

VALIDACIÓN EXTERNA DE UN SOPORTE DE INTELIGENCIA ARTIFICIAL PARA EL DIAGNÓSTICO DE FRACTURAS EN UN SERVICIO DE URGENCIAS

Francisco Javier Sanz Carrio, Àngels Moreno Gutiérrez
Julia Castaños Cortina, Jaime López Prieto, Carlos
Pacios Cerecedo, Diego Hernández Bautista, Elia Pérez
Fernández, José Martel Villagrán.

Hospital Universitario Fundación Alcorcón, Alcorcón,
Madrid.

INTRODUCCIÓN

Las fracturas tienen una elevada incidencia en los servicios de urgencias. Debido a ello, su detección clínica y radiológica es uno de los procedimientos más comunes (1-2). Además, el crecimiento exponencial del uso de pruebas de imagen conlleva una curva de aprendizaje en la interpretación de las mismas. Se calcula que los errores o los retrasos diagnósticos suponen del 3 al 10% de los diagnósticos de fractura, superando el 24% del total de errores diagnósticos en los servicios de urgencias (1-2). Existen estudios (3-4) que demuestran que la experiencia del clínico es un factor determinante en la incidencia de discrepancias diagnósticas, ya que en la organización de los servicios de urgencias es generalmente un facultativo residente quien realiza la primera atención del paciente, sobre todo en determinadas franjas horarias.

A esta realidad se suma el incremento de estudios de imagen musculoesquelética, dado el papel central que ésta tiene tanto en el diagnóstico como en el manejo de las diferentes patologías de su ámbito, aumentando considerablemente la carga de trabajo en los servicios de radiología. Asimismo, de forma paralela, se ha observado un incremento en la demanda de recursos tecnológicos para optimizar los procesos diagnósticos y terapéuticos.

INTRODUCCIÓN

El término Inteligencia Artificial hace referencia a la capacidad de los sistemas informáticos de mimetizar la inteligencia humana en cuanto al procesamiento e integración de datos. Su aplicación en medicina y, particularmente, en la imagen diagnóstica, está experimentando una expansión constante (5). Dentro de las ramas de la Inteligencia Artificial, el aprendizaje automático (en inglés “machine learning”), es capaz de crear algoritmos diagnósticos a partir de datos observacionales, de tal manera que es posible la creación de capas de procesamiento, y posteriormente una extracción progresiva de datos más complejos. Asimismo, el aprendizaje en profundidad (en inglés “deep learning”) con redes neurales convolucionales (en inglés “convolutional neural networks”) permiten la resolución de problemas basados en reconocimiento de patrones, con un rendimiento óptimo en diversos aspectos clave de la imagen médica.

Las herramientas de Inteligencia Artificial desarrollados en los últimos años, aunque aún han de enfrentarse a retos en relación a su implementación en la práctica habitual, poseen el potencial de reducir el riesgo clínico o diagnóstico de determinadas patologías a través de un aumento de la capacidad en la detección de los mismos, de modo que pueden introducir mejoras en el flujo de trabajo y en la productividad (5-7).

OBJETIVOS

El propósito de nuestro trabajo ha sido realizar una validación externa de una herramienta de Inteligencia Artificial en el diagnóstico de fracturas en estudios de radiografía simple (Rx) realizados en el servicio de urgencias de nuestro centro.

MATERIAL Y MÉTODOS

Diseño

Se ha realizado un estudio transversal observacional en el Hospital Universitario Fundación Alcorcón donde dos radiólogos han revisado de forma retrospectiva las radiografías del ámbito de musculoesquelético solicitadas desde el servicio de Urgencias, ya existiese un antecedente traumático o no, y que fueron valoradas inicialmente por una herramienta de Inteligencia Artificial (IA) en el período de tiempo comprendido entre los meses de junio, julio y agosto de 2023.

Patología analizada

La patología identificada en este estudio es la fractura, que consiste en una solución de continuidad de la capa cortical de un hueso. Esto puede ocurrir debido a un traumatismo, como una caída o un golpe, o debido a condiciones médicas que debilitan los huesos, como la osteoporosis o fracturas patológicas, es decir, sobre un hueso previamente afectado, por ejemplo, por un cáncer subyacente, entre otros. Las fracturas pueden variar en gravedad desde pequeñas fisuras hasta fracturas completas, con o sin conminución y con o sin desplazamiento de los fragmentos óseos. Dependiendo del tipo y la ubicación de la fractura, puede requerir tratamiento médico que va desde la inmovilización con un yeso o una férula, hasta cirugía para su estabilización.

MATERIAL Y MÉTODOS

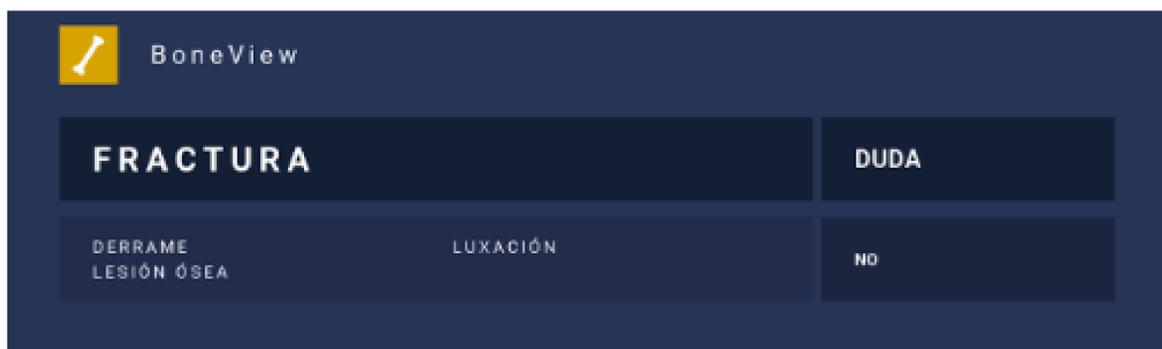
Método de análisis de imagen por la IA

Una vez realizadas las series radiográficas (incluyendo una o varias proyecciones), las imágenes se enviaron al PACS (Picture Archiving and Communication System). Se utilizó un PACS implementado con un sistema de detección asistido por computadora (CAD) basado en IA (AI-PACS). El sistema de CAD basado en IA, analiza las imágenes generando un informe automático que describe la “positividad”, “duda” o “negatividad” para todas las patologías estudiadas, como se describe en las figuras 1-6. El AI-PACS superpone los resultados del CAD basado en IA en las radiografías realizadas minutos después de su adquisición, permitiendo localizar los hallazgos patológicos. El sistema de IA indica la positividad de patología (fractura en nuestro caso) mediante un cuadro formado por una línea continua de color naranja (figuras 1 y 2), la duda con un cuadro formado por una línea discontinua de color naranja (figuras 3 y 4), y la negatividad con ausencia de marcaje (figuras 5 y 6), como se muestra:

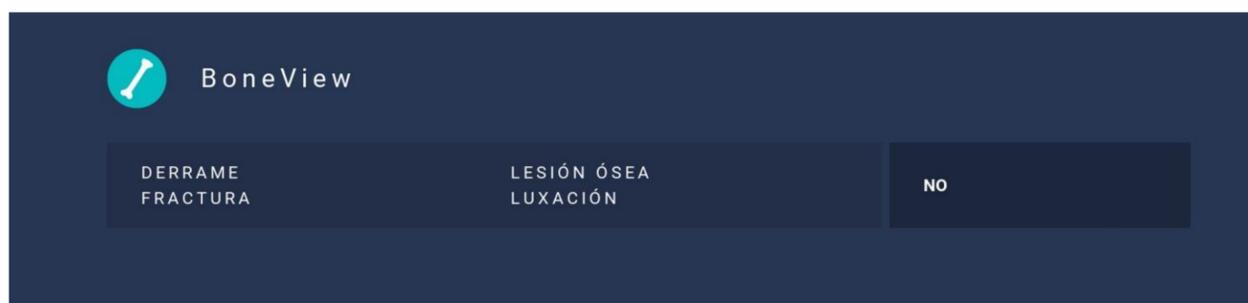


Figuras 1 y 2. Positivo para fractura.

MATERIAL Y MÉTODOS



Figuras 3 y 4. Duda para fractura.



Figuras 5 y 6. Negatividad para fractura.

MATERIAL Y MÉTODOS

En nuestro estudio se consideraron, para un primer análisis de datos, aquellas áreas de interés que el soporte de IA consideraba como “duda” para fractura, como positivas.

Los dos radiólogos participantes valoraron el conjunto de los estudios de forma retrospectiva, asignando VP (verdadero positivo), VN (verdadero negativo), FP (falso positivo) y FN (falso negativo), contrastando así el resultado de la IA.

Criterios de inclusión

Los criterios de inclusión de este estudio han sido los siguientes:

- Pacientes mayores de 18 años.
- Radiografías de musculoesquelético sin alteraciones significativas o con duda/positividad para fractura.

MATERIAL Y MÉTODOS

Los criterios de exclusión de este estudio han sido los siguientes:

- Radiografías de calidad técnica subóptima que supongan una dificultad para su valoración.
- Radiografías de columna cervical, dado que el soporte de IA utilizado no valora dicha localización.
- Radiografías para control de fracturas ya conocidas u otras radiografías de musculoesquelético solicitadas por una vía diferente a la de urgencias (consultas ambulatorias o pacientes ingresados).
- Se excluyeron todas las radiografías donde se detectaban otras patologías también valoradas por nuestra herramienta de IA, es decir, duda/positividad para lesión ósea, luxación y derrame articular.

MATERIAL Y MÉTODOS

Análisis de datos

Para responder al objetivo principal se calculan los índices de validez, sensibilidad (S) y especificidad (E), y utilidad, valor predictivo positivo (VPP) y valor predictivo negativo (VPN) de la IA, confrontado con el del radiólogo revisor. Además se evalúa la concordancia entre ambos métodos (diagnóstico de la IA vs. radiólogos) mediante el índice Kappa (Tabla 1). Para interpretar el índice de concordancia se considera:

Concordancia	Valores del índice Kappa
Muy débil	0 – 0,2
Débil	0,21 – 0,4
Moderado	0,41 – 0,6
Buena	0,61 – 0,8
Muy buena	0,81 - 1

Tabla 1. Valores del índice Kappa.

Todos los resultados se acompañan del correspondiente intervalo de confianza al 95% (IC95%).

Para valorar el impacto que los casos valorados como “duda” tienen en la utilidad de la IA, se realiza un análisis de sensibilidad adicional considerando las valoración dudosa como negativa.

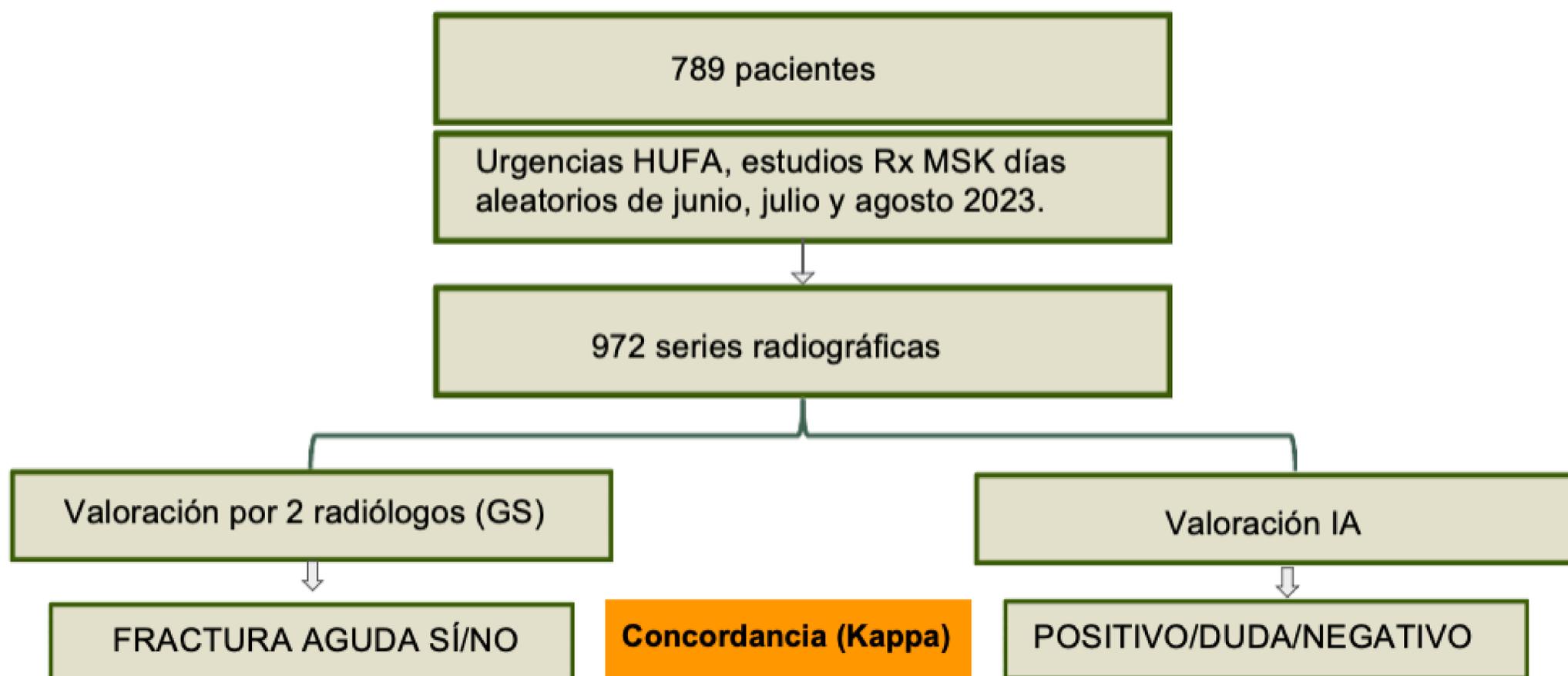


Figura 7. Esquema simplificado de la obtención de datos.

RESULTADOS

En este estudio se han analizado un total de 972 series radiográficas (un total de 1030 proyecciones) de un total de 789 pacientes, de los cuales un 54% fueron mujeres, y de una edad media de 52.6 (DE 24) (Figura 8). La proyección más frecuente fue anteroposterior y lateral (AP-L) (50%) (Figura 9).

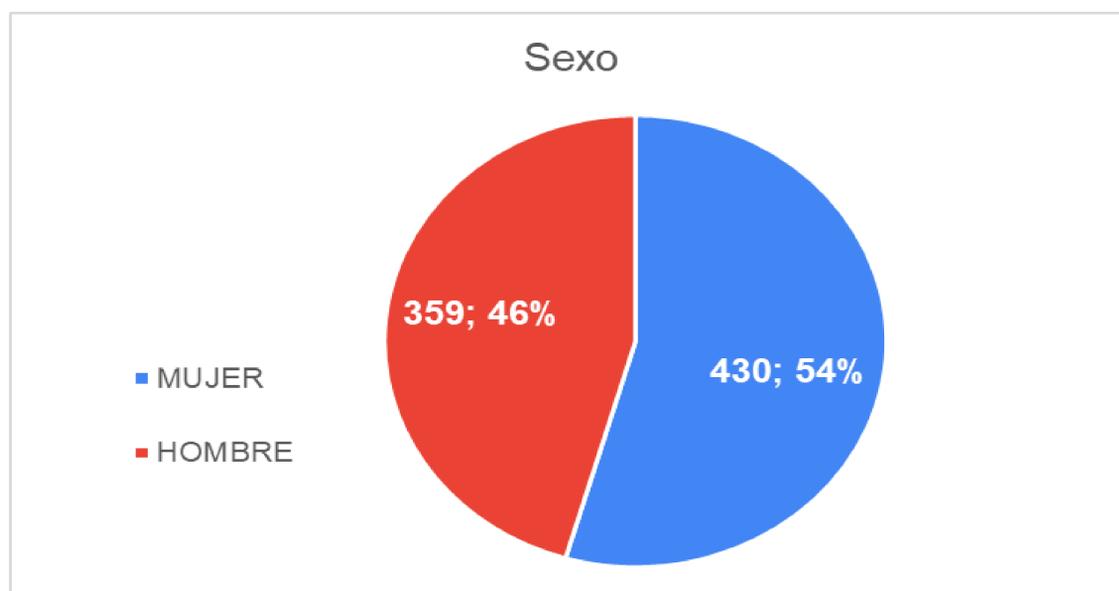


Figura 8. Diagrama de sectores del sexo de la muestra.

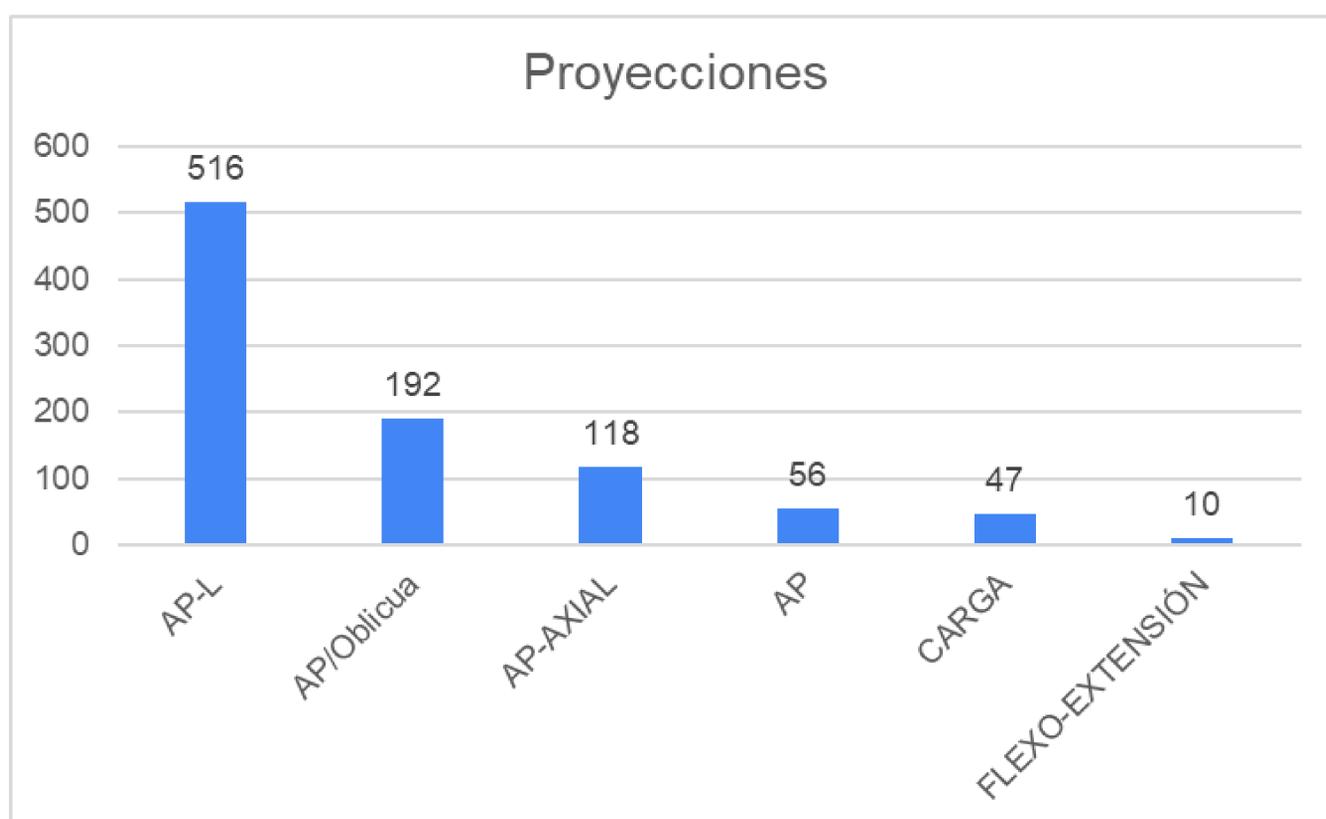


Figura 9. Diagrama de barras con distribución de las proyecciones.

RESULTADOS

La IA clasifica 203 positivos, 728 negativos y 59 dudosos, incluidos en los VP y los FP. En nuestro estudio la sensibilidad estimada fue del 96.2%, la especificidad del 95.7%, el VPP del 86% y el VPN del 98.9% y la exactitud del 95.7% (Tabla 2 y Figura 10).

	Valor	Índices de validez	Valor	IC95%	
VP	203	S	96.20%	92.70%	98%
FP	33	E	95.70%	94%	97%
FN	8	VPP	86%	80.90%	90.20%
VN	728	VPN	98.90%	97.90%	99.50%
VP+VN	931	Exactitud	95.68%		
Prevalencia	95.8%	Concordancia	0.881	0.845	0.917

Tabla 2. Tabla de contingencia.

VP: verdaderos positivos; VN: verdaderos negativos

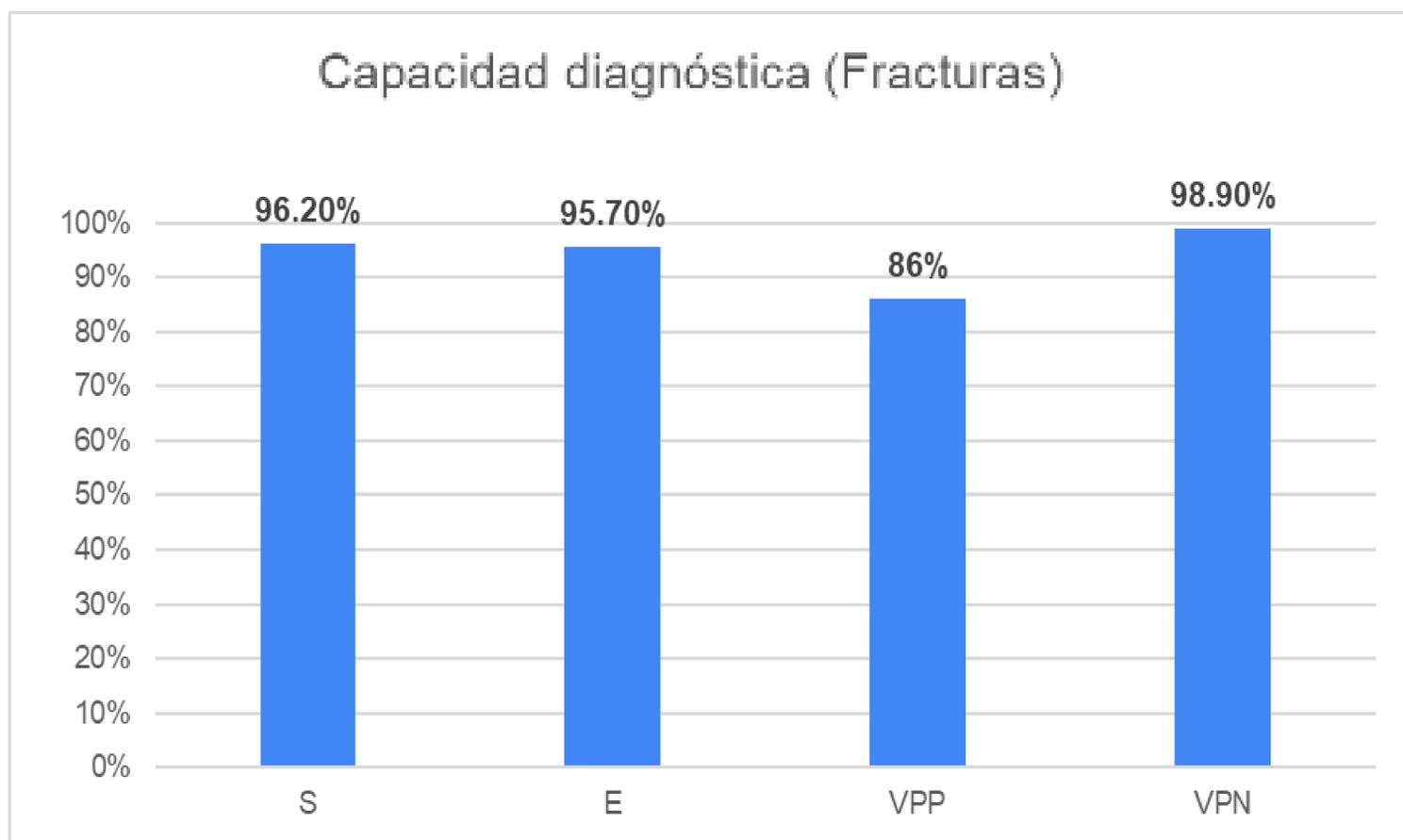


Figura 10. Diagrama de barras con porcentaje de S, E, VPP y VPN.

RESULTADOS

En cuanto a la concordancia entre ambos métodos, se obtiene un índice kappa=0.881 (IC95%:0.845-0.917). Tabla 2.

Análisis de sensibilidad

De los 59 hallazgos marcados como dudosos por la IA, 34 fueron VP (58%) y 25 FP (42%). Para valorar el peso de los casos dudosos, se realizó un análisis de sensibilidad teniendo en cuenta la reconversión de los casos dudosos según la IA a FP y VN, con lo que la sensibilidad estimada fue del 81.6%, la especificidad del 98.3%, el VPP del 93%, el VPN del 95.2% y la exactitud del 94.6% (Tabla 3 y Figura 11).

	Valor	Índices de validez	Valor	IC95%	
VP	169	S	81.60%	75.70%	87%
FP	13	E	98.30%	97%	99%
FN	38	VPP	93%	88.10%	96.10%
VN	752	VPN	95.20%	93.50%	96.60%
VP+VN	921	Exactitud	94.66%		
Prevalencia	94.8%	Concordancia	0.836	0.793	0.88

Tabla 3. Tabla de contingencia considerando los casos dudosos como negativos.

RESULTADOS

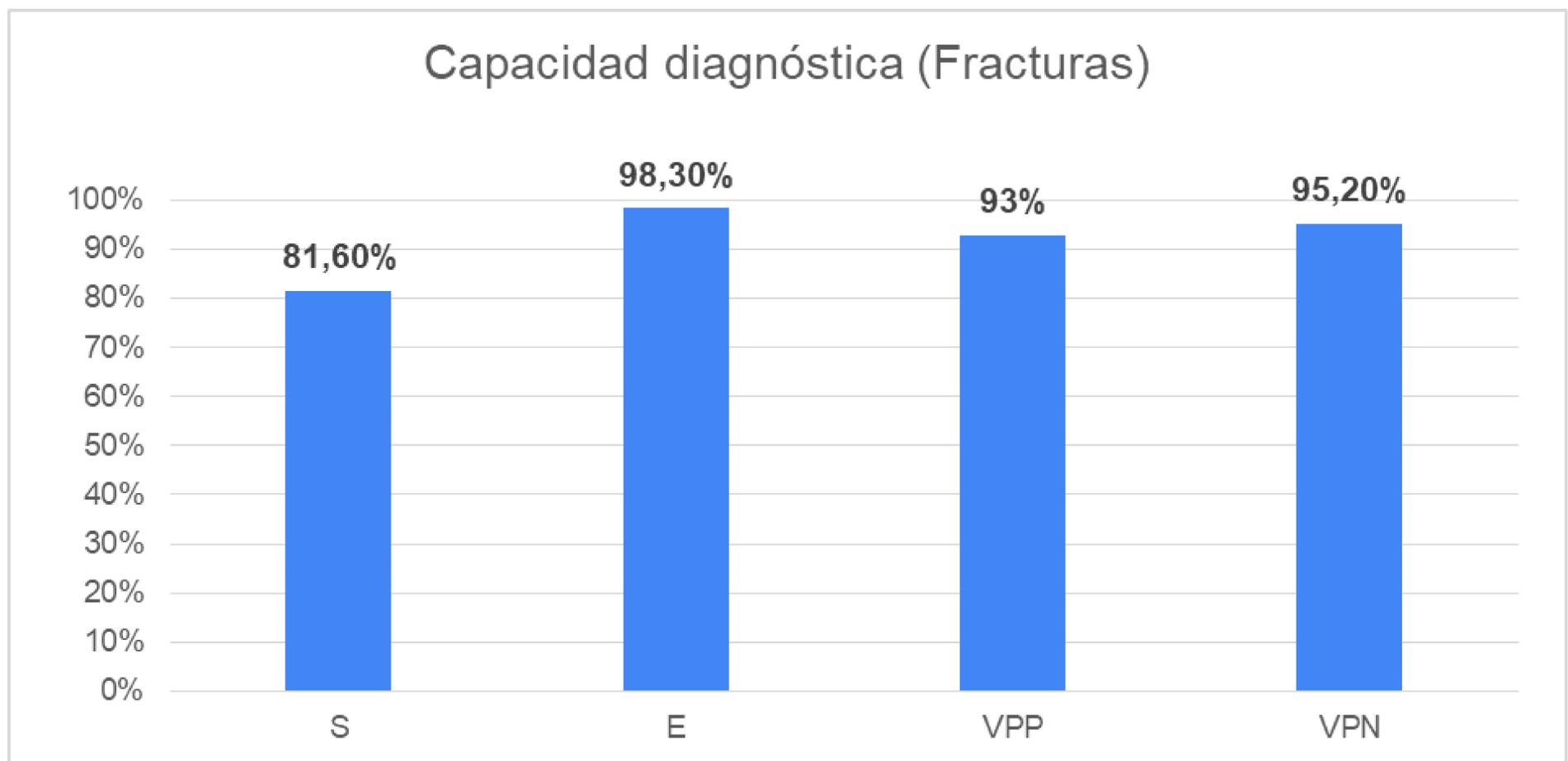


Figura 11. Diagrama de barras con porcentaje de S, E, VPP y VPN considerando los casos dudosos como negativos.

En cuanto a la concordancia entre ambos métodos, se obtiene un índice kappa=0.836 (IC95%:0.793-0.88). Tabla 3.

DISCUSIÓN

En nuestro estudio, la IA mostró una alta precisión diagnóstica, con una S del 96.2% (IC95%: 92.7-98), una E del 95.7% (IC95%: 94-97), un VPP de 86% (IC95%: 80.9-90.2), un VPN de 98.9% (IC95%: 97.9-99.5), y una exactitud del 95.6%, siendo comparables estos resultados a los obtenidos en otros estudios de diseño similar (8), sobre todo en cuanto a la S y el VPN. Se obtuvo asimismo una elevada concordancia entre el diagnóstico arrojado por parte de la IA y el del radiólogo 0.88 (IC95%:0.845-0.917). Con este estudio se ha comprobado que la implementación de sistemas de Inteligencia Artificial en el diagnóstico por imagen musculoesquelética para la detección de fracturas es un soporte eficaz para el clínico, ya que cuenta con altos valores estadísticos de validez y utilidad. Se ha considerado el VPN como el parámetro de utilidad de mayor importancia, debido a su repercusión en la disminución de los posibles errores o retrasos diagnósticos en el manejo del paciente, a tenor de las implicaciones del falso negativo.

Respecto a la clasificación de los casos catalogados por la IA como “duda”, se encontró que de los 59 casos dudosos, el 58% eran VP y el 42% FP, representando estos últimos un porcentaje no despreciable. Se decidió, por tanto, para determinar su peso estadístico, realizar un análisis de sensibilidad tomando la atribución de duda diagnóstica por la IA como un presunto resultado negativo, de manera que se reclasificaron como FP o VN, obteniendo la siguiente variación en los parámetros estadísticos: S 81.6% (IC95%:75.7-87), una E de 98.3% (IC95%:97-99), un VPP 93% (IC95%:88.1-96.1), un VPN de 95.2% (IC95%:93.5-96.6), una exactitud de 94.6%, y una concordancia de 0.83 (IC95%:0.79-0.88).

DISCUSIÓN

Aunque se han obtenido también resultados elevados, con mejora de la E, el descenso en el VPN refuerza la opción del primer análisis, a través del cual interpretamos que, aunque se trate de un caso dudoso, se alberga la posibilidad de existencia de patología en ese estudio, lo cual repercute de forma global en una disminución de los falsos negativos.

Existen estudios que también han analizado la validez y utilidad de la misma herramienta de IA de la que se dispone en nuestro centro (9-10), realizando un estudio comparativo que concluyó en una mejora en la S y E para la detección de fractura en comparación con el diagnóstico realizado por médicos residentes, así como una obtención de valores de VPP y VPN similares. En consonancia con los estudios referidos y una revisión sistemática sobre el papel de la IA en el diagnóstico de fracturas (1), surge la hipótesis sobre la aplicabilidad real de estos resultados, dado que en la organización del servicio de urgencias de nuestro centro no se lleva a cabo una revisión de los estudios de radiología simple musculoesquelética por parte de radiólogos, a no ser que se realice una interconsulta. Sin embargo, y como lo respaldan otros trabajos publicados con anterioridad (10), se sugiere que la implementación de soportes de IA en la práctica habitual de los servicios de urgencias repercute en una mejora en el rendimiento clínico, aumentando su confianza diagnóstica y consecuentemente reduciendo los tiempos de asistencia en urgencias en comparación con la etapa pre-IA. Queda abierto el campo de estudio para futuros trabajos.

DISCUSIÓN

En cuanto a las limitaciones de nuestro trabajo, se encontraron deficiencias en los métodos de estudio que podrían limitar la extrapolación de los hallazgos a la práctica clínica. Al obtener el análisis de la IA en el PACS junto con la imagen, no ha sido posible realizar un estudio de ciego sobre el radiólogo revisor, pudiendo incurrir en sesgos de clasificación.

No se han realizado análisis de subgrupos en cuanto al sexo, la edad o la localización de la fractura. No se tuvo en cuenta el grado de experiencia del clínico lector de la prueba diagnóstica.

Tampoco se han considerado diferencias de concordancia entre ambos radiólogos, y no se ha recurrido a un tercer radiólogo experto en radiología musculoesquelética como *gold standard*. Aunque no se trata de una limitación del estudio, el hecho de que el soporte de IA no realice distinción entre fractura aguda y crónica, incurriría también en un sesgo de clasificación por sobrediagnóstico, generando un incremento en el número de consultas a Diagnóstico por Imagen o Traumatología.

CONCLUSIONES

- La demanda de herramientas de IA para el soporte diagnóstico en el ámbito de la radiología musculoesquelética está actualmente en expansión, debido al incremento en la solicitud de pruebas de imagen que repercuten sobre el volumen de carga laboral. La IA puede ser una aliada a la hora de mejorar la capacidad diagnóstica del clínico, repercutiendo en una minimización de los errores o retrasos diagnósticos de fracturas en urgencias.
- Se ha obtenido una alta fiabilidad de resultados, dado su elevado VPN, lo que resulta en una elevada utilidad en el servicio de Urgencias para discriminar entre radiografías patológicas de las no patológicas, reduciendo así la tasa de fracturas no detectadas.
- La herramienta de IA utilizada en nuestro centro ha demostrado buenos resultados para descartar la existencia de fractura en un primer nivel de atención del paciente.

EJEMPLO DE CASOS:

1) Radiografías de fractura de meseta tibial lateral previas (Figuras 12 y 14) y posteriores a la evaluación de la IA (Figuras 13,15 y 16).



Figura 12. Proyección AP de rodilla.



Figura 13. Proyección AP de rodilla con marcaje positivo para fractura.



Figura 14. Proyección LA de rodilla.



Figura 15. Proyección LA de rodilla con marcaje positivo para fractura y derrame.

BoneView		
FRACTURA	sí	
DERRAME	sí	
LESIÓN ÓSEA	LUXACIÓN	NO

Figura 16. Confirmación de hallazgos visualizados por el programa de la IA.

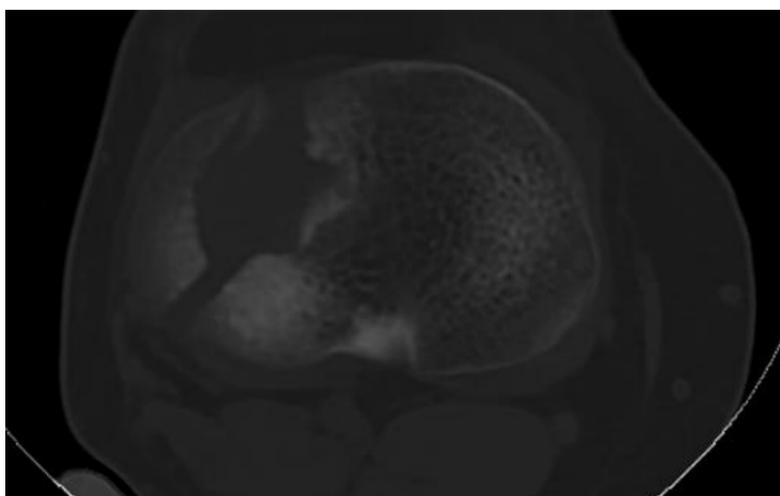


Figura 17. Corte axial TC de rodilla sin CIV.



Figura 18. Corte sagital TC de rodilla sin CIV.

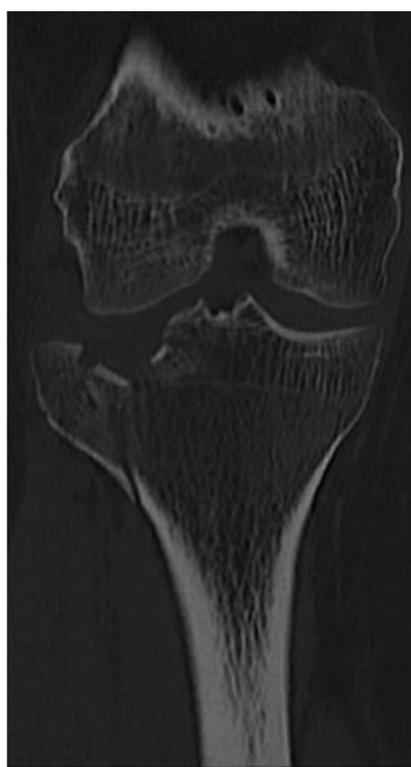


Figura 19. Corte coronal TC de rodilla sin CIV.



Figura 20. Reconstrucción 3D.

TC de rodilla derecha urgente realizado sin administración de contraste. Se adjuntan reconstrucciones multiplanares (Figuras 17,18 y19) y 3D (Figura 20).

Comentario: Fractura-hundimiento (de 12 mm) de la meseta tibial lateral derecha con trazo de predominio vertical en su vertiente posterior, sin visualizarse desplazamiento significativo (tipo II de Schatzker). No se observan otras líneas de fractura. Espinas tibiales conservadas. Hemartros, derrame articular y aumento de las partes blandas adyacentes en relación con cambios inflamatorios secundarios a su antecedente traumático.

2) Radiografías de fractura de radio distal derecho con imágenes previas (Figuras 21 y 23) y posteriores a la evaluación de la IA (Figuras 22 y 24).



Figura 21. Proyección AP de muñeca.



Figura 22. Proyección AP de muñeca con marcaje positivo para fractura y derrame.



Figura 23. Proyección LA de muñeca.

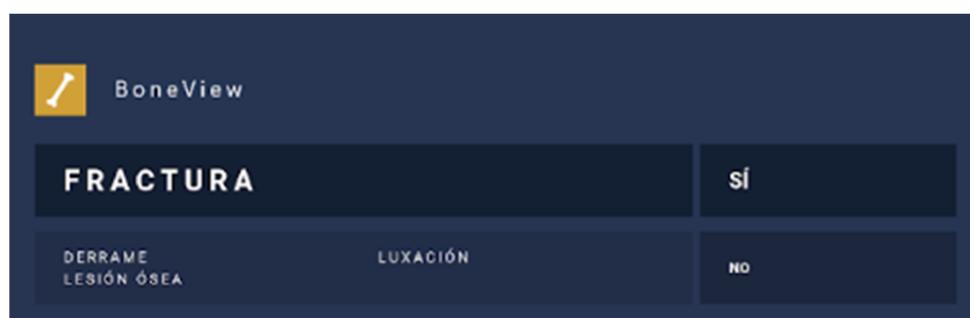


Figura 24. Confirmación de hallazgos visualizados por el program de la IA.



Figura 25. Corte sagital TC de rodilla sin CIV.



Figura 26. Corte coronal TC de rodilla sin CIV.

TC DE MUÑECA DERECHA realizado sin administración de contraste. Se adjuntan reconstrucciones multiplanares y 3D. Comentario: Fractura de radio distal con trazo de predominio horizontal en su vertiente dorsal que involucra a la superficie articular radiocarpiana, sin objetivarse un desplazamiento significativo de los fragmentos óseos. Pequeño escalón articular (0,5 mm). No se observan otras líneas de fractura. Alineación de los huesos del carpo conservada. Derrame articular y aumento de las partes blandas adyacentes en relación con cambios inflamatorios secundarios a su antecedente traumático. Yeso cerrado. Sin otros hallazgos relevantes.

3) Radiografías y TC de fractura trimaleolar de tobillo derecho. Fractura espiroidea de tercio distal de diáfisis tibial, con desplazamiento anterior del fragmento distal de 1 cm. Fractura transindesmal de peroné tipo B de Weber.



Figura 27. Proyección AP de tobillo con fractura trimaleolar.



Figura 28. Proyección lateral de tobillo.



Figura 29. Marcaje positivo de las áreas afectadas.



Figura 30. Proyección lateral de tobillo con marcaje positivo de las áreas afectadas.



Figura 31. Correlación con TC sin CIV. Plano sagital.



Figura 32. Correlación con TC sin CIV. Plano sagital.

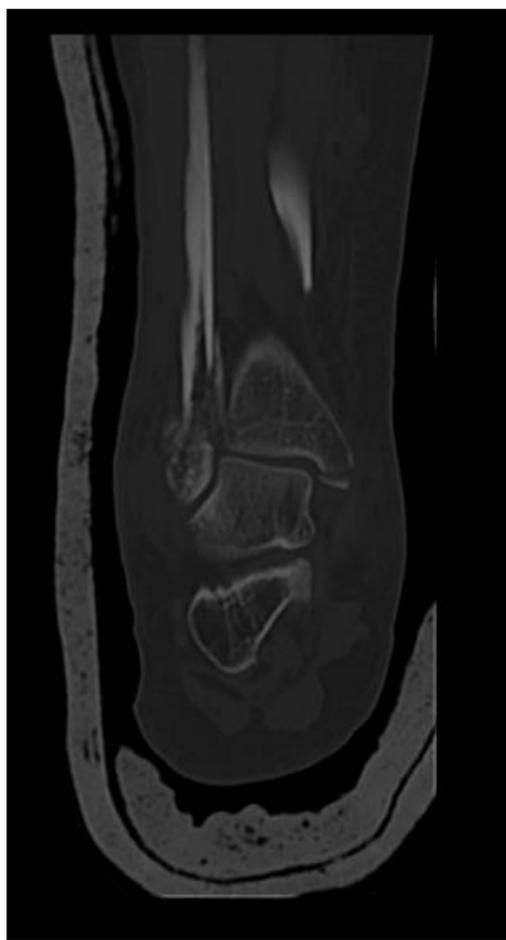


Figura 33. Correlación con TC sin CIV. Plano coronal.



Figura 34. Correlación con TC sin CIV. Plano coronal.



Figura 35. Reconstrucciones 3D. Plano coronal.



Figura 36. Reconstrucciones 3D. Plano sagital.

REFERENCIAS BIBLIOGRÁFICAS

1. Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, Stewart M, Collins GS, Furniss D. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. *Radiology*. 2022 Jul;304(1):50-62. doi: 10.1148/radiol.211785. Epub 2022 Mar 29. PMID: 35348381; PMCID: PMC9270679.
2. Guermazi A, Tannoury C, Kompel AJ, Murakami AM, Ducarouge A, Gillibert A, Li X, Tournier A, Lahoud Y, Jarraya M, Lacave E, Rahimi H, Pourchot A, Parisien RL, Merritt AC, Comeau D, Regnard NE, Hayashi D. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. *Radiology*. 2022 Mar;302(3):627-636. doi: 10.1148/radiol.210937. Epub 2021 Dec 21. PMID: 34931859.
3. Fernholm R, Pukk Härenstam K, Wachtler C, Nilsson GH, Holzmann MJ, Carlsson AC. Diagnostic errors reported in primary healthcare and emergency departments: A retrospective and descriptive cohort study of 4830 reported cases of preventable harm in Sweden. *Eur J Gen Pract*. 2019 Jul;25(3):128-135. doi: 10.1080/13814788.2019.1625886. Epub 2019 Jul 1. PMID: 31257959; PMCID: PMC6713141.
4. Wood G, Knapp KM, Rock B, Cousens C, Roobottom C, Wilson MR. Visual expertise in detecting and diagnosing skeletal fractures. *Skeletal Radiol*. 2013 Feb;42(2):165-72. doi: 10.1007/s00256-012-1503-5. Epub 2012 Sep 1. PMID: 22940835.
5. Debs P and Fayad LM (2023) The promise and limitations of artificial intelligence in musculoskeletal imaging. *Front. Radiol*. 3:1242902. doi: 10.3389/fradi.2023.1242902.

REFERENCIAS BIBLIOGRÁFICAS

6. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current applications and future impact of machine learning in radiology. *Radiology*. (2018) 288(2):318–28. doi: 10.1148/radiol.2018171820
7. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol*. 2018 May;73(5):439-445. doi: 10.1016/j.crad.2017.11.015. Epub 2017 Dec 18. PMID: 29269036.
8. Dupuis M, Delbos L, Rouquette A, Adamsbaum C, Veil R. External validation of an artificial intelligence solution for the detection of elbow fractures and joint effusions in children. *Diagn Interv Imaging*. 2024 Mar;105(3):104-109. doi: 10.1016/j.diii.2023.09.008. Epub 2023 Oct 8. PMID: 37813759.
9. 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 Jan 13;13(1):223. doi: 10.3390/life13010223. PMID: 36676172; PMCID: PMC9864518.
10. Zech JR, Santomartino SM, Yi PH. Artificial Intelligence (AI) for Fracture Diagnosis: An Overview of Current Products and Considerations for Clinical Adoption, From the *AJR* Special Series on AI Applications. *AJR Am J Roentgenol*. 2022 Dec;219(6):869-878. doi: 10.2214/AJR.22.27873. Epub 2022 Jun 22. PMID: 35731103.