

# Radiografía ósea en la urgencia: ¿Quién diagnostica mejor, la Inteligencia Artificial o el residente?

Hospital Universitario  **quirónsalud**  
Madrid

Alejandro Díaz Moreno<sup>1</sup>  
Julia López Alcolea<sup>1</sup>  
David García Castellanos<sup>1</sup>  
Raquel Cano Alonso<sup>1</sup>  
Ana Fernández Alfonso<sup>1</sup>  
Cristina Andreu Vázquez<sup>2</sup>  
Israel John Thuissard Vasallo<sup>2</sup>  
Lucía Sanabria Greciano<sup>1</sup>  
Vicente Martínez De Vega Fernández<sup>1</sup>

<sup>1</sup>Hospital Universitario QuirónSalud Madrid, Pozuelo de Alarcón (Madrid)

<sup>2</sup>Universidad Europea de Madrid. Facultad de Ciencias Biomédicas y de la Salud, Villaviciosa de Odón (Madrid).

# OBJETIVO

- Objetivos del estudio:

## **Objetivo principal:**

- Determinar la sensibilidad (S) y especificidad (E) del software de IA en el diagnóstico de fracturas, luxaciones y derrame articular en codo en radiografías simples óseas urgentes (RxU) respecto al Gold Standard (GS, radiólogo senior).

## **Objetivos secundarios:**

- Evaluar el grado de concordancia entre las lecturas realizadas por la IA y por residentes de las fracturas, luxaciones y derrame articular en codo en RxU de pacientes que acuden al servicio de Urgencias del Hospital Universitario Quirónsalud Madrid entre el 15 de octubre y el 15 de noviembre de 2022.
- Determinar la proporción de exploraciones en los que la IA y el residente registran resultados dudosos, y la proporción que fueron confirmados por el GS

# MATERIAL Y MÉTODO

- Diseñamos un estudio de tipo observacional, retrospectivo, descriptivo y doble ciego para evaluar todas las RxU de adultos de nuestro centro de octubre a noviembre de 2022.
- Los lectores no tuvieron acceso a la historia clínica, examen físico ni a la lectura aportada por la IA.
- De una muestra inicial de 2487 pacientes elegibles, obtuvimos una muestra final de 792 pacientes tras aplicar los criterios de inclusión y de exclusión.



## DISEÑO DEL ESTUDIO

### • MATERIAL Y MÉTODO

#### 2487 PACIENTES ELEGIBLES

- Referidos desde el servicio de Urgencias
- Evaluados mediante Radiografías óseas digitales
- Del 15 de octubre al 15 de noviembre de 2022

#### CRITERIOS DE INCLUSIÓN: 819

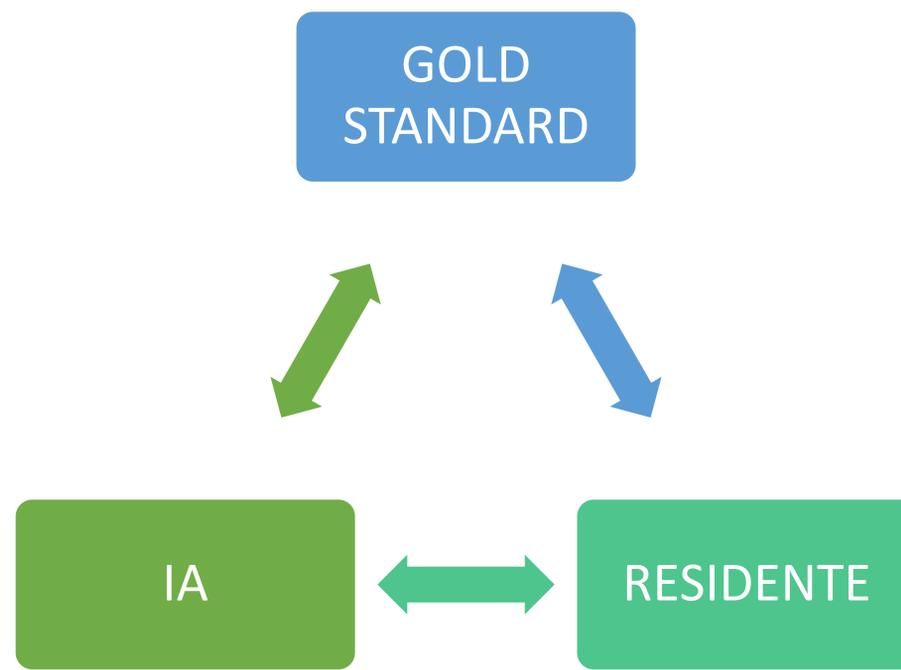
- Edad mayor o igual a 18 años
- Radiografía de miembros superiores, inferiores o de pelvis.

#### CRITERIOS DE EXCLUSIÓN: 27

- Lectura de IA no disponible (25)
- Pobre calidad de imagen (3 y 1 de estos también con lectura de IA no disponible).

MUESTRA FINAL: 792

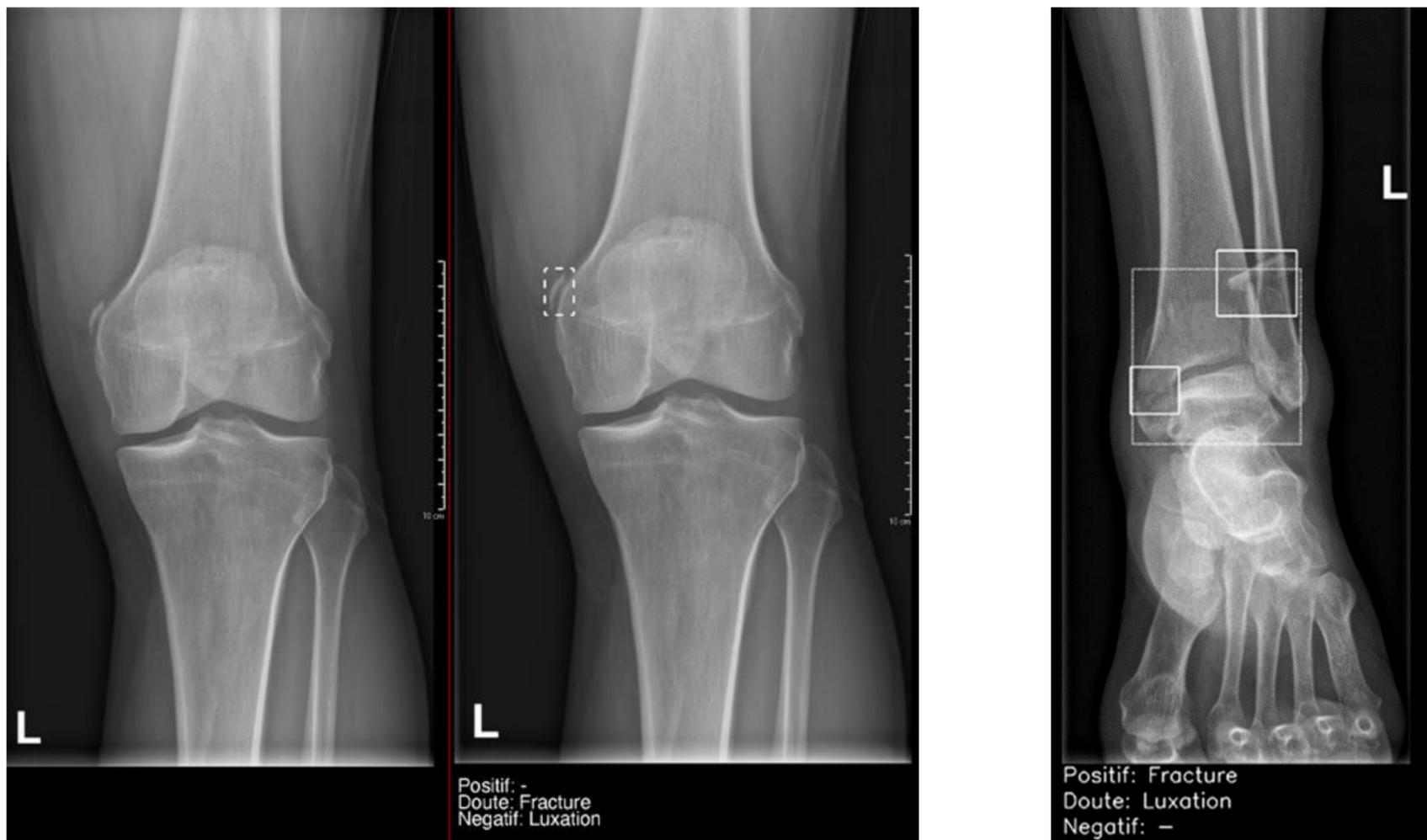
# MATERIAL Y MÉTODO



- Los lectores de las RxU fueron:
  - **Gold Standard (GS):** un radiólogo senior que leyó independientemente cada estudio y recogió el resultado como positivo o negativo en una tabla de datos. Las posibles dudas fueron resueltas en consenso junto con un segundo radiólogo senior.
  - **IA:** Cada hallazgo fue marcado mediante un cuadro (ver siguiente diapositiva) y fue registrado como positivo, negativo o dudoso. Un residente recogió estos resultados en una tabla de datos.
  - **Residente de radiodiagnóstico (RR)** (2 años de experiencia): también realiza una lectura independiente de cada estudio recogiendo cada hallazgo como positivo, negativo o dudoso en una tabla de datos.

# MATERIAL Y MÉTODO

- En el informe de la IA nos aparece el resultado:
  - **Negativo:** no aparece ninguna marca sobre la radiografía.
  - **Positivo:** el hallazgo positivo aparece marcado con una caja con línea continua sobre la radiografía.
  - **Dudoso:** el hallazgo dudoso aparece marcado con una caja con línea discontinua sobre la radiografía.

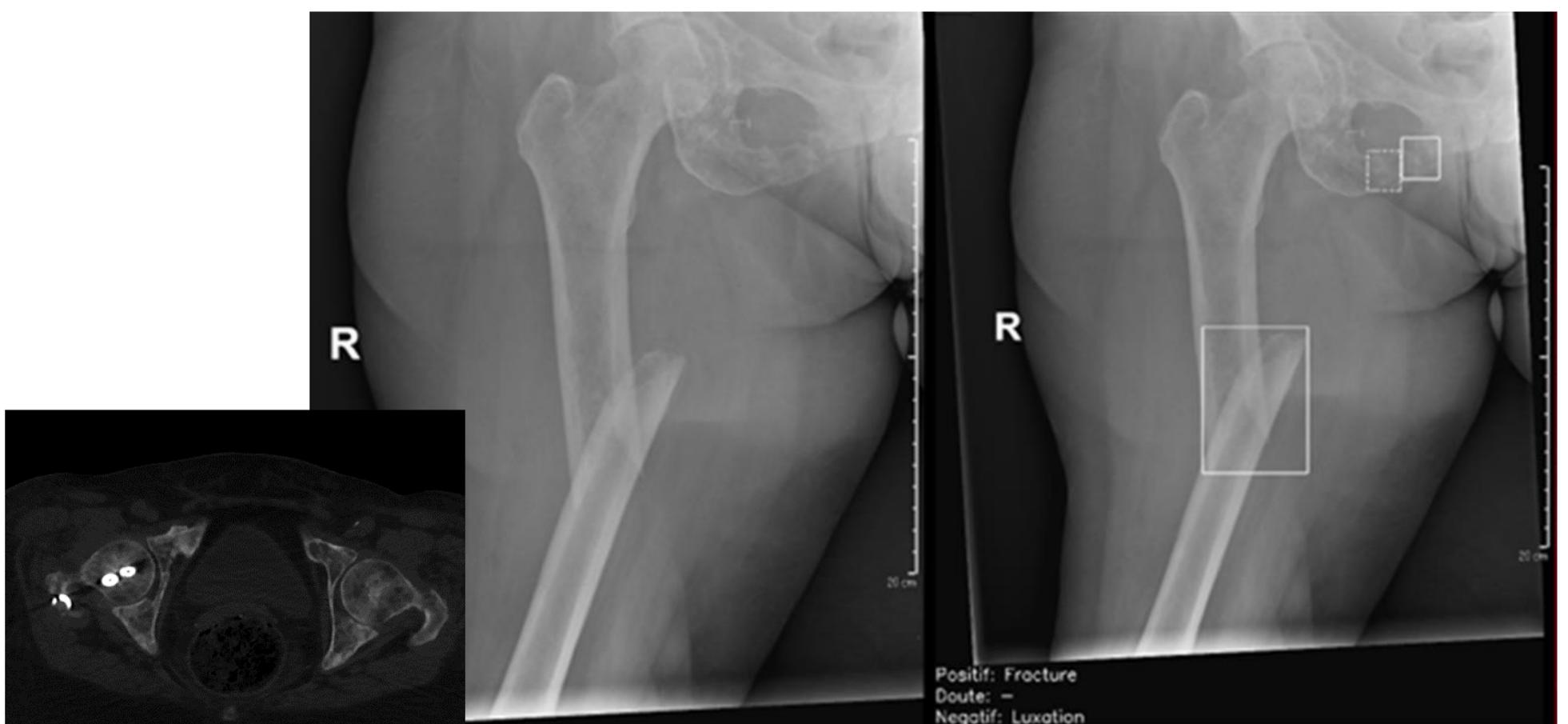


**Izquierda:** Ejemplo de fractura dudosa marcada mediante una caja con línea discontinua en la radiografía AP de rodilla, que correspondió con una lesión de Pellegrini-Stieda.

**Derecha:** Ejemplos de fracturas marcadas como positivas mediante una caja con línea continua en la radiografía AP de tobillo. También se identifica una luxación dudosa marcada por la IA.

# MATERIAL Y MÉTODO

- Al ser variables categóricas todas ellas fueron recogidas con frecuencias absolutas y relativas.
- Para validar la capacidad diagnóstica de la IA usamos tablas de contingencia y calculamos S, E e intervalos de confianza del 95%.
- Para evaluar la concordancia interobservador entre IA y RR usamos el coeficiente de Kappa.



Fractura patológica del fémur derecho por metástasis blástica.

# RESULTADOS

## 1. CARACTERÍSTICAS DE LA MUESTRA

- La mediana de edad de la muestra de nuestro estudio fue de 48 años, siendo un 48,6% varones. La mayoría de los estudios se componían de 2 proyecciones (90,2%) y la calidad de imagen fue considerada óptima en la mayoría de los casos (97,2%). (**Tabla 1**)

	<b>Casos (n=792)</b>
<b>Edad del paciente (años cumplidos) mediana [Q1; Q3]</b>	48,0 [33,0 ; 61,5]
<b>Sexo del paciente (n,%)</b>	
Hombre	385 (48,6)
Mujer	407 (51,4)
<b>Proyecciones radiográficas (n,%)</b>	
1	62 (7,8)
2	714 (90,2)
3	12 (1,5)
4	4 (0,5)
<b>Calidad de la radiografía (n,%)</b>	
Óptima	774 (97,8)
Regular	18 (2,2)
<b>Desglose por grupos de articulaciones (n,%)</b>	
Pequeñas articulaciones	407 (51,4)
Grandes articulaciones	343 (43,3)
Huesos planos/largos	42 (5,3)
<b>Desglose por articulaciones (n,%)</b>	
Rodilla	151 (19,1)
Tobillo	127 (16)
Hombro	99 (12,5)
Pie	77 (9,7)
Muñeca	68 (8,6)
Pelvis-Cadera	65 (8,2)
Mano	53 (6,7)
Dedos mano	47 (5,9)
Dedos pie	31 (3,9)
Codo	28 (3,5)
Tibia	11 (1,4)
Antebrazo	10 (1,3)
Húmero	9 (1,1)
Clavícula	6 (0,8)
Fémur	6 (0,8)
Calcáneo	4 (0,5)

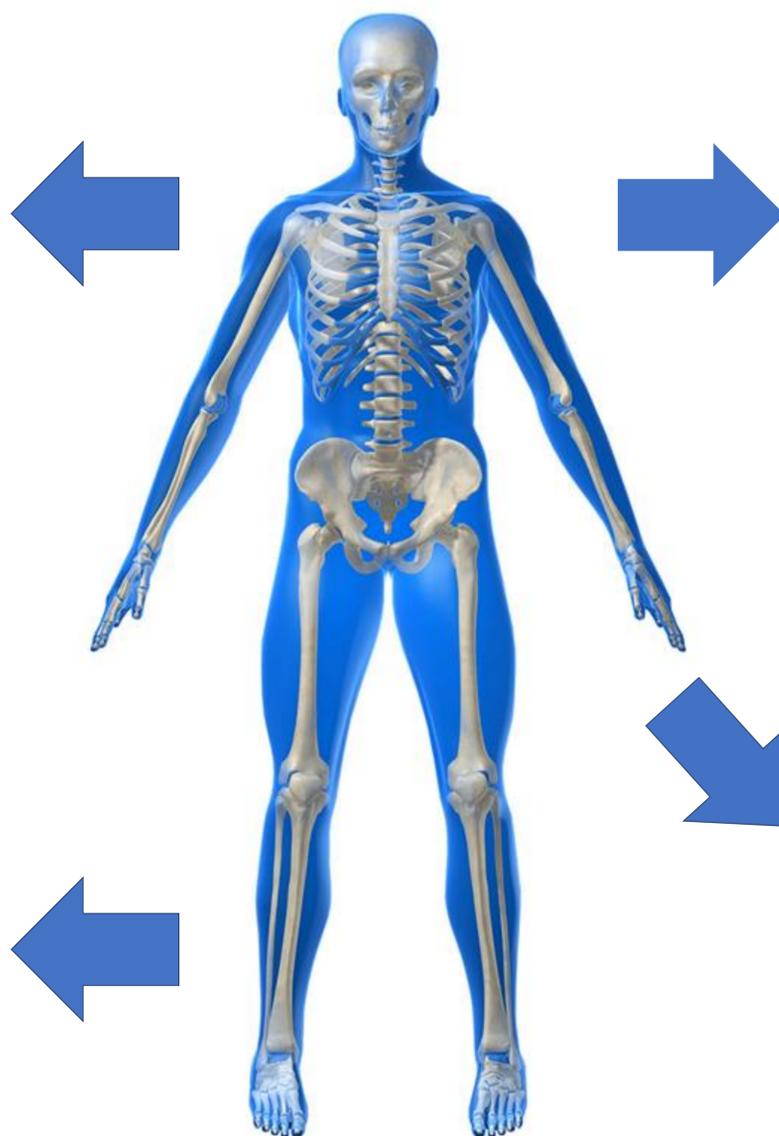
Q1: primer cuartil = percentil 25. Q3: tercer cuartil = percentil 75

**Tabla 1:** características de la muestra.

# RESULTADOS

## 1. CARACTERÍSTICAS DE LA MUESTRA

- Para un análisis más interesante dividimos los exámenes en tres grupos:
  1. Pequeñas articulaciones (51,4%): mano, pie, dedos, calcáneo, muñeca y tobillo.
  2. Grandes articulaciones (43,3%): hombro, codo, pelvis o cadera y rodilla.
  3. Huesos planos/huesos largos (5,3%): clavícula, húmero, antebrazo, fémur y tibia.



# RESULTADOS

## 2. PREVALENCIA

- De la lectura del GS inferimos una estimación de la prevalencia global y por subgrupos (**Tabla 2**). Hallamos 134 fracturas (prevalencia del 16,9% del total), la mayoría en el grupo de las pequeñas articulaciones (84 casos, 20.6%).
- La prevalencia de la luxación fue menor (20 casos, 2,5%) y fue más frecuente en el grupo de las grandes articulaciones (14 casos, 4,1%).
- Tan sólo encontramos derrame articular en 7 de las 28 radiografías de codo (prevalencia del 25%).

	Global de casos (n=792)	Grandes articulaciones (n=343)	Pequeñas articulaciones (n=407)	Huesos planos/largos (n=42)
<b>Prevalencia* (n, % [IC 95%])</b>				
Fractura aguda	134 (16,9 [14,0-19,2])	37 (10,8 [7,1-13,8])	84 (20,6 [16,0-24,5])	13 (30,9 [18-48,1])
Luxación aguda	20 (2,5 [1,4-3,6])	14 (4,1[2-6,4])	3 (0,7[0,15-2,1])	3 (7,1 [0,6-16,5])
Fractura crónica	25 (3,2 [2-4,5])	9 (2,6[1-4,6])	14 (3,4[1,9-5,7])	2 (4,8 [0,6-16,2])
Derrame (solo en codo, N=28)	7 (25,0[11-44,9])			

\*en base al diagnóstico emitido por el radiólogo adjunto

IC: Intervalo de confianza al 95% de la prevalencia

**Tabla 2:** Prevalencia global y por subgrupo.

# RESULTADOS

## 3. RENDIMIENTO DIAGNÓSTICO DE LA IA Y DEL RR

- Todos los resultados que mostramos a continuación comparan a la IA y al RR con la lectura del GS.
- Incluimos la frecuencia absoluta de los casos registrados como dudosos por el RR y por la IA y la proporción de los mismos que fue confirmada por el GS.
- Para el cálculo de S, E, los valores predictivos, el área bajo la curva (AUC) y sus respectivos intervalos de confianza del 95% se excluyeron los casos dudosos.

## VARIABLES ESTUDIADAS

- Evaluamos las variables que el software de IA disponible en nuestro centro está entrenado para detectar:
  - Fracturas.
  - Luxaciones.
  - Derrame articular en codo.
- También consideramos la variable fractura crónica (sólo evaluada por GS y RR) para indirectamente analizar si podía tener un impacto significativo en los resultados de detección de las fracturas agudas.

# RESULTADOS

## 3.1. FRACTURAS

- La S y la E fueron muy altas tanto para la IA como para el RR, con áreas bajo la curva excelentes para ambos. **(Tabla 3)**.
- Si prestamos atención al análisis por subgrupos la tendencia es similar, destacando especialmente los resultados del RR en el subgrupo de los huesos largos/planos.

	Global de casos (n=792)		Grandes articulaciones (n=343)		Pequeñas articulaciones (n=407)		Huesos planos/largos (n=42)	
	Validez diagnóstica del Residente	diagnóstica del Software de IA	diagnóstica del Residente	diagnóstica del Software de IA	diagnóstica del Residente	diagnóstica del Software de IA	diagnóstica del Residente	diagnóstica del Software de IA
Fractura aguda (ratio diagnósticos dudosos* / seguros)	12 / 780	58 / 734	5 / 338	13 / 330	6 / 401	41 / 366	1/41	4 / 38
Positividad de los casos dudosos (n, %)	6 (50)	15 (25,9)	3 (60)	4 (30,8)	3 (50)	9 (21,9)	0 (0)	2 (50)
Sensibilidad (% IC 95%)	90,6 (84,2 - 95,1)	95,8 (90,5 - 98,6)	94,1 (80,3 - 99,3)	93,9 (79,8-99,3)	87,7 (78,5-93,9)	96 (88,8-99,2)	100 (75,3-100)	100 (71,5-100)
Especificidad (% IC 95%)	98,0 (96,6 - 98,9)	97,6 (96,0 - 98,6)	100 (98,8-100)	99,7 (98,1-100)	95,9 (93,2-97,8)	95,5 (92,5-97,6)	100 (87,7-100)	96,3 (81-99,9)
VPP (% IC 95%)	89,9 (83,4 - 94,5)	88,4 (81,5 - 93,3)	100 (89,1- 100)	96,9 (83,8- 99,9)	84,5 (75-91,5)	84,7 (75,3-91,6)	100 (75,3-100)	91,7 (61,5-99,8)
VPN (% IC 95%)	98,2 (96,8 - 99,0)	99,2 (98,1 - 99,7)	99,3 (97,7-99,9)	99,3 (97,6-99,9)	96,8 (94,3-98,5)	98,9 (96,9-99,8)	100 (87,7-100)	100 (86,8-100)
Área bajo la curva (IC 95%)	0,943 ( 0,917 - 0,969)	0,967 ( 0,948 - 0,986)	0,971 (0,93-1)	0,968 (0,927-1)	0,918 (0,88-0,956)	0,958 (0,932-0,983)	1 (1-1)	0,981 (0,945-1)

\*Los casos con diagnóstico dudoso se excluyen en el cálculo de la S, E, VPP, VPN y área bajo la curva.

IC: Intervalo del confianza al 95%

NA: No aplica

NC: No calculable

**Tabla 3: Fractura aguda.**

# RESULTADOS

## 3.1. FRACTURAS

Globalmente el RR registró 12 casos dudosos de fractura, de los cuales el 50% fueron confirmados por el GS como casos positivos.

La IA registró un número mayor de casos dudosos (58 casos), pero la proporción de casos que finalmente resultaron positivos fue menor (sólo un 25,9%).



Fractura aguda de peroné distal registrada correctamente tanto por la IA como por el RR.

# RESULTADOS

## 3.1. FRACTURAS



Fractura aguda de la falange proximal del 5º dedo, marcada como dudosa por la IA, pero diagnosticada correctamente por el RR.



Ejemplo de variante anatómica (sesamoideo bipartito) marcada como fractura dudosa por la IA, y como negativa para fractura aguda por parte del RR.

# RESULTADOS

## 3.2. FRACTURA CRÓNICA (VARIABLE NO ANALIZADA POR IA)

Para analizar si las fracturas crónicas pueden ser un factor de confusión para la detección de fractura aguda realizamos un breve análisis estadístico de esta variable, para la que el software de IA no estaba entrenado, y que exponemos en la **Tabla 4**.

	Global de casos (n=792)		Grandes articulaciones (n=343)		Pequeñas articulaciones (n=407)		Huesos planos/largos (n=42)	
	Validez diagnóstica del Residente	diagnóstica del Software de IA	diagnóstica del Residente	diagnóstica del Software de IA	diagnóstica del Residente	diagnóstica del Software de IA	diagnóstica del Residente	diagnóstica del Software de IA
Fractura crónica (ratio diagnósticos dudosos* / seguros)	4/788	NA	2/341	NA	2/405	NA	0/42	NA
Positividad de los casos dudosos (n, %)	1 (25)		1 (50)		0 (0)		0 (0)	
Sensibilidad (% IC 95%)	29,2 (12,6-51,1)		12,5 (0,316-52,7)		35,7 (12,8-64,9)		50 (1,26-98,7)	
Especificidad (% IC 95%)	99,9 (99,3-100)		100 (98,9-100)		99,7 (98,6-100)		100 (91,2-100)	
VPP (% IC 95%)	87,5 (47,3-99,7)		100 (2,5-100)		83,3 (35,9-99,6)		100 (2,5-100)	
VPN (% IC 95%)	97,8 (96,5-98,7)		97,9 (95,8-99,2)		97,7 (95,8-99)		97,6 (87,1-99,9)	
Área bajo la curva (IC 95%)	0,645 (0,552-0,738)		0,563 (0,44-0,685)		0,677 (0,547-0,808)		0,75 (0,26-1)	

\*Los casos con diagnóstico dudoso se excluyen en el cálculo de la S, E, VPP, VPN y área bajo la curva.

IC: Intervalo del confianza al 95%

NA: No aplica

NC: No calculable

**Tabla 4:** Fractura crónica (variable no analizada por la IA).

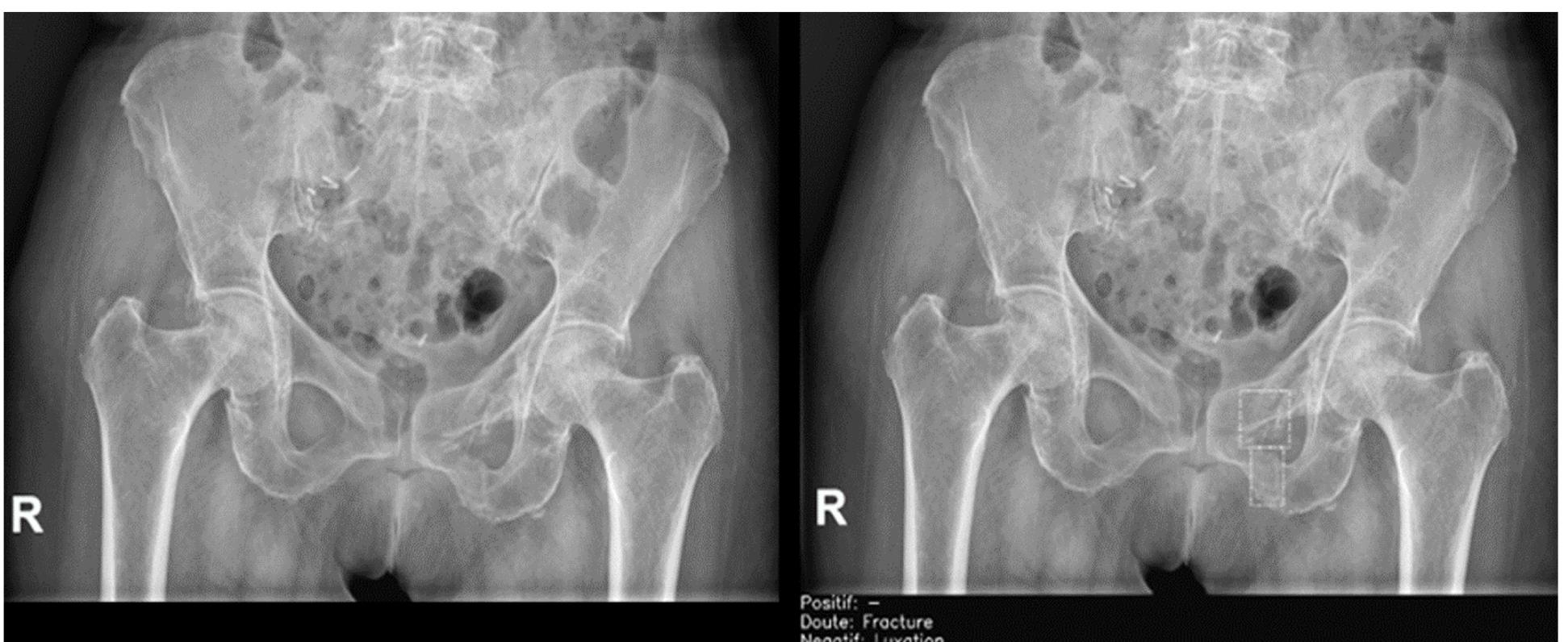
# RESULTADOS

## 3.2. FRACTURA CRÓNICA (VARIABLE NO ANALIZADA POR LA IA)

La baja prevalencia de esta variable (25 casos, 3.2%) condujo a bajos valores de sensibilidad y a intervalos de confianza anchos. Sin embargo, la E y los valores predictivos resultaron aceptables. **(Tablas 2 y 4).**

El RR registró 4 casos con resultado dudoso para fractura crónica, de los cuáles tan sólo uno fue confirmado por el GS.

Además los bajos valores de S tuvieron impacto sobre el cálculo del área bajo la curva, que sólo resultó aceptable para el subgrupo de los huesos largos/planos. **(Figura 6)**



Fractura crónica en pelvis registrada como dudosa tanto por la IA como por el RR.

# RESULTADOS

## 3.3. LUXACIÓN AGUDA

- La prevalencia de esta variable fue muy baja (20 casos, 2.5%), con 3 casos dudosos registrados por el RR de los que 2 fueron confirmados por el GS.
- La IA sólo registró dos casos dudosos, pero ninguno de ellos fue confirmado por el GS.
- No se registraron casos dudosos en el subgrupo de las pequeñas articulaciones.

	Global de casos (n=792)		Grandes articulaciones (n=343)		Pequeñas articulaciones (n=407)		Huesos planos/largos (n=42)	
	Validez diagnóstica del Residente	diagnóstica del Software de IA	diagnóstica del Residente	diagnóstica del Software de IA	diagnóstica del Residente	diagnóstica del Software de IA	diagnóstica del Residente	diagnóstica del Software de IA
Luxación aguda (ratio diagnósticos dudosos* / seguros)	3 / 789	2 / 790	2 / 341	2 / 341	0 / 407	0 / 407	1/41	0/42
Positividad de los casos dudosos (n, %)	2 (66,7)	0 (0)	1 (50)	0 (0)	0 (0)	0 (0)	1 (100)	0 (0)
Sensibilidad (%; IC 95%)	77,8 (52,4- 93,6)	35,0 (15,4 - 59,2)	84,6 (54,6-98,1)	35,7 (12,8-64,9)	66,7 (9,43-99,2)	66,7 (9,43-99,2)	50 (1,26-98,7)	NC
Especificidad (%; IC 95%)	100 (99,5-100)	99,7 (99,1-100)	100 (98,9-100)	100 (98,9-100)	100 (99,1-100)	99,5 (98,2-99,9)	100 (91-100)	NC
VPP (%; IC 95%)	100 (76,8-100)	77,8 (40-7,2)	100 (71,5-100)	100 (47,8-100)	100 (15,8-100)	50 (6,76-93,2)	100 (2,5-100)	NC
VPN (%; IC 95%)	99,5 (98,7-99,9)	98,3 (97,2-99,1)	99,4 (97,8-99,9)	97,3 (95-98,8)	99,8 (98,6-100)	99,8 (98,6-100)	97,5 (86,8-99,9)	NC
Área bajo la curva (IC 95%)	0,889 (0,79-0,988)	0,674 (0,566-0,781)	0,923 (0,821-1)	0,679 (0,548-0,809)	0,833 (0,507-1)	0,831 (0,504-1)	0,75 (0,26-1)	NC

\*Los casos con diagnóstico dudoso se excluyen en el cálculo de la S, E, VPP, VPN y área bajo la curva.

IC: Intervalo del confianza al 95%

NA: No aplica

NC: No calculable

Tabla 5: Luxación aguda.

# RESULTADOS

## 3.3. LUXACIÓN AGUDA

- Debido a su baja prevalencia los valores de S solo fueron aceptables en la lectura del RR dentro del subgrupo de las grandes articulaciones, sin embargo, sus IC resultaron muy anchos.
- Los valores de E fueron muy elevados en todos los subgrupos tanto para la IA como para el RR, excepto para el subgrupo de los huesos largos/planos en la lectura de la IA. En este subgrupo no se registró ninguna duda, sin embargo la IA tampoco detectó ninguna de las 3 luxaciones agudas, por lo que no se pudo realizar el análisis estadístico.
- El área bajo la curva mostró valores superiores a 0,7 para la lectura del RR en cada subgrupo, sin embargo, para la lectura de la IA sólo se alcanzaron dichos valores en el subgrupo de las pequeñas articulaciones. **(Tabla 5)**

# RESULTADOS



Luxación aguda de la articulación glenohumeral correctamente detectada por la IA y por el RR.

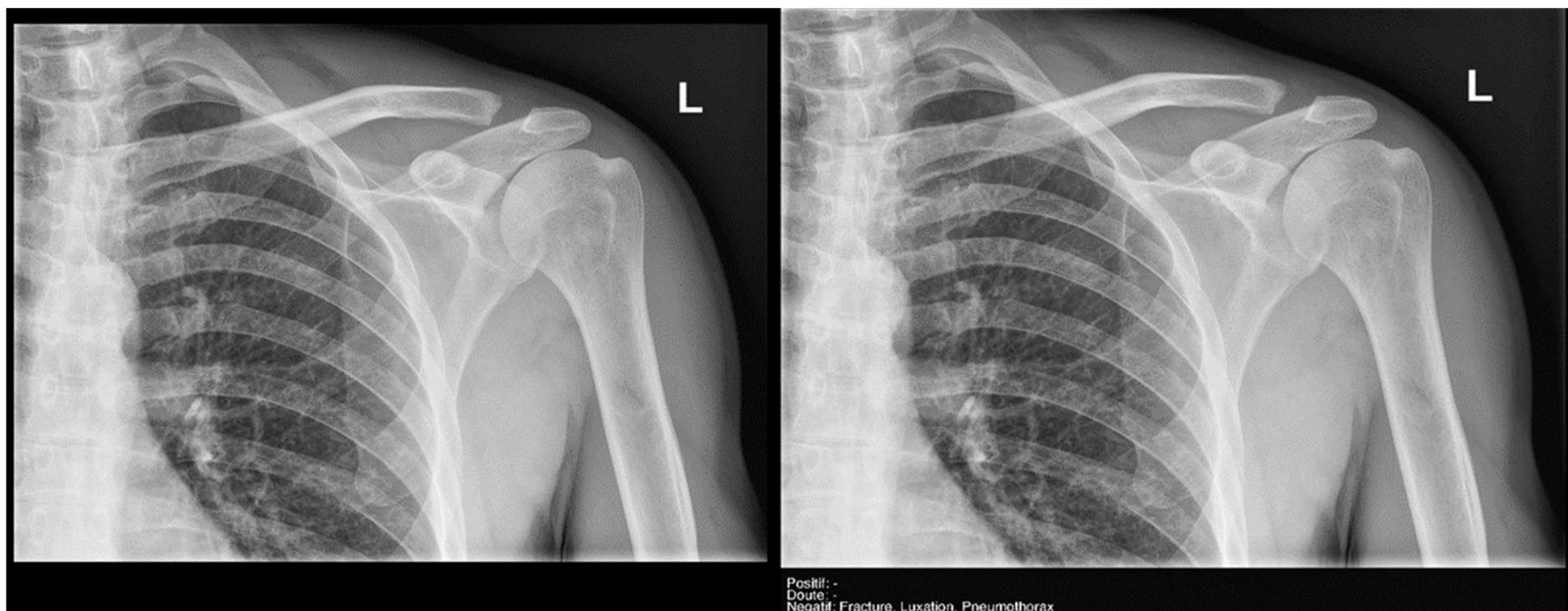


Fractura pélvica y luxación aguda coxofemoral, con resultado negativo para luxación aguda para la IA, pero correctamente detectada por el RR.

# RESULTADOS

## 3.3. LUXACIÓN AGUDA ACROMIOCLAVICULAR

Un caso especial es la luxación de la articulación acromioclavicular pues ninguno de los casos de nuestro estudio fueron detectados por la IA (todas se leyeron como negativas).



Luxación aguda de la articulación acromioclavicular con resultado negativo de la IA, pero correctamente detectada por el RR.

# RESULTADOS

## 3.4. DERRAME ARTICULAR (EN CODO)

La muestra incluyó 28 exámenes de codo, pero tan sólo 7 presentaron derrame articular.

La IA registró 4 casos dudosos para derrame articular, de estos sólo uno fue confirmado por el GS.

El RR no registró casos dudosos.

A pesar de que los valores de S para ambos (IA y RR) fueron del 100%, los IC del 95% resultaron muy amplios. En contraposición, la E y los valores predictivos negativos fueron más fiables, así como sus IC con áreas bajo la curva excelentes (**Tabla 6**).

	Global de casos (n=792)	
	Validez diagnóstica del Residente	diagnóstica del Software de IA
Derrame (solo en codo, ratio diagnósticos dudosos* / seguros)	0/28	4 / 24
Positividad de los casos dudosos (n, %)	0 (0)	1 (25)
Sensibilidad (% IC 95%)	100 (59-100)	100 (54,1-100)
Especificidad (% IC 95%)	90,5 (69,6-98,8)	94,4 (72,7-99,9)
VPP (% IC 95%)	77,8 (40-97,2)	85,7 (42,1-99,6)
VPN (% IC 95%)	100 (82,4-100)	100 (80,5-100)
Área bajo la curva (IC 95%)	0,952 (0,888-1)	0,972 (0,918-1)

\*Los casos con diagnóstico dudoso se excluyen en el cálculo de la S, E, VPP, VPN y área bajo la curva.

IC: Intervalo del confianza al 95%

NA: No aplica

NC: No calculable

**Tabla 6:** Derrame articular en codo.

# RESULTADOS



Fractura de la cabeza radial y derrame articular, ambos hallazgos correctamente detectados tanto por el RR como por la IA. En este la caja con línea discontinua que aparece sobre la imagen indica duda, pero en el informe de la IA se marca como positivo, lo que puede generar confusión.



Fractura de la cabeza radial, fractura de la apófisis coronoides y derrame articular, con resultado dudoso para ambas fracturas por parte de la IA y positivo para el derrame articular. El RR acierta en todos los hallazgos.

# RESULTADOS

## 4. CONCORDANCIA ENTRE LAS LECTURAS DE LA IA Y DEL RR

El grado de concordancia entre ambos lectores fue valorado mediante el cálculo del coeficiente de Kappa, que representamos mediante un mapa de calor.

**Kappa concordancia Residente - IA (IC 95%)**

Global	0.699 (0.686 - 0.736)	0.420 (0.340 - 0.523)	0.719 (0.684 - 0.814)
Pequeñas articulaciones	0.641 (0.604 - 0.653)	0.329 (0.159 - 0.817)	NA
Grandes articulaciones	0.797 (0.692 - 0.835)	0.489 (0.193 - 0.772)	0.719 (0.684 - 0.814)
Huesos largos / planos	0.712 (0.611 - 0.840)	NC	NA
		Fractura Aguda	Luxación aguda
			Derrame (codo)

Índice Kappa					
1	0,75	0,5	0,25		
Perfecta	Considerable	Moderada	Aceptable	Leve	Pobre

NA: No aplica

NC: No calculable

Heat map (mapa de calor) que representa el grado de correlación entre las lecturas de la IA y del RR. Para el global de la muestra en la detección de fracturas el coeficiente de Kappa resultó 0.69 y fue especialmente alto para el subgrupo de las grandes articulaciones (0.79).

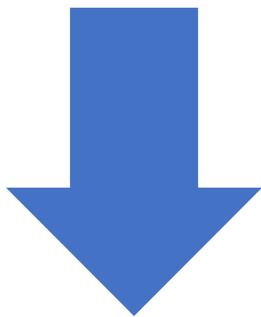
# RESULTADOS

## 4. CONCORDANCIA ENTRE LAS LECTURAS DE LA IA Y DEL RR

- Para las **fracturas** el coeficiente de Kappa global fue de 0.69 y fue especialmente alto para el subgrupo de las grandes articulaciones (0.79). El grado de concordancia fue aceptable para el resto de subgrupos.
- En cuanto a la **luxación aguda** la concordancia global y para el subgrupo de las grandes articulaciones logró un Kappa de alrededor de 0.4 (concordancia moderada).
- Para el subgrupo de las pequeñas articulaciones el grado de concordancia resultó aceptable-leve (0.3).
- El grado de concordancia no pudo ser calculado para el subgrupo de los huesos largos/planos pues la IA no detectó ninguna luxación aguda en este subgrupo.
- Encontramos un grado de concordancia considerable para el **derrame articular en codo**, con un coeficiente de Kappa de 0.719.
- Debemos tener en cuenta que el valor del coeficiente de Kappa depende de la prevalencia de la variable estudiada. En nuestro estudio la prevalencia de fractura aguda fue mucho mayor que la de luxación aguda y que la de derrame articular en codo.

# DISCUSIÓN

- La **radiografía simple** es la prueba de imagen más solicitada de primera línea para el diagnóstico de la patología osteoarticular más frecuente (fracturas y luxaciones).<sup>[1]</sup>
- En la práctica clínica habitual los médicos de la urgencia leen las radiografías por sí mismos, sin solicitar o esperar al informe del radiólogo, lo que lleva a un **importante número de fracturas no detectadas**, que según el estudio de Guly representan hasta un 80% de los errores en las lecturas de las radiografías óseas urgentes.<sup>[2]</sup>



- Varias casas comerciales han desarrollado **softwares de IA** capaces de detectar diferentes ítems en las radiografías simples, con **muy buenos resultados**.
- La IA en la radiografía simple ha venido para quedarse, pues según la literatura es una herramienta útil tanto para el screening como para la toma de decisiones aportando una segunda lectura de la prueba de imagen.<sup>[3]</sup>

# DISCUSIÓN

## FRACTURAS

- Nuestro estudio demuestra que la IA consigue valores altos de S y E para la detección de **fracturas agudas** tanto de manera global como en los diferentes subgrupos analizados.
- Nuestros resultados apoyan lo ya descrito por otros autores en la literatura, como en el metaanálisis de Kuo et al <sup>[3]</sup> donde obtienen valores medios muy elevados de S (91%) y de E (91%), tras analizar 42 estudios.
- Valores similares obtuvieron Chen et al. <sup>[4]</sup> tras el análisis con IA de 100 radiografías óseas y Zhang et al. <sup>[5]</sup> tras analizar otras 300 radiografías óseas obteniendo S de 94% y 92% respectivamente; y E de 98% en ambos estudios.

# DISCUSIÓN

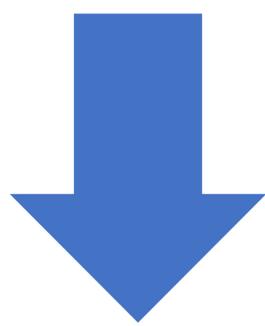
## LUXACIONES

- La baja prevalencia de **luxaciones agudas** en nuestra muestra llevó a valores de S y E con intervalos de confianza muy anchos.
- Estos valores también se vieron influidos por el error sistemático de la IA en la detección de luxación de la articulación acromioclavicular.
- En consecuencia encontramos valores de S más bajos que los descritos en la literatura, como en el estudio de Regnard et al <sup>[7]</sup> donde obtienen S elevadas de 89,9% tras analizar 4774 estudios de pelvis y extremidades.

# DISCUSIÓN

## LUXACIONES

- Sin embargo, sí obtenemos valores altos de E, similares a los descritos en el estudio de Regnard et al <sup>[7]</sup> (E = 99,1%).
- En el estudio de Wei et al. <sup>[8]</sup> analizan 106 radiografías de codo y otras 140 de hombro, obteniendo la IA AUC elevadas (>0,97), mientras que en nuestro estudio el AUC global resultó de 0,67, pues su cálculo se ve influido por valores bajos de S.



Postulamos que las diferencias entre nuestros resultados y las S y AUC publicadas en la literatura se deben tanto a una baja prevalencia de luxaciones agudas en nuestra muestra como a la incapacidad de la IA para la detección de luxaciones de la articulación acromioclavicular.

# DISCUSIÓN

## DERRAME ARTICULAR EN CODO

Encontramos valores de S, E y AUC muy elevados y similares a los publicados en la literatura.

Por ejemplo, en el estudio de Regnard et al <sup>[7]</sup> obtienen valores de S (91,5%) y E (99,8%) muy elevados en la detección de derrame articular en codo.

También publican resultados similares en el estudio de Huhtanen et al <sup>[9]</sup>, en el que analizan 4423 radiografías tanto de adultos como pediátricas. En el análisis de las radiografías de codo la IA obtuvo S (88,8%) y E (90,5%) muy elevadas.

El estudio de England et al <sup>[10]</sup> publica datos similares con AUC (0,985), S (90%) y E (90%) muy elevadas, tras analizar radiografías de codo urgentes de 882 pacientes pediátricos.

Con todo ello se demuestra que la IA es una herramienta muy útil también para la detección de este hallazgo.

# DISCUSIÓN

## LIMITACIONES

- No se trata de un estudio multicéntrico.
- Sólo incluye pacientes adultos.
- El GS se basa en la lectura de un radiólogo senior, sin considerar otras posibles pruebas confirmatorias.
- La baja prevalencia de algunas variables tuvo impacto sobre el cálculo de algunos parámetros estadísticos como el AUC, IC del 95% y los coeficientes de Kappa.
- La lectura de las radiografías se realizó sin tener acceso a la exploración física ni a la historia clínica de los pacientes, lo que es esencial en la vida real al evaluar radiografías simples solicitadas desde el servicio de urgencias.

# CONCLUSIONES

- En nuestro estudio la IA logra resultados prometedores en la detección de fracturas (en especial en el subgrupo de las grandes articulaciones).
- La IA obtuvo más resultados dudosos que el residente de radiología, y éstos fueron finalmente “positivos” también en menor proporción (25% para fracturas y derrame articular en codo) que los resultados dudosos del residente.
- La concordancia interobservador fue entre moderada y considerable para cada variable en todos los grupos.
- El principal punto débil de la IA es el error sistemático en la detección de la luxación acromioclavicular.
- En conclusión, la IA podría ser una herramienta útil como screening en la lectura de radiografías simples en los servicios de Urgencia, pues podría ayudar a reducir las tasas de fracturas, luxaciones y derrames articulares no detectados en la práctica clínica.

# REFERENCIAS

- [1] Duron L, Ducarouge A, Gillibert A, Lainé J, Allouche C, Cherel N, et al. Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study. *Radiology* 2021;300(1):120–9.
- [2] Guly HR. Diagnostic errors in an accident and emergency department. *Emerg Med J* 2001;18(4):263–9.
- [3] Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. *Radiology* 2022;304(1):50–62.
- [4] Chen Y, Wang X, Zhang Y, Li L. Artificial intelligence for fracture detection in radiographs: a systematic review and meta-analysis. *J Med Imaging Radiat Oncol* 2022;66(1): 11–20.
- [5] Zhang Y, Yang L, Chen Y, Zhou X. Deep learning in orthopedic research and clinical practice. *Front Bioeng Biotechnol* 2021;9.
- [6] Wang Z, Cui H, Zhang Q, Liu L, Zhang Y. Deep learning-based automatic classification of distal radius fractures on computed tomography. *Med Phys* 2021;48(1), 90–102.
- [7] Regnard NE, Lanseur, B, Ventre, J, Ducarouge, A, Clovis, L, Lassalle L, et al. Assessment of performances of a deep learning algorithm for the detection of limbs and pelvic fractures, dislocations, focal bone lesions, and elbow effusions on trauma X-rays. *Eur J Radiol* 2022;154 (110447):110447.
- [8] Wei J, Li D, Sing DC, Beeram I, Puvanesarajah V, Tornetta P 3rd, et al. Detecting upper extremity native joint dislocations using deep learning: A multicenter study. *Clin Imaging* 2022; 92:38–43.
- [9] Huhtanen JT, Nyman M, Doncenco D, Hamedian M, Kawalya D, Salminen L, et al. Deep learning accurately classifies elbow joint effusion in adult and pediatric radiographs. *Sci Rep.* 2022;12(1): 11803.
- [10] England JR, Gross JS, White EA, Patel DB, England JT, & Cheng PM. Detection of Traumatic Pediatric Elbow Joint Effusion Using a Deep Convolutional Neural Network. *AJR. American journal of roentgenology* 2018;211(6), 1361–1368.