. Assessment of deep learning segmentation for real-time free-breathing cardiac magnetic resonance imaging at rest and under exercise stress. *Sci Rep* 2024; 14: 3754.
7. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019; 170: 51 - 58.
8. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019; 380: 1347 - 1358.
9. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024; 385: e078378.
10. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019; 1: e271-97.
11. Cai Y, Cai YQ, Tang IY, et al. Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review. *BMC Med* 2024; 22: 56.
12. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; 368: m441.
13. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge: MIT Press; 2016. DOI: 10.1007/s10710-017-9314-z

14. Ying X. An overview of overfitting and its solutions. *J Phys Conf Ser* **2019**; 1168: 022022.
15. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* **2014**; 15: 1929-58.
16. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* **2016**; 69: 245 - 247.
17. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* **2018**; 286: 800 - 809.
18. Futoma J, Simons M, Panch T, et al. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* **2020**; 2: e489-92.
19. Yu AC, Mohlin B, Eaton J. External validation of deep learning algorithms for cardiovascular imaging: a systematic review of performance degradation. *Eur Heart J Digit Health* **2024**; 5: 245 - 256.
20. Litjens G, Ciompi F, Wolterink JM, et al. State-of-the-art deep learning in cardiovascular image analysis. *JACC Cardiovasc Imaging* **2019**; 12: 1549 - 1565.
21. Glocker B, Robinson R, Castro DC, et al. Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects. *arXiv*. **2019**. Disponible en: <https://doi.org/10.48550/arXiv.1910.04597> Acceso 30 de Agosto de 2026.
22. Gianfrancesco MA, Tamang S, Yazdany J, et al. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* **2018**; 178: 1544 - 1547.
23. Ghassemi M, Naumann T, Schulam P, et al. A review of challenges and opportunities in machine learning for health. *AMIA Jt Summits Transl Sci Proc* **2020**; **2020**: 191 -200.
24. Khera R, Aminorroaya A, Dhingra LS, et al. Reporting standards for cardiovascular artificial intelligence research: a scientific statement from the American Heart Association. *Circulation* **2024**; 150: e438 - e453.
25. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* **2020**; 26: 1320 - 1324.
26. Sendak MP, D'Arcy J, Kashyap S, et al. A path for translation of machine learning products into healthcare delivery. *EMJ Innov* **2020**; 4: 19-00172.
27. Morales-Galán A, Lopez-Gutierrez P, Garrido-Oliver J, et al. Artificial intelligence for automatic echocardiography image view detection and left ventricle segmentation and ejection fraction prediction. *Eur Heart J Cardiovasc Imaging* **2025**; 26 (Supplement 1): jeac333.031.
28. Arafati A, Morisawa D, Avendi MR, et al. Generalizable fully automated multi-label segmentation of four-chamber view echocardiograms based on deep convolutional adversarial networks. *J R Soc Interface* **2020**; 17: 20200267.
29. Geven BW, Zhao D, Creamer SA, et al. Automated quality-controlled left heart segmentation from 2D echocardiography. En: *Statistical Atlases and Computational Models of the Heart. Vol. 14507*. Springer Nature Switzerland; **2024**. p. 98-107. Disponible en: https://doi.org/10.1007/978-3-031-52448-6_10
30. Gonciar D, Berciu AG, Dulf EH, et al. Computer-assisted algorithm for quantification of fibrosis by native cardiac CT: a pilot study. *J Clin Med* **2024**; 13: 4807.
31. Curiale AH, Cabrera F, Jimenez P, et al. Detection of fibrosis in cine magnetic resonance images using artificial intelligence techniques. *arXiv*. **2022**. Disponible en: <https://doi.org/10.48550/ARXIV.2206.01745> Acceso 28 de Agosto de 2026.
32. Campese S, Agostini F, Sciarretta T, et al. Myocardial fibrosis detection using kernel methods: preliminary results from a cardiac magnetic resonance study. *Eur Heart J Cardiovasc Imaging* **2022**; 23 (Supplement 2): jeac141.005.
33. Mota D, Filho FNB, Sousa EB, et al. Comparative analysis of classical artificial intelligence and deep learning models for atrial fibrillation detection in electrocardiograms. *Europace* **2024**; 26 (Supplement 1): euae102.583.
34. Jin Y, Ko B, Chang W, et al. Explainable paroxysmal atrial fibrillation diagnosis using an artificial intelligence-enabled electrocardiogram. *Korean J Intern Med* **2025**; 40: 251 - 261.
35. Wu C, Hwang M, Huang TH, et al. Application of artificial intelligence ensemble learning model in early prediction of atrial fibrillation. *BMC Bioinformatics* **2021**; 22: 93.
36. Taniguchi H, Takata T, Takechi M, et al. Explainable artificial intelligence model for diagnosis of atrial fibrillation using Holter electrocardiogram waveforms. *Int Heart J* **2021**; 62: 534 - 539.
37. Zhao X, Zhang J, Gong Y, et al. Reliable detection of myocardial ischemia using machine learning based on temporal-spatial characteristics of electrocardiogram and vectorcardiogram. *Front Physiol* **2022**; 13: 854191.
38. Braun T, Spiliopoulos S, Veltman C, et al. Detection of myocardial ischemia due to clinically asymptomatic coronary artery stenosis at rest using supervised artificial intelligence-enabled vectorcardiography. *J Electrocardiol* **2020**; 59: 100 - 105.
39. Jiang Z, Song W, Yan Y, et al. Automated valvular heart disease detection using heart sound with a deep learning algorithm. *IJC Heart Vasc* **2024**; 51: 101368.
40. Rancier MA, Israel I, Monickam V, et al. Abstract 13244: real world evaluation of an artificial intelligence enabled digital stethoscope for detecting undiagnosed valvular heart disease in primary care. *Circulation* **2023**; 148 (Suppl 1). Disponible en: https://doi.org/10.1161/circ.148.suppl_1.13244 Acceso 28 de Agosto de 2026.
41. Begic E, Gurbeta Pokvic L, Begic Z, et al. From heart murmur to echocardiography - congenital heart defects diagnostics using machine-learning algorithms. *Psychiatria Danubina* **2021**; 33(Suppl 13): 236 - 246.
42. Zhang WL, Song B, Huang QJ, et al. Artificial intelligence for automated left ventricular systolic dysfunction detection using a wearable cardiac patch incorporated with synchronized phonocardiogram and electrocardiogram. *Eur Heart J* **2024**; 45 (Supplement 1): ehac666.3551.
43. Shiraga T, Makimoto H, Kohlmann B, et al. Improving valvular pathologies and ventricular dysfunction diagnostic efficiency using combined auscultation and electrocardiography data: a multimodal AI approach. *Sensors* **2023**; 23: 9834.
44. Almarie B, Gonzalez-Gonzalez LF, dos Santos Barbosa LA, et al. Machine learning-enabled medical devices authorized by the US Food and Drug Administration in 2024: regulatory characteristics, predicate lineage, and transparency reporting. *Biomedicines* **2025**; 13: 3005.
45. Gerke S, Babic B, Evgeniou T, et al. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digit Med* **2020**; 3: 53.
46. Larson DB, Harvey H, Rubin DL, et al. Regulatory frameworks for development and evaluation of artificial intelligence-based diagnostic imaging algorithms: summary and recommendations. *J Am Coll Radiol* **2021**; 18: 413 - 424.
47. Feng J, Phillips RV, Malenica I, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit Med* **2022**; 5: 66.
48. Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. *Lancet Digit Health* **2021**; 3: e337 - e338.
49. Pesapane F, Volonté C, Codari M, et al. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging* **2018**; 9: 745 - 753.
50. Char DS, Shah NH, Magnus D. Implementing machine learning in health care: addressing ethical challenges. *N Engl J Med* **2018**; 378: 981 - 983.
51. Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell* **2019**; 1: 501 - 507.
52. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**; 366: 447 - 453.
53. Rajkomar A, Hardt M, Howell MD, et al. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* **2018**; 169: 866 - 872.
54. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; **2021**. Disponible en: <https://www.who.int/publications/i/item/9789240029200> Acceso 30 de Agosto de 2028.
55. Floridi L. The European legislation on AI: a brief analysis of its philosophical approach. *Philos Technol* **2023**; 36: 27.
56. Ioannidis JPA. The challenge of reforming nutritional epidemiologic research. *JAMA* **2018**; 320: 969 - 970.
57. Heus P, Damen JAA, Pajouheshnia R, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ Open* **2020**; 9: e025611.
58. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell* **2020**; 2: e200029.
59. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* **2019**; 25: 44 - 56.