ORIGINAL

Fiabilidad y validez de un sistema asistido por inteligencia artificial para la detección de anomalías en las radiografías de tórax y óseas en un servicio de urgencias hospitalario

Raissa de Fátima Silva Afonso[1,7], Pilar Gallardo-Rodríguez[1,7], Begoña Espinosa[2,7], Alejandro Bautista[2], Javier Serrano[2], Mónica Veguillas[2], María Corell[2], Raúl Garrido Chamorro[2], Juan Arenas Jiménez[3,7], Celia Astor Rodríguez[3], Álvaro Abellón Fernández[3], Álvaro Palazón Ruíz de Tremiño[3], María Javiera Garfias Baladrón[3], Víctor Marquina Arribas[3], Pablo Chico-Sánchez[1,7,8], Paula Gras Valenti[1,7,8], Miguel Cabrer González[4], Carlos Martínez Riera[5], David Moliner Mateu[5], José María Salinas Serrano[6], Emilio Vivancos Rubio[5,9], Bernardo Valdivieso Martínez[5], Luis Concepción-Aramendia[3,7,8], José Sánchez-Payá[1,7], Pere Llorens[2,7,8]

Introducción. Evaluar el rendimiento diagnóstico para la detección de anomalías de dos sistemas comerciales de inteligencia artificial (IA), ChestView para radiografías de tórax (RxT) y BoneView para radiografías óseas (RxO), en un servicio de urgencias hospitalario (SUH), y comparar su validez con la de observadores de diferente perfil profesional y experiencia: urgenciólogos, radiólogos en formación y radiólogos expertos.

Método. Estudio de evaluación de pruebas diagnósticas en una selección aleatoria de 346 RxT y 261 RxO solicitadas en urgencias. Las exploraciones fueron analizadas de forma independiente por los sistemas de IA y los diferentes observadores. El diagnóstico de referencia *(gold standard)* fue establecido mediante consenso por tres radiólogos, recurriendo a otras pruebas de imagen disponibles o información clínica cuando era necesario. Se calcularon y compararon la sensibilidad, especificidad, y los valores predictivos positivo y negativo (VPN).

Resultados. Para la RxT, la IA (ChestView) mostró una sensibilidad global (64,4%) significativamente superior a la de los médicos de urgencias (49,2%; p = 0,018), aunque inferior a la del radiólogo experto (83,9%; p < 0,001). El rendimiento fue notable para la detección de nódulos/masas (sensibilidad 80,0%) y neumotórax (VPN 99,7%), pero inferior para consolidaciones (sensibilidad 40,4%). Para RxO, la IA (BoneView) alcanzó una sensibilidad para la detección de fracturas (87,5%) superior a la del radiólogo experto (77,1%), con un VPN del 96,9%. Sin embargo, su rendimiento fue menor para la detección de luxaciones (sensibilidad 60,0%) y derrames articulares (sensibilidad 25,0%).

Conclusión. Los sistemas de IA evaluados demuestran un rendimiento clínicamente relevante en el entorno de urgencias, y mejorar significativamente la capacidad diagnóstica de los urgenciólogos. Su elevada sensibilidad para la detección de fracturas y su elevado VPN para los nódulos pulmonares, neumotórax y fracturas hacen que se consolide como un sistema de seguridad de alto impacto.

Palabras clave: Radiología. Radiografía de tórax. Radiografía ósea. Inteligencia artificial. Urgencias.

Reliability and validity of an artificial intelligence-assisted system for the detection of abnormalities in chest and bone radiographs in an emergency department

Introduction. To evaluate the diagnostic performance of two commercial artificial intelligence (AI) systems—ChestView for chest radiographs (CXR) and BoneView for bone radiographs (BXR)—in an emergency department (ED), and compare their validity with that of observers with different professional profiles and levels of experience: emergency physicians, radiology trainees, and expert radiologists.

Method. We conducted a diagnostic test evaluation study on a random selection of 346 CXRs and 261 BXRs requested in the ED. Examinations were independently analysed by the AI systems and the various observers. The reference diagnosis (gold standard) was established by consensus among 3 radiologists, resorting to additional imaging tests or clinical information when necessary. Sensitivity, specificity, and positive and negative (NPV) predictive values were then calculated and compared.

Results. For CXRs, AI (ChestView) showed overall sensitivity (64.4%) significantly higher than that of emergency physicians (49.2%; *P* = .018), although lower than that of the expert radiologist (83.9%; *P* < .001). Performance was notable for the detection of nodules/masses (sensitivity 80.0%) and pneumothorax (NPV, 99.7%), but lower for consolidations (sensitivity, 40.4%). For BXRs, AI (BoneView) achieved sensitivity for fracture detection (87.5%) higher

Filiación de los autores:
[1]Servicio de Medicina Preventiva, Hospital General Doctor Balmis, Alicante, España.
[2]Servicio de Urgencias, Unidad de Corta Estancia y Hospitalización a Domicilio, Hospital General Doctor Balmis, Alicante, España.
[3]Servicio de Radiodiagnóstico, Hospital General Doctor Balmis, Alicante, España.
[4]Plataforma Digital Idonia, Barcelona, España.
[5]Secretaría Autonómica de Planificación, Información, y Transformación Digital, Conselleria de Sanitat, Generalitat Valenciana.
[6]Servicio de Informática, Hospital San Juan, Alicante, España.
[7]Instituto de Investigación Sanitaria y Biomédica de Alicante, ISABIAL, España.
[8]Universidad de Alicante, Alicante, España.
[9]Valencian Research Institute for Artificial Intelligence (VRAIN), España.

Contribución de los autores:
Todos los autores han confirmado su autoría en el documento de responsabilidades del autor, acuerdo de publicación y cesión de derechos a EMERGENCIAS.

Autor para correspondencia:
Pere Llorens
Servicio de Urgencias
Hospital General Universitario Dr. Balmis
C/Pintor Baeza, 12
03010, Alicante, España

Correo electrónico:
llorens_ped@gva.es

Información del artículo:
Recibido: 22-8-2025
Aceptado: 17-9-2025
Online: 22-10-2025

Editor responsable:
Agustín Julián-Jiménez

DOI:
XXXX

than that of the expert radiologist (77.1%), with an NPV of 96.9%. However, its performance was lower for detecting dislocations (sensitivity 60.0%) and joint effusions (25.0%).

Conclusions. The evaluated AI systems demonstrate clinically relevant performance in the emergency setting, significantly enhancing the diagnostic capacity of emergency physicians. Their high sensitivity for fracture detection and high NPV for pulmonary nodules, pneumothorax, and fractures establish them as a high-impact safety tool.

Keywords: Radiology. Chest radiograph. Bone radiograph. Artificial intelligence. Emergency medicine.

DOI: XXXXX

Introducción

Las radiografías óseas (RxO) y de tórax (RxT) son las pruebas de imagen más solicitadas en los servicios de urgencias hospitalarios (SUH), y generan una carga asistencial importante sobre los servicios de radiodiagnóstico[1]. Esta demanda creciente choca con una escasez estructural de radiólogos, lo cual crea un desequilibrio crítico entre el volumen de trabajo y el déficit laboral actual[2]. La presión por emitir informes con celeridad, sumada a la fatiga y a la delegación de la interpretación inicial en personal menos experto, elevan la tasa de error diagnóstico, que se sitúa en torno al 3%[3].

Los errores más frecuentes son precisamente las fracturas sutiles, que representan hasta el 80% de los fallos diagnósticos, y las enfermedades torácicas críticas como nódulos pulmonares, consolidaciones o neumotórax. Estos errores no suelen deberse a una falta de conocimiento, sino a fallos de percepción y sesgos cognitivos exacerbados por el entorno de alta presión de las urgencias, el cese prematuro de la búsqueda tras un hallazgo evidente *(satisfaction of search)* o la fatiga visual[4]. Esta situación crea un ámbito idóneo para la aplicación tecnológica.

La inteligencia artificial (IA), basada en aprendizaje profundo, ha surgido como una herramienta con el potencial de transformar este paradigma[5-8]. Sus aplicaciones van desde la optimización del flujo de trabajo mediante la priorización inteligente de casos urgentes hasta actuar como un "segundo lector", lo que aumenta la sensibilidad diagnóstica, especialmente para los médicos con menor experiencia[9]. Múltiples estudios han demostrado que la IA reduce los tiempos de lectura y mejora la eficiencia, por lo que aborda directamente los problemas de sobrecarga y retrasos[10].

La evidencia científica ha validado el alto rendimiento de la IA en la detección de fracturas en diversas localizaciones y en la identificación de anomalías torácicas clave como el neumotórax, el derrame pleural y las consolidaciones[8,9]. Sin embargo, el rendimiento puede variar según la enfermedad y los factores técnicos, y gran parte de la investigación existente se ha realizado en entornos controlados o se ha centrado en una única tarea diagnóstica[11]. Es por ello que en la actualidad existe una brecha entre la evidencia publicada y la utilidad clínica real.

Este estudio se justifica por la necesidad de realizar una validación externa e independiente de una plataforma de IA comercial (Gleamer©) capaz de analizar simultáneamente un amplio espectro de enfermedades óseas y torácicas en un entorno de un SUH no seleccionado. El objetivo del estudio fue validar su rendimiento diagnóstico y compararlo con la práctica clínica habitual y valorar su utilidad clínica en el SUH[1,2].

Método

Se ha realizado un estudio observacional de tipo transversal de evaluación de pruebas diagnósticas radiológicas (RxT y RxO) realizadas en un SUH de un hospital terciario durante el 2024.

Criterios de inclusión y exclusión

Se determinaron como criterios de inclusión las RxT solicitadas por el SUH por sospecha de enfermedad cardiopulmonar o torácica y RxO por sospecha de enfermedad traumatológica. Los criterios de exclusión que se establecieron fueron los siguientes: radiografías que por cuestiones técnicas se consideren que no tienen la calidad suficiente como para ser evaluadas, imágenes de seguimiento de enfermedad cardiopulmonar o torácica conocida, radiografías no completas de tórax y radiografías esqueléticas con otras investigaciones (por ejemplo, enfermedad inflamatoria, radiografías postquirúrgicas, etc.) y radiografías de cráneo o cara.

Cálculo del tamaño muestral

Partiendo de que en el año 2023 se habían realizado aproximadamente 44.350 RxT solicitadas por el SUH por sospecha de enfermedad cardiopulmonar o torácica, que aproximadamente el 34% de ellas mostraba anomalías patológicas y teniendo en cuenta un error alfa del 5%, y que un 5% tenga algún criterio de exclusión, el número estimado de RxT a incluir en el estudio fue de 360. Tras la selección mediante un muestreo aleatorio simple de estas, y que estas cumplieran el criterio de inclusión y ninguno de exclusión se incluyeron en el estudio 346 RxT. A su vez, partiendo de que en el año 2023 se habían realizado aproximadamente 40.515 RxO solicitadas por el SUH por sospecha de enfermedad traumatológica de las que aproximadamente el 22 % contenían anomalías patológicas y teniendo en cuenta un error alfa del 5%, y que un 5% tengan algún criterio de exclusión, el número estimado de RxO a incluir en el estudio fue de 275. Tras la selección mediante un muestreo aleatorio simple de estas, y que estas cumplieran el criterio de inclusión y ninguno de exclusión, se incluyeron en el estudio 261 RxO.

Variables

Para las RxT se recogieron las siguientes variables: observador (ChestView –este emplea el algoritmo de *Deep Learning Detectron2*–, informe de urgencias, facultativo urgencias, radiólogo en formación, radiólogo experto, radiólogo de referencia en RxT, estándar –casos concordantes entre el radiólogo en formación y el radiólogo experto, y las discordancias resueltas por el radiólogo de referencia–), presencia de anomalía (Sí/No) y tipo de anomalía (neumotórax, derrame pleural, consolidación y nódulo o masa pulmonar). Para las RxO: observador (BoneView, informe de urgencias, facultativo urgencias, radiólogo en formación, radiólogo experto y radiólogo de referencia en RxO), presencia de anomalía (Sí/No) y tipo de anomalía (fractura, luxación y derrame). Cada observador realizó la lectura de las Rx seleccionadas de manera independiente. El informe de urgencias hace referencia al informe radiológico que realizó un urgenciólogo o médico en formación durante su actividad asistencial en urgencias, ya que en este SU no se realizan informes radiológicos sistemáticos por parte del radiólogo, y solo bajo demanda del urgenciólogo. El radiólogo de referencia únicamente lee las Rx en las cuales hay discrepancias entre el radiólogo en formación y el radiólogo experto y para su evaluación utilizó toda la información clínica disponible (historia clínica, otras pruebas de diagnóstico por imagen disponibles, etc.).

Anállsis estadístico

En el análisis de los datos, en primer lugar se ha realizado el estudio de la fiabilidad. Para ello, con las 49 primeras RxT y las 47 RxO que fueron leídas por las herramientas ChestView y BoneView respectivamente dos veces, se ha calculado el Índice Kappa ponderado (IK) con sus intervalos de confianza al 95% (IC 95%). A continuación, se ha procedido al estudio de la validez, y para ello, se ha calculado la sensibilidad (S), especificidad (E), valor predictivo positivo (VPP) y valor predictivo negativo (VPN) de cada uno de los observadores, y se comparó la frecuencia de detección de anomalías de cada uno de ellos con el estándar. A continuación se ha comparado la S, E, VPP y VPN de cada observador y de cada tipo de lesión, con la S, E, VPP y VPN de las herramientas ChestView y BoneView, y para ello se ha utilizado la prueba de la ji cuadrado o la prueba exacta de Fisher en caso necesario. El nivel de significación estadística utilizado en los contrastes de hipótesis realizadas ha sido de $p < 0,05$ y los programas de análisis estadístico que se han utilizado son el IBM-SPSS V.25.0 y el programa Epidat 3.1.

Se siguieron los principios éticos de la Declaración de Helsinki sobre investigación en humanos. El protocolo del estudio fue aprobado por el Comité de Ética e Investigación Clínica con medicamentos (CEIm) del Hospital General Universitario Dr. Balmis de Alicante (número de referencia: 2024-0042) con exención del consentimiento informado.

Resultados

Estudio de fiabilidad

Los estudios de fiabilidad iniciales demostraron una alta consistencia de los sistemas de IA, con un Índice Kappa ponderado para la detección de anomalías del 0,93 para ChestView y del 0,96 para BoneView (Tabla 1 y Tabla 2).

Rendimiento de ChestView para radiografías de tórax

La sensibilidad global del sistema ChestView para la detección de anomalías fue del 64,4%. Esta sensibilidad fue estadísticamente superior a la del facultativo de urgencias y al informe de urgencias (49,2%, $p = 0,018$ y 39,8, $p < 0,001$ respectivamente), e inferior a la del radiólogo experto (83,9%, $p < 0,001$). Para los nódulos/masas pulmonares, la sensibilidad del sistema alcanzó el 80,0%, un valor equivalente al del radiólogo experto. En contraste, la sensibilidad para la detección de consolidaciones fue del 40,4%, un valor inferior al del radiólogo experto y en formación (59,6%, $p = 0,050$ y 71,2%, $p < 0,001$, respectivamente).

El VPN del sistema ChestView para el neumotórax fue del 99,7%, del 99,4% para el nódulo o masa pulmonar, del 93,2% para la masa hiliomediastínica y del 90% para la consolidación pulmonar. El VPN global de ChestView (83,0%) fue estadísticamente superior al del informe de urgencias (75,8%, $p = 0,040$), pero inferior al del radiólogo experto (92,0%, $p = 0,003$) (Tabla 3).

Rendimiento de BoneView para radiografías óseas

La sensibilidad global del sistema BoneView para la detección de anomalías fue del 84,5%, un valor comparable al del radiólogo en formación y al del radiólogo

Tabla 1. Fiabilidad de las lecturas de la herramienta de inteligencia artificial ChestView (n = 49)

	Casos concordantes n (%)	Casos discordantes n (%)	Índice Kappa (IC 95%)
Anomalías	48 (98,0)	1 (2,0)	0,93 (0,79-1,00)
Neumotórax	49 (100,0)	0 (0,0)	1,0 (–)
Derrame pleural	48 (98,0)	1 (2,0)	0,88 (0,64-1,00)
Consolidación	49 (100,0)	0 (0,0)	1,0 (–)
Masa hiliomediastínica	49 (100,0)	0 (0,0)	1,0 (–)
Nódulo/masa pulmonar	49 (100,0)	0 (0,0)	1,0 (–)

IC 95%: intervalo de confianza al 95%.

Tabla 2. Fiabilidad de las lecturas de la herramienta de inteligencia artificial BoneView (n = 47)

	Casos concordantes n (%)	Casos discordantes n (%)	Índice Kappa (IC 95%)
Anomalías	46 (97,9)	1 (2,1)	0,96 (0,87-1,00)
Fractura	46 (97,9)	1 (2,1)	0,95 (0,85-1,00)
Luxación	47 (100,0)	0 (0,0)	1,0 (-)
Derrame	47 (100,0)	0 (0,0)	1,0 (-)

IC 95%: intervalo de confianza al 95%.

Tabla 3. Validez de distintos observadores en la detección de anomalías en radiografías de tórax solicitadas en un servicio de urgencias (n = 346)

	VP (n)	FN (n)	FP (n)	VN (n)	Sensibilidad (%)	p1	Especificidad (%)	p2	VPP (%)	p3	VPN (%)	p4
Anomalías (N = 118)												
Inteligencia artificial (ChestView)	76	42	23	205	64,4		89,9		76,8		83,0	
Informe de urgencias	47	71	6	222	39,8	< **0,001**	97,4	**0,001**	88,7	0,075	75,8	**0,040**
Facultativo urgencias	58	60	35	193	49,2	**0,018**	84,6	**0,092**	62,4	**0,030**	76,3	0,063
Radiólogo en formación	88	30	18	210	74,6	0,090	92,1	0,413	83,0	0,264	87,5	0,162
Radiólogo experto	99	19	9	219	83,9	< **0,001**	96,1	**0,010**	91,7	**0,003**	92,0	**0,003**
Neumotórax (N = 4)												
Inteligencia artificial (ChestView)	3	1	4	338	75,0		98,8		42,9		99,7	
Informe de urgencias	2	2	0	342	50,0	0,465	100	0,368	100	1,000	99,4	0,571
Facultativo urgencias	2	2	0	342	50,0	0,465	100	0,368	100	1,000	99,4	0,571
Radiólogo en formación	3	1	1	341	75,0	1,000	99,7	0,178	75,0	0,303	99,7	0,995
Radiólogo experto	4	0	1	341	100	0,530	99,7	0,369	80,0	0,198	100	0,483
Derrame pleural (N = 58)												
Inteligencia artificial (ChestView)	31	27	5	283	53,4		98,3		86,1		91,3	
Informe de urgencias	27	31	6	282	46,6	0,458	97,9	0,761	81,8	0,627	90,1	0,608
Facultativo urgencias	24	34	14	274	41,4	0,193	95,1	**0,036**	63,2	**0,024**	89,0	0,332
Radiólogo en formación	41	17	19	269	70,7	0,056	93,4	**0,004**	68,3	0,052	94,1	0,197
Radiólogo experto	50	8	14	274	86,2	< **0,001**	95,1	**0,036**	78,1	0,329	97,2	**0,003**
Consolidación (N = 52)												
Inteligencia artificial (ChestView)	21	31	14	280	40,4		95,2		60,0		90,0	
Informe de urgencias	11	41	4	290	21,2	**0,034**	98,6	**0,017**	73,3	0,368	87,6	0,332
Facultativo urgencias	19	33	24	270	36,5	0,687	91,8	0,094	44,2	0,165	89,1	0,708
Radiólogo en formación	37	15	22	272	71,2	**0,002**	92,5	0,169	62,7	0,794	94,8	**0,030**
Radiólogo experto	31	21	6	288	59,6	0,050	98,0	0,069	83,8	0,024	93,2	0,154
Masa hiliomediastínica (N = 29)												
Inteligencia artificial (ChestView)	7	22	16	301	24,1		95,0		30,4		93,2	
Informe de urgencias	3	26	2	315	10,3	0,164	99,4	< **0,001**	60,0	0,211	92,4	0,686
Facultativo urgencias	5	24	7	310	17,2	0,517	97,8	0,056	41,7	0,506	92,8	0,851
Radiólogo en formación	5	24	1	316	17,2	0,517	99,7	< **0,001**	83,3	**0,019**	92,9	0,900
Radiólogo experto	7	22	1	316	24,1	1,000	99,7	< **0,001**	87,5	**0,005**	93,5	0,876
Nódulo/masa pulmonar (N = 10)												
Inteligencia artificial (ChestView)	8	2	22	314	80,0		93,5		26,7		99,4	
Informe de urgencias	2	8	1	335	20,0	**0,007**	99,7	< **0,001**	66,7	0,151	97,7	0,075
Facultativo urgencias	1	9	16	320	10,0	**0,002**	95,2	0,316	5,9	0,082	97,3	**0,039**
Radiólogo en formación	2	8	7	329	20,0	**0,007**	97,9	**0,004**	22,2	0,789	97,6	0,070
Radiólogo experto	8	2	7	329	80,0	1,000	97,9	**0,004**	53,3	0,078	99,4	0,963

VP: verdadero positivo; FN: falso negativo; FP: falso positivo; VN: verdadero negativo.
Sensibilidad = VP/VP + FN. Especificidad = VN/VN + FP. VPP: VP/VP + FP. VPN: VN/VN + FN.
p1: diferencia entre la sensibilidad de los distintos observadores y la inteligencia artificial (ChestView). p2: diferencia entre la especificidad de los distintos observadores y la inteligencia artificial (ChestView). p3: diferencia entre el VPP de los distintos observadores y la inteligencia artificial (ChestView). p4: diferencia entre el VPN de los distintos observadores y la inteligencia artificial (ChestView).
Los valores en negrita denotan significación estadística (p < 0,05).

experto (86,2% y 81,0% respectivamente). Para la detección de fracturas, el sistema alcanzó una sensibilidad del 87,5%, que fue estadísticamente superior a la del informe de urgencias (70,8%, p = 0,044), y equivalente a la del facultativo de urgencias y la del radiólogo en formación. En cuanto a otras anomalías, la sensibilidad del sistema para luxaciones fue del 60,0% y para derrames articulares fue del 25,0%.

El VPN del sistema BoneView para el global de anomalías fue del 95%, para fracturas del 96,9%, del 99,2% para luxaciones y del 96,4% para el derrame articular, sin encontrar diferencias significativas con el resto de observadores (Tabla 4).

En la Tabla 3 para radiografías de tórax y la Tabla 4 para radiografías óseas se detallan todos los parámetros de validez y las comparaciones estadísticas.

Discusión

Este estudio evaluó la validez y fiabilidad de dos sistemas de IA, ChestView y BoneView, para la interpretación de RxT y RxO en un SUH. Los resultados revelan un rendimiento diagnóstico con un elevado VPN como parámetro de mayor relevancia clínica para descartar lesiones pulmonares y óseas con un alto grado de certeza, con patrones de gran interés que varían según la alteración que se investiga y el sistema evaluado. Al comparar la IA con observadores humanos de distinta experiencia (médicos de urgencias, radiólogos en formación y radiólogos expertos), emergen implicaciones significativas para la práctica clínica, que perfilan el rol de la IA como una herramienta de apoyo[15,16].

Tabla 4. Validez de distintos observadores en la detección de anomalías en radiografías óseas solicitadas en un servicio de urgencias (n = 261)

	VP (n)	FN (n)	FP (n)	VN (n)	Sensibilidad (%)	p1	Especificidad (%)	p2	VPP (%)	p3	VPN (%)	p4
Anomalías (N = 58)												
Inteligencia artificial (BoneView)	49	9	31	172	84,5		84,7		61,3		95,0	
Informe de urgencias	41	17	13	190	70,7	0,075	93,6	**0,004**	75,9	0,076	91,8	0,203
Facultativo urgencias	45	13	15	188	77,6	0,344	92,6	**0,012**	75,0	0,087	93,5	0,531
Radiólogo en formación	50	8	15	188	86,2	0,793	92,6	**0,012**	76,9	**0,044**	95,9	0,677
Radiólogo experto	47	11	18	185	81,0	0,623	91,1	**0,048**	72,3	0,162	94,4	0,782
Fractura (N = 48)												
Inteligencia artificial (BoneView)	42	6	26	187	87,5		87,8		61,8		96,9	
Informe de urgencias	34	14	12	201	70,8	**0,044**	94,4	**0,017**	73,9	0,177	93,5	0,112
Facultativo urgencias	42	6	13	200	87,5	1,000	93,9	**0,029**	76,4	0,084	97,1	0,909
Radiólogo en formación	43	5	13	200	89,6	0,749	93,9	**0,029**	76,8	0,073	97,6	0,684
Radiólogo experto	37	11	9	204	77,1	0,181	95,8	0,003	80,4	**0,034**	95,8	0,311
Luxación (N = 5)												
Inteligencia artificial (BoneView)	3	2	2	254	60,0		99,2		60,0		99,2	
Informe de urgencias	4	1	3	253	80,0	0,490	98,8	0,653	57,1	0,921	99,6	0,567
Facultativo urgencias	3	2	3	253	60,0	1,000	98,8	0,653	50,0	0,740	99,2	0,997
Radiólogo en formación	5	0	11	245	100	0,853	95,7	**0,012**	31,3	0,248	100	0,972
Radiólogo experto	5	0	6	250	100	0,853	97,7	0,154	45,5	0,590	100	0,986
Derrame (N = 12)												
Inteligencia artificial (BoneView)	3	9	6	243	25,0		97,6		33,3		96,4	
Informe de urgencias	0	12	1	248	0,0	0,527	99,6	0,057	0,0	0,712	95,4	0,552
Facultativo urgencias	0	12	2	247	0,0	0,527	99,2	0,154	0,0	0,480	95,4	0,546
Radiólogo en formación	7	5	16	233	58,3	0,098	93,6	**0,029**	30,4	0,874	97,9	0,329
Radiólogo experto	12	0	14	235	100	**0,003**	94,4	0,068	46,2	0,503	100	**0,033**

VP: verdadero positivo; FN: falso negativo; FP: falso positivo; VN: verdadero negativo.
Sensibilidad = VP/VP + FN. Especificidad = VN/VN + FP. VPP: VP/VP + FP. VPN: VN/VN + FN.
p1: diferencia entre la sensibilidad de los distintos observadores y la inteligencia artificial (BoneView). p2: diferencia entre la especificidad de los distintos observadores y la inteligencia artificial (BoneView). p3: diferencia entre el VPP de los distintos observadores y la inteligencia artificial (BoneView). p4: diferencia entre el VPN de los distintos observadores y la inteligencia artificial (BoneView).
Los valores en negrita denotan significación estadística ($p < 0,05$).

El rendimiento global de ChestView para la detección de cualquier anomalía torácica lo sitúa en una posición óptima y estratégicamente valiosa. Los datos muestran que el sistema de IA supera de forma estadísticamente significativa la sensibilidad de los urgenciólogos, que a menudo son los primeros en interpretar estas imágenes en un entorno de alta presión. Este hallazgo es consistente con la literatura, que ha demostrado que la IA puede elevar el rendimiento de los profesionales no radiólogos al nivel de los especialistas en formación en radiología, actuando como un soporte crucial en entornos sin cobertura radiológica 24/7[17,18]. Aunque la sensibilidad de ChestView no alcanza a la de los radiólogos, su capacidad para mejorar el diagnóstico inicial en la primera línea asistencial es una de sus mayores fortalezas. Con un VPN global del 83,0% para cualquier anomalía, el sistema proporciona una confianza considerablemente mayor que la interpretación inicial realizada en urgencias[4].

ChestView demuestra una alta sensibilidad para la detección de nódulos/masas pulmonares. Este es un hallazgo de gran relevancia clínica, dado que la omisión de nódulos pulmonares es una de las principales causas de demandas judiciales en los servicios de urgencias y radiodiagnóstico[19]. Estudios prospectivos han confirmado que la IA mejora la detección de nódulos, de manera que funciona como un sistema de seguridad eficaz[15]. El resultado del presente estudio refuerza la idea de que la IA puede ser particularmente útil para detectar lesiones focales y sutiles que pueden ser pasadas por alto por el ojo humano, especialmente en lectores menos experimentados[20].

Para el neumotórax y los nódulos/masas pulmonares, ChestView alcanza un VPN extraordinariamente alto. Clínicamente, esto significa que un resultado negativo de la IA para estas dos condiciones es extremadamente fiable como sospecha diagnóstica final. En un SUH, donde un neumotórax a tensión no diagnosticado puede ser fatal y un nódulo pulmonar omitido puede retrasar un diagnóstico oncológico, esta capacidad de descarte *(rule-out)* es de una importancia crucial. Permite al médico de urgencias redirigir su atención hacia otras posibles causas con mayor confianza y optimiza el proceso diagnóstico en tiempo real. Para el neumotórax, ChestView alcanzó una sensibilidad del 75%. La literatura muestra un rendimiento excelente de la IA para esta enfermedad, con un aumento absoluto de la sensibilidad de hasta el 26% cuando se utiliza como asistente[10]. Metanálisis recientes sitúan la sensibilidad conjunta de los algoritmos de IA para neumotórax en un 87% y otros estudios publican sensibilidades superiores al 94%[12]. Si bien el resultado en este estudio es ligeramente inferior, sigue representando una mejora sustancial sobre la práctica de urgencias y subraya el

potencial de la IA para priorizar estos casos críticos[21]. En contraste, el rendimiento del sistema fue notablemente inferior para la detección de consolidaciones y masas hiliomediastínicas. Esta variabilidad de rendimiento según la lesiones o enfermedad es un fenómeno descrito. La dificultad para detectar consolidaciones, que a menudo presentan bordes mal definidos y se superponen con otras estructuras, es un desafío conocido para los algoritmos. Aunque otros estudios han comunicado un rendimiento superior de la IA para la neumonía[22], los datos actuales sugieren que, para esta alteración específica, el sistema evaluado no supera al juicio del radiólogo y debe ser utilizado con cautela.

En el ámbito osteoarticular, el hallazgo más destacado es su excepcional rendimiento en la detección específica de fracturas. Este resultado está fuertemente respaldado por una amplia evidencia. Múltiples estudios prospectivos y metanálisis han confirmado que los sistemas de IA para fracturas, y BoneView en particular, alcanzan una sensibilidad independiente en torno al 87-92%[23]. Más importante aún, se ha demostrado que la asistencia de la IA eleva la sensibilidad de los radiólogos en formación a más del 91%, sin una pérdida significativa de especificidad. Por ello, actúa como un potente sistema de apoyo que reduce los errores de omisión[3]. Este estudio corrobora que la IA puede detectar fracturas que incluso un experto pasa por alto, con lo que se posiciona como una herramienta necesaria para mejorar la seguridad del paciente.

Pero su verdadero poder reside en su elevado VPN. Para la detección específica de fracturas, este valor asciende a un 96,9%. Este dato es, posiblemente, el hallazgo más destacado del estudio. Un VPN tan elevado significa que si BoneView no detecta una fractura, la probabilidad de que realmente no exista es altísima. En la práctica habitual, una sospecha de fractura no resuelta a menudo conduce a inmovilizaciones innecesarias, consultas de seguimiento, pruebas de imágenes adicionales o litigios por reclamaciones patrimoniales o judiciales. Este hallazgo podría reducir drásticamente estas disfunciones, y agilizan las altas desde urgencias con cierta seguridad y optimizar el uso de recursos. La evidencia actual apoya este potencial, con múltiples estudios prospectivos, que confirman la alta sensibilidad y el elevado VPN de BoneView, y demuestran que su integración reduce la tasa de errores diagnósticos sin comprometer la especificidad del radiólogo[24,25].

Además, el rendimiento de BoneView para hallazgos no relacionados con fracturas fue de menor impacto, con una sensibilidad de solo el 60,0% para luxaciones y del 25,0% para derrames articulares. Esto subraya una limitación fundamental de las soluciones de IA actuales: son herramientas altamente especializadas, entrenadas para una tarea específica (en este caso, la detección de fracturas)[11]. Y no pueden ni deben considerarse un sustituto de la evaluación clínica y radiológica integral, que incluye la valoración de partes blandas, alineación y otros hallazgos secundarios.

Entre las limitaciones del estudio cabe mencionar que, en primer lugar, que al ser un análisis de una serie retrospectiva, no se pueden extraer conclusiones sobre el impacto real en el flujo de trabajo o los tiempos de informe. Sin embargo, su diseño basado en una muestra aleatoria de casos de un SUH real garantiza una alta validez externa, así como la inclusión de un espectro completo de observadores y un amplio abanico de enfermedades y permite una evaluación específica y de práctica habitual del rendimiento de la IA. En segundo lugar, al ser un estudio unicéntrico, la generalización de los resultados debe ser cautelosa, ya que el rendimiento de la IA puede variar en función de las características de la población, la idiosincrasia del SUH y los equipos de adquisición[13]. A su vez, hay que tener en cuenta que para las anomalías de baja frecuencia (neumotórax, masa hiliomediastínica y nódulo/masa pulmonar o luxaciones/derrame articular) los resultados obtenidos en S, E, VPP y VPN van a tener una gran variabilidad, de ahí que sea difícil obtener diferencias estadísticamente significativas. Aunque el número limitado de casos positivos para enfermedades de baja frecuencia como el neumotórax (4 casos) o los nódulos pulmonares (10 casos) afecta a la fiabilidad de la sensibilidad, el valor clínico de la IA reside en su elevado VPN, que se ha evidenciado en este estudio, con lo que el sistema actúa como una red de seguridad, permitiendo a los médicos de urgencias descartar con gran certeza posibles hallazgos graves y redirigir la atención de forma segura y eficiente. En tercer lugar, la existencia del sesgo del estándar de referencia establecido mediante consenso de radiólogos en lugar de otras pruebas radiológicas. Aunque la tomografía computarizada puede considerarse el patrón oro para determinadas enfermedades tales como nódulos pulmonares pequeños o fracturas sutiles, su uso sistemático en todos los pacientes no sería ético ni factible por la radiación y el coste asociado. El consenso radiológico refleja mejor la práctica clínica habitual en un SUH y permite una validación externa más realista de las herramientas de IA[12,26]. Finalmente, la IA solo detecta las enfermedades para las que fue entrenada, lo que supone un riesgo inherente de omitir hallazgos no previstos en su algoritmo[13,14].

En conclusión, este estudio de validación externa demuestra que las herramientas de IA ChestView y BoneView tienen un elevado VPN, y son instrumentos válidos para descartar fracturas y lesiones pulmonares en los SUH. Deben considerarse como un lector asistente sin sustituir al radiólogo, para aumentar significativamente la sensibilidad diagnóstica, reducir errores críticos en la fase inicial de atención al paciente urgente y evitar consultas frecuentes y potencialmente innecesarias.

Conflicto de intereses: Los autores declaran no tener conflicto de interés en relación con el presente artículo.

Financiación: Los autores declaran la existencia de financiación en relación con el presente artículo.

Responsabilidades éticas: Todos los autores han confirmado el mantenimiento de la confidencialidad y respeto de los derechos de los pacientes en el documento de responsabilidades del autor, acuerdo de publicación y cesión de derechos a EMERGENCIAS. El protocolo del estudio fue aprobado por el Comité de Ética e Investigación Clínica con medicamentos (CEIm) del Hospital Universitario General Universitario Dr. Balmis de Alicante (número de referencia: 2024-0042) con exención del consentimiento informado.

Artículo no encargado por el Comité Editorial y con revisión externa por pares.

Bibliografía

1 Poyiadji N, Beauchamp N, Myers DT, Krupp S, Griffith B. Diagnostic Imaging Utilization in the Emergency Department: Recent Trends in Volume and Radiology Work Relative Value Units. J Am Coll Radiol. 2023;20:1207-14.

2 Saket DD. The provision of emergency radiology services and potential radiologist workforce crisis: is there a role for the emergency-dedicated radiologist? Semin Ultrasound CT MR. 2007;28:81-4.

3 Petinaux B, Bhat R, Boniface K, Aristizabal J. Accuracy of radiographic readings in the emergency department. Am J Emerg Med. 2011;29:18-25.

4 Pinto A, Reginelli A, Pinto F, Lo Re G, Midiri F, Muzj C, et al. Errors in imaging patients in the emergency setting. Br J Radiol. 2016;89:20150914.

5 Kutbi M. Artificial Intelligence-Based Applications for Bone Fracture Detection Using Medical Images: A Systematic Review. Diagnostics (Basel). 2024;14:1879.

6 Sáenz-Abad D, Sachi Martínez-Mihara M, Lahoza-Pérez MC. La inteligencia artificial como herramienta de apoyo diagnóstico en urgencias. Emergencias. 2025;37:78-9.

7 González-Martínez F, Garrido NJ, Mateo J. Inteligencia artificial en la práctica clínica de urgencias: más realidad que fascinación. Emergencias. 2025;37:159-8.

8 Gordo-Vidal F, Gordo-Herrera N. Inteligencia artificial y sistemas de aprendizaje automático: fascinación versus realidad. Emergencias. 2025;37:03-4.

9 Hwang EJ, Nam JG, Lim WH, Park SJ, Jeong YS, Kang JH, et al. Deep Learning for Chest Radiograph Diagnosis in the Emergency Department. Radiology. 2019;293:573-80.

10 Bennani S, Regnard NE, Ventre J, Lassalle L, Nguyen T, Ducarouge A, et al. Using AI to Improve Radiologist Performance in Detection of Abnormalities on Chest Radiographs. Radiology. 2023;309:e230860. Erratum in: Radiology. 2024;311:e249015.

11 Oppenheimer J, Lüken S, Hamm B, Niehues SM. A Prospective Approach to Integration of AI Fracture Detection Software in Radiographs into Clinical Workflow. Life (Basel). 2023;13:223.

12 Katzman BD, Alabousi M, Islam N, Zha N, Patlas MN. Deep Learning for Pneumothorax Detection on Chest Radiograph: A Diagnostic Test Accuracy Systematic Review and Meta Analysis. Can Assoc Radiol J. 2024;75:525-33.

13 Kim C, Yang Z, Park SH, Hwang SH, Oh YW, Kang EY, et al. Multicentre external validation of a commercial artificial intelligence software to analyse chest radiographs in health screening environments with low disease prevalence. Eur Radiol. 2023;33:3501-9.

14 Omoumi P, Ducarouge A, Tournier A, Harvey H, Kahn CE Jr, Louvet-de Verchère F, et al. To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). Eur Radiol. 2021;31:3786-96.

15 Cellina M, Cè M, Irmici G, Ascenti V, Caloro E, Bianchi L, et al. Artificial Intelligence in Emergency Radiology: Where Are We Going? Diagnostics (Basel). 2022;12:3223.

16 Estella A, Armengol de la Hoz MA, González del Castillo J, Grupo de trabajo INFURG-SEMES. Datos abiertos e inteligencia artificial: una ventana de oportunidad para pacientes sépticos en los servicios de urgencias. Emergencias. 2025;37:373-81.

17 Rudolph J, Huemmer C, Ghesu FC, Mansoor A, Preuhs A, Fieselmann A, et al. Artificial Intelligence in Chest Radiography Reporting Accuracy: Added Clinical Value in the Emergency Unit Setting Without 24/7 Radiology Coverage. Invest Radiol. 2022;57:90-8.

18 Rudolph J, Huemmer C, Preuhs A, Buizza G, Hoppe BF, Dinkel J, et al.Nonradiology Health Care Professionals Significantly Benefit From AI Assistance in Emergency-Related Chest Radiography Interpretation. Chest. 2024;166:157-70.

19 Nam JG, Hwang EJ, Kim J, Park N, Lee EH, Kim HJ, et al. AI Improves Nodule Detection on Chest Radiographs in a Health Screening Population: A Randomized Controlled Trial. Radiology. 2023;307:e221894.

20 Woodhouse P, Paez R, Meyers P, Lentz RJ, Shojaee S, Sharp K, et al. Leveraging Artificial Intelligence as a Safety Net for Incidentally Identified Lung Nodules at a Tertiary Center. J Am Coll Surg. 2025;240:417-22.

21 Hillis JM, Bizzo BC, Mercaldo S, Chin JK, Newbury-Chaet I, Digumarthy SR, et al. Evaluation of an Artificial Intelligence Model for Detection of Pneumothorax and Tension Pneumothorax in Chest Radiographs. JAMA Netw Open. 2022;5:e2247172.

22 Ippolito D, Maino C, Gandola D, Franco PN, Miron R, Barbu V, et al. Artificial Intelligence Applied to Chest X-ray: A Reliable Tool to Assess the Differential Diagnosis of Lung Pneumonia in the Emergency Department. Diseases. 2023;11:171.

23 Brady AP. Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging. 2017;8:171-82.

24 Castro-Delgado R, Pardo Ríos M. La inteligencia artificial y los servicios de urgencias y emergencias: debemos dar un paso adelante. Emergencias. 2024;36:145-7.

25 Romero Olóriz C. Inteligencia artificial en incidentes con múltiples víctimas: estado actual y perspectivas. Emergencias. 2025;37:159-60.

26 Husarek J, Hess S, Razaeian S, Ruder TD, Sehmisch S, Müller M, et al. Artificial intelligence in commercial fracture detection products: a systematic review and meta-analysis of diagnostic test accuracy. Sci Rep. 2024;14:23053.