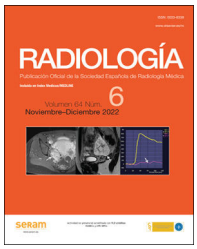
Radiología 67 (2025) 273–286

RADIOLOGÍA

www.elsevier.es/rx



ORIGINAL

Usar una herramienta comercial de inteligencia artificial no entrenada para COVID-19 mejora ligeramente la interpretación de las radiografías con neumonía COVID-19, especialmente entre lectores inexpertos

M. Pérez Laencina[a], J.M. Plasencia Martínez[b,*], M. Sánchez Canales[b], C. Jiménez Pulido[b], R. Rodríguez Mondéjar[b], L. Martínez Encarnación[b], C. García Hidalgo[a], D. Galdo Galián[a], P. Hernández Madrid[a], L. Chico Caballero[a], E. Guillén García[a], M.N. Plasencia Martínez[c], S. Martínez Romero[d], J. García Molina[e] y J.M. García Santos[b]

[a] *Facultad de Medicina, Universidad de Murcia, Murcia, España*
[b] *Servicio de Radiodiagnóstico, Hospital Universitario Morales Meseguer, Murcia, España*
[c] *Medicina de Familia, Consultorio El Albujón, Cartagena; Gerencia Área II de Salud, Cartagena, Murcia, España*
[d] *Medicina de Familia, Centro de Salud Cieza Oeste, Hospital Vega Lorenzo Guirao, Cieza, Murcia, España*
[e] *Medicina de Familia, Centro de Salud La Flota, Vistalegre; Gerencia Área VI de Salud, Murcia, España*

Recibido el 28 de septiembre de 2023; aceptado el 8 de enero de 2024
Disponible en Internet el 4 de abril de 2024

PALABRAS CLAVE
COVID-19;
Inteligencia artificial;
Diagnóstico;
Radiografía;
Estudiantes de medicina

Resumen

Introducción: Nuestos objetivo es evaluar la utilidad de una herramienta de inteligencia artificial (IA) para los lectores de radiografías de tórax con distintos niveles de experiencia para diagnosticar la neumonía COVID-19 cuando la herramienta ha sido entrenada en patología diferente a neumonía COVID-19.

Métodos: Se recogieron datos de los pacientes que se habían sometido previamente a una radiografía de tórax y a una tomosíntesis digital por sospecha de neumonía COVID-19. El estándar de referencia consistió en las lecturas de dos radiólogos expertos que evaluaron la presencia y la distribución de la neumonía COVID-19 en las imágenes. Seis estudiantes de medicina, dos residentes de radiología y otros dos radiólogos torácicos expertos participaron como lectores adicionales. Se realizaron dos lecturas radiográficas sin emplear la herramienta, y una tercera con el apoyo de la herramienta de IA Thoracic Care Suite. Se evaluaron la distribución y la probabilidad de la neumonía COVID-19 junto con la contribución de la IA. Se analizaron la concordancia y el rendimiento diagnóstico.

* Autor para correspondencia.
Correo electrónico: plasen79@gmail.com (J.M. Plasencia Martínez).

https://doi.org/10.1016/j.rx.2024.01.007

Resultados: La muestra estaba formada por 113 casos, de los cuales 56 presentaban opacidades pulmonares; el 52,2% eran mujeres y la edad media era de 50,70 ± 14,9 años. La concordancia con el estándar de referencia difirió entre estudiantes, residentes y radiólogos. Hubo una mejora no significativa para cuatro de los seis estudiantes cuando se utilizó la IA. El uso de la IA por parte de los estudiantes no mejoró el rendimiento diagnóstico de la neumonía COVID-19, pero sí redujo la diferencia en el rendimiento diagnóstico con los radiólogos más expertos. Además, influyó más en la interpretación de la neumonía leve que en la de la grave y de los hallazgos radiográficos normales. La IA resolvió más dudas de las que generó, especialmente entre los estudiantes (31,30 frente al 8,32%), seguidos de los residentes (14,45 frente al 5,7%) y los radiólogos (10,05% frente al 6,15%).
Conclusión: Tanto para los radiólogos expertos como para los menos experimentados, esta herramienta comercial de IA no ha mostrado ningún impacto en las lecturas de radiografías de tórax de pacientes con sospecha de neumonía COVID-19. Sin embargo, ayudó a la evaluación de los lectores inexpertos y en los casos de neumonía leve.

KEYWORDS
COVID-19;
Artificial intelligence;
Diagnosis;
Radiograph;
Medical Students

A commercial AI tool untrained for COVID-19 demonstrates slight improvement in the interpretation of COVID-19 pneumonia x-rays, especially among inexperienced readers

Abstract
Introduction: Our objective is to evaluate how useful an artificial intelligence (AI) tool is to chest radiograph readers with various levels of expertise for the diagnosis of COVID-19 pneumonia when the tool has been trained on a non-COVID-19 pneumonia pathology.
Methods: Data was collected for patients who had previously undergone a chest radiograph and digital tomosynthesis due to suspected COVID-19 pneumonia. The gold standard consisted of the readings of two expert radiologists who assessed the presence and distribution of COVID-19 pneumonia on the images. Six medical students, two radiology trainees, and two other expert thoracic radiologists participated as additional readers. Two radiograph readings and a third supported by the AI Thoracic Care Suite tool were performed. COVID-19 pneumonia distribution and probability were assessed along with the contribution made by AI. Agreement and diagnostic performance were analysed.
Results: The sample consisted of 113 cases, of which 56 displayed lung opacities, 52.2% were female, and the mean age was 50.70 ± 14.9. Agreement with the gold standard differed between students, trainees, and radiologists. There was a non-significant improvement for four of the six students when AI was used. The use of AI by students did not improve the COVID-19 pneumonia diagnostic performance but it did reduce the difference in diagnostic performance with the more expert radiologists. Furthermore, it had more influence on the interpretation of mild pneumonia than severe pneumonia and normal radiograph findings. AI resolved more doubts than it generated, especially among students (31.30 vs 8.32%), followed by trainees (14.45 vs 5.7%) and radiologists (10.05% vs 6.15%).
Conclusion: For expert and lesser experienced radiologists, this commercial AI tool has shown no impact on chest radiograph readings of patients with suspected COVID-19 pneumonia. However, it aided the assessment of inexperienced readers and in cases of mild pneumonia.

Introducción

La primera prueba de imagen que debe realizarse cuando se sospecha una neumonía COVID-19 es la radiografía de tórax, debido a su menor coste y su mayor disponibilidad[1]. Aunque se trata de una prueba muy accesible, hay muchos lugares en el mundo que no disponen del equipo necesario ni de radiólogos expertos formados para interpretar correctamente las pruebas[2]. Por otro lado, en los países desarrollados la sobrecarga de trabajo y los limitados recursos de los radiólogos hacen que muchos departamentos de radiología no informen sistemáticamente las radiografías procedentes del servicio de urgencias, circunstancia que puede ocurrir especialmente en los picos pandémicos, cuando la sobrecarga puede aumentar. Por lo tanto, la inteligencia artificial (IA) puede resultar una herramienta útil para ayudar a los profesionales sanitarios con menos experiencia en radiología a interpretar las radiografías de tórax.

Los datos disponibles de publicaciones anteriores sobre IA no se traducen necesariamente en resultados clínicos[3], y

los programas informáticos comerciales no están entrenados para procesos específicos, sino que responden a preguntas básicas basadas en la detección de lesiones como nódulos o consolidaciones[4]. Por este motivo, es interesante evaluar el impacto de estas herramientas de IA disponibles en el mercado, que no están específicamente entrenadas para diagnosticar las complicaciones torácicas de la COVID-19. Para ello, será necesario incluir en el análisis a lectores con diferentes grados de experiencia para investigar el impacto de la herramienta de IA en cada uno de ellos.

Nuestra hipótesis es que un software comercial de IA tendrá un impacto en la lectura de radiografías de tórax de pacientes con sospecha de neumonía COVID-19 en todos los lectores, independientemente de su experiencia, de menor a mayor experiencia (estudiantes, residentes de radiología, radiólogos).

Nuestros objetivos son analizar: 1) la concordancia intraobservador para determinar la «probabilidad de neumonía COVID-19» con radiografía y la «probabilidad de afectación de cada zona pulmonar»; 2) la concordancia de cada lector con el estándar de referencia para determinar la «probabilidad de neumonía COVID-19» con radiografía y la «probabilidad de afectación de cada zona pulmonar», con y sin apoyo de la IA; 3) el rendimiento diagnóstico de los lectores antes y después de aplicar la IA, y 4) la opinión de los lectores sobre el beneficio global de la IA.

Materiales y métodos

Este estudio ha sido autorizado por nuestro Comité Ético Institucional (Código interno EST: 38/20). Debido a las características del proyecto, el Comité no consideró necesario que los pacientes incluidos firmaran un consentimiento informado.

Pacientes

La muestra se seleccionó a partir de una base de datos creada durante las tres primeras oleadas de la enfermedad. Los criterios de inclusión fueron: 1) sospecha/confirmación de infección por COVID-19; 2) sospecha de neumonía, y 3) realización de una radiografía de tórax y una tomosíntesis digital torácica 3D para que pudiera servir como prueba de referencia. El único criterio de exclusión fue que las imágenes no tuvieran la calidad suficiente para ser evaluadas.

Dos radiólogos con 14 y 11 años de experiencia de la Unidad de Radiología de Urgencias evaluaron la probabilidad de afectación (sí/no) de cada zona pulmonar en las radiografías de tórax de 480 pacientes utilizando tanto radiografías como imágenes de tomosíntesis 3D. De ellos, se seleccionaron aquellos casos en los que los dos radiólogos coincidieron en su veredicto de afectación (sí/no) para todas las zonas pulmonares (fig. 1). Así, el grupo de radiografías patológicas estaba formado por 56 casos con concordancia en la distribución de las opacidades pulmonares. Cuando los radiólogos discreparon en la determinación de la probabilidad de neumonía COVID-19, se tomó una decisión final por consenso. Además, se añadió una muestra aleatoria de 57 casos con radiografías de tórax sin opacidades según ambos radiólogos, que constituyeron el grupo de radiografías normales (fig. 2).

Análisis de las imágenes

Participaron en el análisis los siguientes lectores: dos estudiantes de medicina de cuarto año (estudiante 4-1 y estudiante 4-2), dos estudiantes de medicina de quinto año (estudiante 5-1 y estudiante 5-2), dos estudiantes de medicina de sexto año (estudiante 6-1 y estudiante 6-2), un residente de radiología de segundo año, un residente de radiología de tercer año y dos radiólogos expertos distintos del estándar de referencia (radiólogo 1 y radiólogo 2), con 20 y 13 años de experiencia en radiología torácica, respectivamente. El número total de radiografías se evaluó en 3-5 sesiones de 3-4 horas cada una para cada evaluación.

Parte 1 (figura 3)
Antes de la primera evaluación de las radiografías se realizó un breve seminario para explicar los aspectos técnicos de la evaluación. A continuación se accedió a las radiografías de forma anónima mediante el número de historia clínica y la fecha de la prueba.

Las variables analizadas en la primera evaluación, descritas en la tabla 1, fueron:

Probabilidad de neumonía COVID-19 según las radiografías, basada en la clasificación de la Sociedad Británica de Radiología Torácica[5] (figs. 4-5).

Probabilidad de afectación en cada zona pulmonar. Para ello, se analizaron las radiografías de tórax tras dividir la imagen en 6 zonas pulmonares (fig. 1).

Parte 2 (figura 3)
Esta parte se separó en el tiempo de la parte 1 al menos 4 semanas.

Se llevó a cabo una breve fase de preparación, consistente en un vídeo realizado exclusivamente para este proyecto por General Electric Healthcare sobre cómo utilizar el software de IA Thoracic Care Suite (GE Healthcare. Milwaukee, WI, EE.UU.).

A continuación, para poder determinar la variabilidad intraobservador y no atribuir esta variación por completo a la interpretación de la imagen procesada por IA, se realizó una segunda evaluación de las radiografías en la que se analizaron las mismas variables que en la primera evaluación.

Inmediatamente después, la radiografía de tórax y la imagen con los resultados del software de IA se abrieron juntas en una tercera evaluación (fig. 3). Las variables evaluadas fueron las mismas que en la primera y la segunda evaluación, y dos más para describir sus opiniones sobre la contribución de la imagen de IA: a) cambio de la primera impresión tras observar la imagen de IA en cada zona pulmonar, y b) contribución global de la IA (tabla 1).

Análisis y estadísticas realizados

La evaluación de la radiografía de tórax y la tomosíntesis por los dos radiólogos de la Unidad de Radiología de Urgencias se consideró el estándar de referencia.

Las variables «probabilidad de neumonía COVID-19» y «probabilidad de afectación en cada zona pulmonar» se dicotomizaron (tabla 1).

Los análisis estadísticos realizados fueron los siguientes (fig. 3):

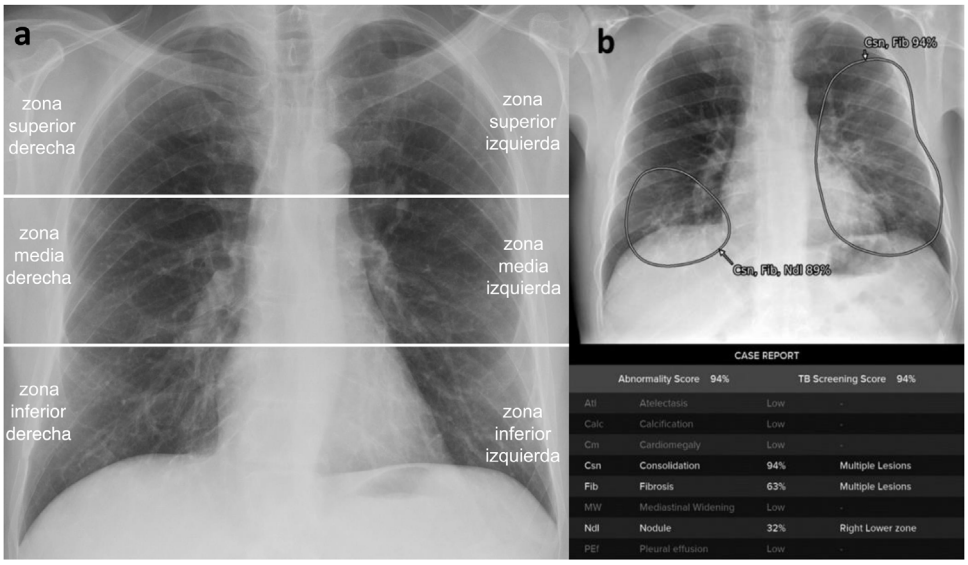


Figura 1 a) Radiografía posteroanterior de tórax con hallazgos normales. División de la radiografía en 6 campos pulmonares. Para ello se trazó una línea horizontal superior (borde inferior del arco aórtico) y una línea horizontal inferior (borde inferior de la vena pulmonar inferior derecha).

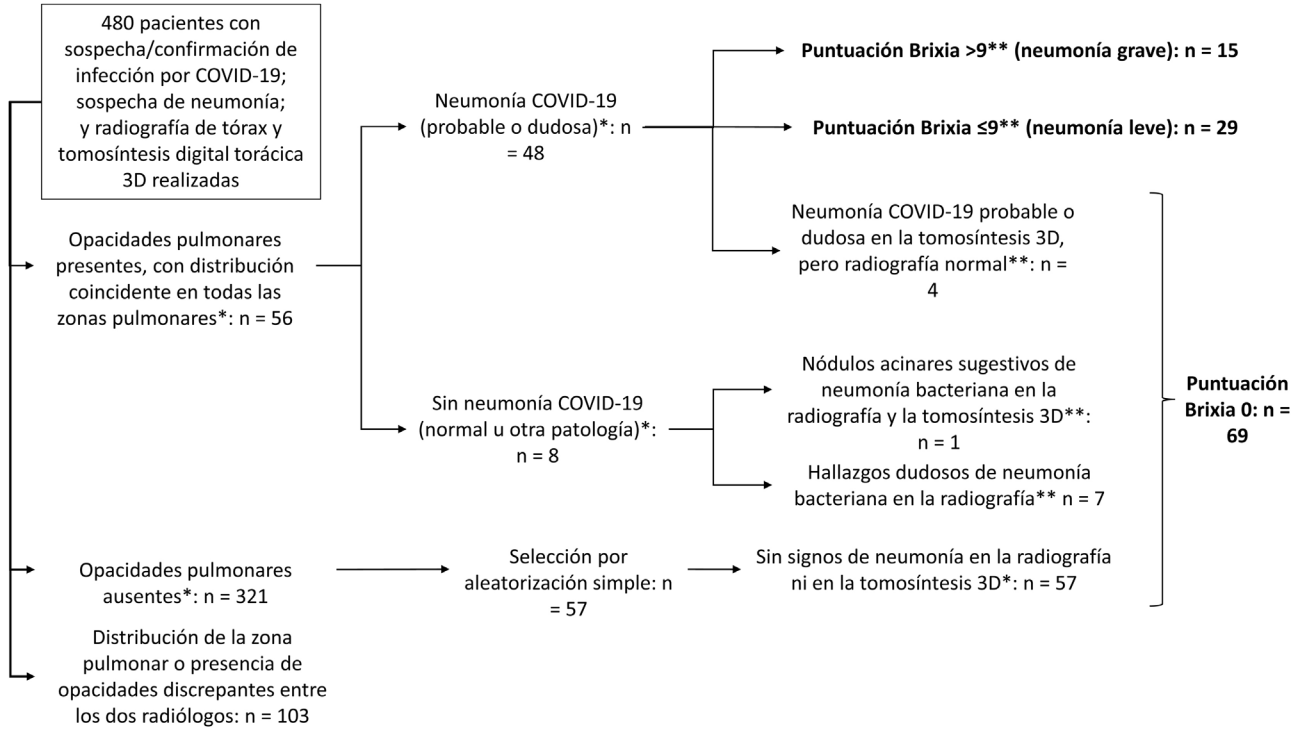


Figura 2 Selección de muestras.
Se muestran en negrita los casos en los que se evaluó la contribución de la inteligencia artificial en función de la gravedad de la neumonía y en los casos normales frente a los patológicos.
* Según los dos radiólogos expertos que compusieron la prueba de referencia tras evaluar a ciegas las radiografías y las tomosíntesis 3D.
** Según el radiólogo más experimentado que haya evaluado la radiografía.

Tabla 1 Definición de variables, dicotomización y evaluación donde se reclutaron las variables

Variables	Opciones y dicotomizaciones (flechas)		Evaluaciones de los lectores
Probabilidad de neumonía COVID-19	COVID-19 clásico/probable (figura 2) COVID-19 indeterminado (figura 3)	→ Neumonía COVID-19 (probable o dudosa)	1.ª radiografía 2.ª radiografía 3.ª radiografía + IA
	Normal (figura 2) No COVID-19/otra patología (figura 3)	→ Neumonía no COVID-19 (improbable)	
Probabilidad de afectación de cada zona pulmonar	Tengo claro que la zona pulmonar es normal Creo que la zona pulmonar es normal No sé si la zona pulmonar es normal o patológica (excluido)	→ Afectación	1.ª radiografía 2.ª radiografía 3.ª radiografía + IA
	Creo que la zona pulmonar es anormal Tengo claro que la zona pulmonar es anormal	→ Sin afectación	
Cambio en la primera impresión tras evaluar la imagen de IA en cada zona pulmonar	No, no he cambiado mi primera impresión Arroja dudas sobre los hallazgos Sí, he cambiado mi primera impresión		3.ª radiografía + IA
¿Qué me ha aportado la IA en general (contribución general)?	Nada Me ha generado más dudas Me ha resuelto dudas		3.ª radiografía + IA

IA: inteligencia artificial.

Análisis de concordancias
Se analizaron las siguientes:

- Concordancia intraobservador para «probabilidad de neumonía COVID-19» y «probabilidad de afectación en cada zona pulmonar» de la primera y segunda evaluación.
- Concordancia interobservador con el estándar de referencia para estudiantes, residentes y radiólogos para «probabilidad de neumonía COVID-19» y «probabilidad de afectación en cada zona pulmonar» en la primera evaluación, la segunda evaluación y en la evaluación de la radiografía de tórax y la de IA combinadas. Comparamos los resultados de la concordancia de las tres evaluaciones de cada lector para determinar la mejora potencial con la IA.

Se utilizó el coeficiente Kappa de Cohen (K) con intervalos de confianza del 95% (IC 95%). El grado de concordancia se estableció según el método de Landis y Koch[6]. La concordancia se consideró deficiente (inferior a 0,2), débil (0,21-0,4), moderada (0,41-0,6), buena (0,61-0,8) o excelente (superior a 0,80).

Se consideró que el software de IA contribuía al lector cuando el valor K al utilizar la IA aplicada a la radiografía superaba el valor K de la primera y la segunda evaluación de cada lector, para las que también presentamos los porcentajes de mejora.

Rendimiento diagnóstico
El rendimiento diagnóstico para la «probabilidad de neumonía COVID-19» de la primera evaluación, la segunda evaluación y la evaluación con IA para cada lector se analizó con la Receiver Operating Curve-ROC, y las áreas bajo la curva (AUC) resultantes se compararon con la prueba de DeLong. Las diferencias de proporciones entre grupos (entre las distintas evaluaciones de estudiantes, residentes y radiólogos) para cada evaluación se analizaron con la prueba de ji al cuadrado (χ^2) o la prueba exacta de Fischer. La evolución a lo largo de las tres evaluaciones sucesivas del número de diferencias estadísticamente significativas en el rendimiento diagnóstico (DESRD) de los estudiantes entre sí, y con residentes y radiólogos.

Además, para determinar si la contribución de la evaluación de la imagen de IA fue favorable o desfavorable, se contabilizó el número de veces que cambiaba la interpretación del lector en la «probabilidad de afectación de cada zona pulmonar», así como el número de veces que este cambio concuerda con el estándar de referencia y el número de veces que discrepa. Además, para determinar si sería útil disponer de la imagen IA en los casos de neumonía con manifestaciones radiológicas sutiles, se hizo una distinción para la neumonía COVID-19 con puntuación Brixia en el grupo de 56 pacientes con radiografías patológicas. Para evaluar la puntuación Brixia, la radiografía de tórax se divide en seis zonas pulmonares (fig. 1a), y las opacidades de cada zona se cuantifican como 0: normal; 1: opacidades en vidrio deslustrado (OVD); 2: OVD y consolidaciones con predominio de OVD; 3: consolidaciones con o sin OVD, con predominio de consolidaciones. La puntuación oscila entre 0 y 18 puntos[7]. Se comparó el rendimiento diagnóstico para una puntuación brixia inferior o igual a 9 puntos con una puntuación superior a 9 puntos, según la valoradión del radiólogo de urgencias con 14 años de experiencia, ya que el punto medio de gravedad según la puntuación del índice Brixia es de 9 puntos (fig. 2). Las frecuencias de este recuento se analizaron de forma descriptiva.

Estudio de las frecuencias de opinión en la lectura conjunta de la radiografía con IA
Se analizaron descriptivamente las frecuencias de las respuestas de cada lector a las preguntas relacionadas con su percepción de la contribución de la IA.

Los programas estadísticos utilizados para el análisis fueron IBM SPSS Statistics versión 20 y MedCalc versión 12.7.0.0. Un valor de $p < 0,05$ se consideró un resultado significativo.

Resultados

Muestra

Se obtuvo una muestra final de 113 pacientes: 56 con opacidades pulmonares, 48 que sugerían neumonía COVID-19 y

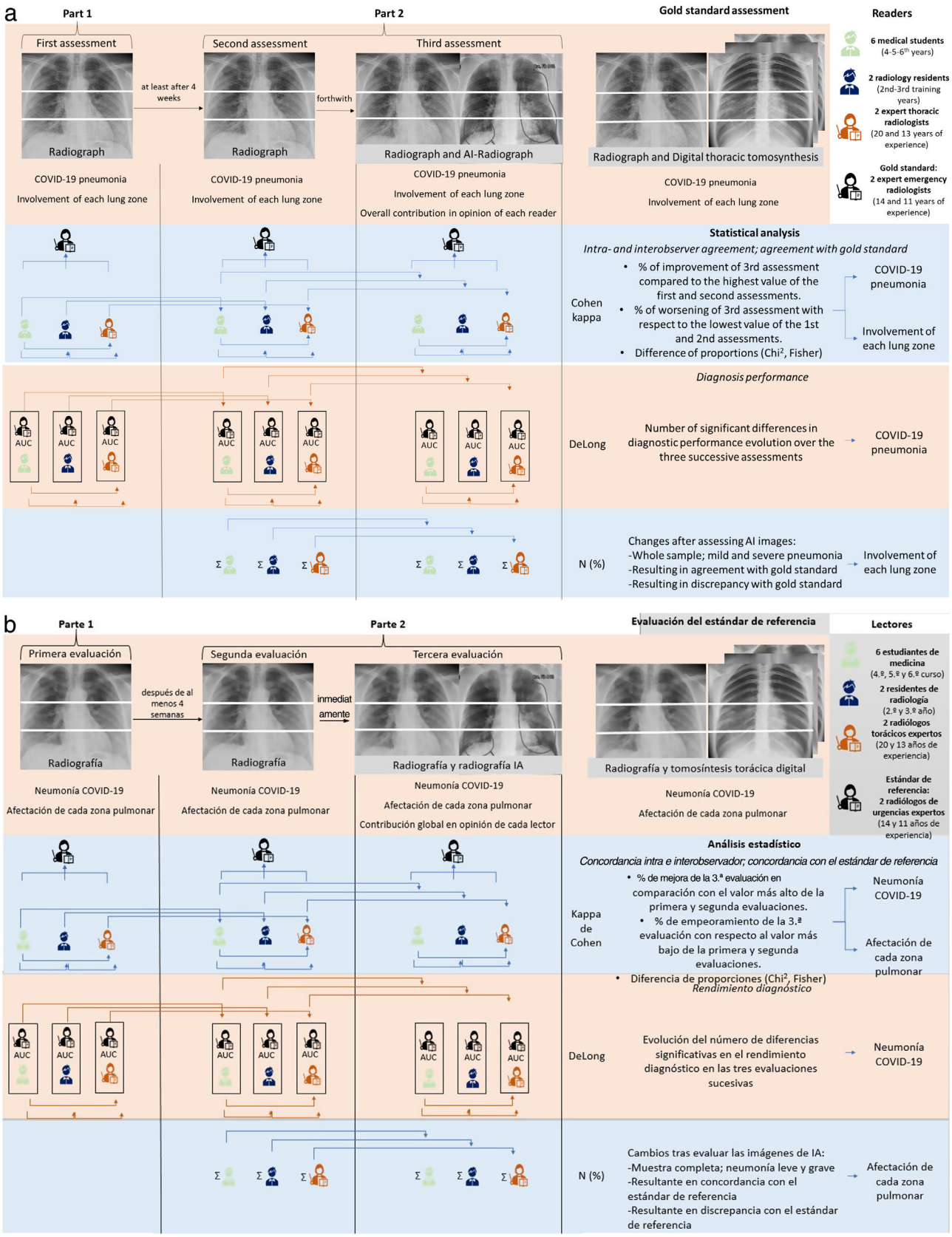


Figura 3 Descripción del proceso de recogida de datos, los lectores, las variables analizadas en cada evaluación y los análisis estadísticos realizados para cada comparación (flechas entre lectores y el estándar de referencia).
AUC [*area under the curve*]: área bajo la curva; IA: inteligencia artificial.

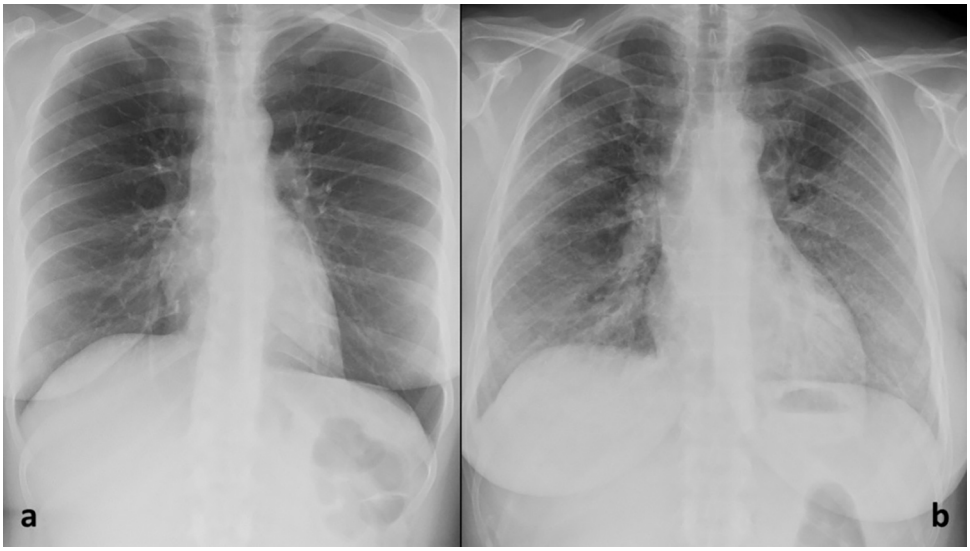


Figura 4 a) Radiografía posteroanterior de tórax con hallazgos normales. b) Radiografía posteroanterior de tórax con hallazgos compatibles con neumonía COVID-19 clásica.

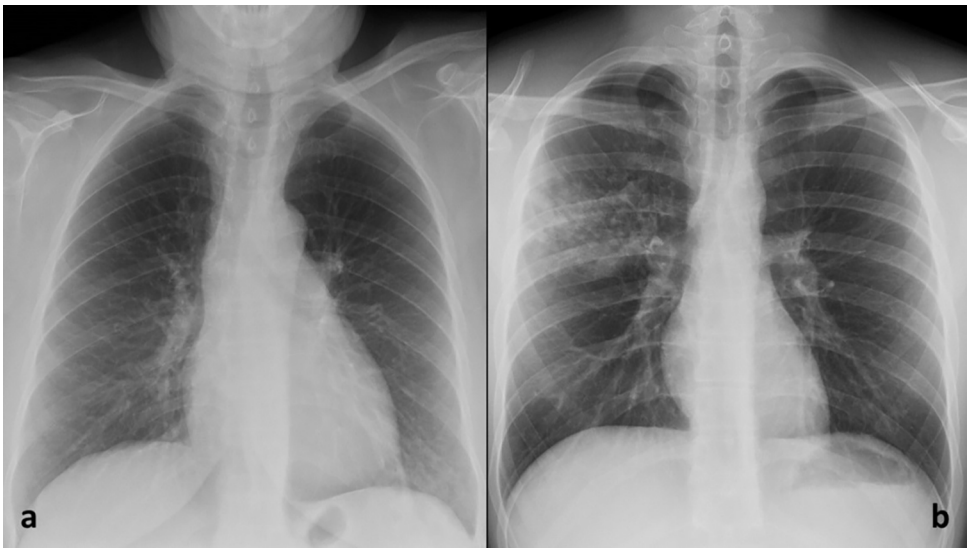


Figura 5 a) Radiografía posteroanterior de tórax con hallazgos compatibles con neumonía COVID-19 indeterminada. b) Radiografía posteroanterior de tórax con hallazgos compatibles con enfermedad no COVID-19 (neumonía bacteriana).

15 con puntuación Brixia > 9 puntos (fig. 2). El 52,2% eran mujeres, con una edad media ± desviación estándar, una mediana de edad [rango intercuartílico] y un rango de edad de 50,70 ± 14,9, 49 [40,5-62] y 19-88 años.

No se excluyó ningún estudio.

Para cada evaluación (fig. 3) se realizaron entre 3 y 5 sesiones de 3-4 horas cada una.

Análisis de concordancias

Teniendo en cuenta todos los lectores, se obtuvieron 10 valores K para la «probabilidad de neumonía COVID-19» en cada una de las tres evaluaciones, y 60 valores K para la «probabilidad de afectación de cada zona pulmonar», respectivamente.

Concordancias intraobservador entre la primera y la segunda evaluación (tabla 2)

Para la «probabilidad de neumonía COVID-19» con radiografía, las concordancias intraobservador fueron predominantemente buenas (7/10 [70%]) frente a concordancias excelentes (2/10 [20%]) y moderadas (1/10 [10%]).

Para la «probabilidad de afectación de cada zona pulmonar» con la radiografía, la concordancia intraobservador fue predominantemente buena (37/60 [61,66%]), con una proporción menor de concordancias excelentes (18/60

Tabla 2 Concordancias intraobservador

	Probabilidad de neumonía COVID-19	Zona superior derecha	Zona central derecha	Zona inferior derecha	Zona superior izquierda	Zona central izquierda	Zona inferior izquierda
Estudiante 4-1	0,848 (0,724-0,939)	0,822 (0,641-0,960)	0,856 (0,725-0,957)	0,840 (0,704-0,939)	0,755 (0,538-0,918)	0,860 (0,732-0,954)	0,838 (0,706-0,942)
Estudiante 4-2	0,520 (0,370-0,665)	0,395 (0,056-0,672)	0,674 (0,528-0,811)	0,706 (0,550-0,841)	0,633 (0,299-0,873)	0,557 (0,415-0,687)	0,530 (0,373-0,687)
Estudiante 5-1	0,656 (0,495-0,799)	0,836 (0,694-0,948)	0,771 (0,637-0 0,881)	0,755 (0,621-0,879)	0,710 (0,546-0,851)	0,769 (0,627-0,890)	0,797 (0,671-0,905)
Estudiante 5-2	0,691 (0,539-0,825)	0,701 (0,538-0,846)	0,721 (0,573-0,839)	0,680 (0,542-0,806)	0,479 (0,301-0,663)	0,757 (0,628-0,876)	0,725 (0,594-0,840)
Estudiante 6-1	0,629 (0,504-0,764)	0,810 (0,652-0,939)	0,639 (0,487-0,775)	0,685 (0,545-0,816)	0,816 (0,668-0,939)	0,781 (0,647-0,898)	0,617 (0,468-0,760)
Estudiante 6-2	0,604 (0,457-0,750)	0,732 (0,536-0,886)	0,676 (0,534-0,822)	0,632 (0,481-0,770)	0,565 (0,320-0,779)	0,676 (0,539-0,808)	0,664 (0,529-0,782)
Residente 2	0,683 (0,541-0,806)	0,724 (0,576-0,854)	0,716 (0,577-0,844)	0,716 (0,585-0,841)	0,805 (0,661-0,928)	0,792 (0,675-0,903)	0,771 (0,636-0,885)
Residente 3	0,855 (0,747-0,946)	0,926 (0,831-1,000)	0,880 (0,783-0,962)	0,941 (0,871-1,000)	0,806 (0,668-0,918)	0,887 (0,791-0,963)	0,904 (0,809-0,980)
Radiólogo 1	0,610 (0,460-0,749)	0,718 (0,507-0,872)	0,691 (0,539-0,821)	0,759 (0,632-0,871)	0,727 (0,555-0,878)	0,725 (0,582-0,857)	0,815 (0,689-0,920)
Radiólogo 2	0,704 (0,574-0,824)	0,815 (0,679-0,928)	0,780 (0,661-0,888)	0,826 (0,707-0,928)	0,738 (0,592-0,863)	0,784 (0,657-0,889)	0,711 (0,570-0,839)

Coeficiente kappa (K) con intervalos de confianza del 95% entre paréntesis para la concordancia intraobservador en la primera y segunda lecturas de las radiografías de tórax para cada zona pulmonar y para la probabilidad de neumonía COVID-19.

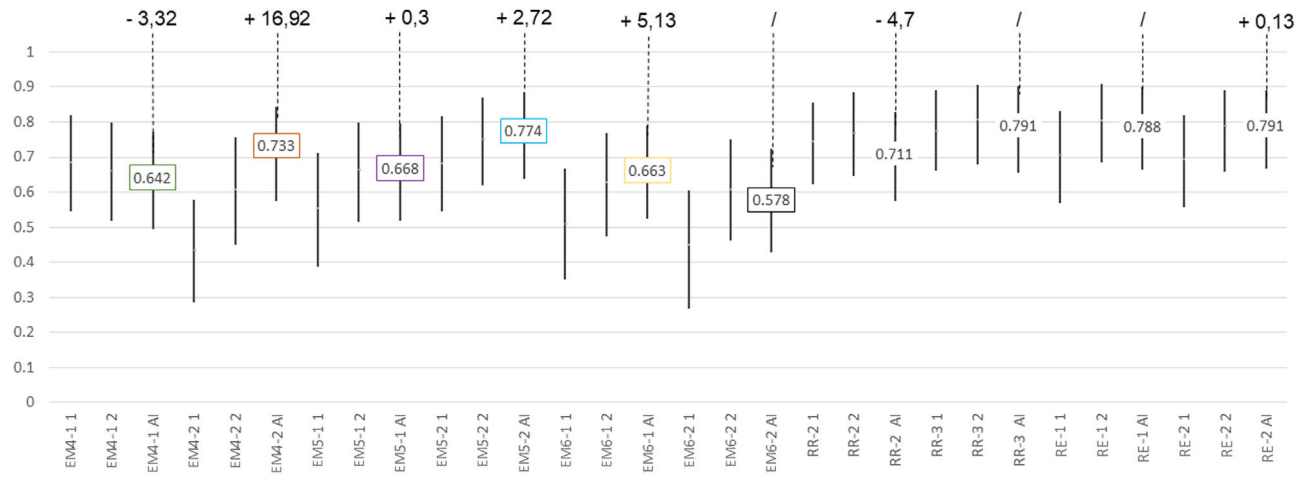


Figura 6 Concordancia para «probabilidad de neumonía COVID-19 con radiografía» entre cada lector y el estándar de referencia. Para cada evaluación, el coeficiente kappa corresponde al centro de la barra y sus intervalos de confianza del 95% a los límites de la barra.
Para la evaluación con inteligencia artificial (IA) se muestra el valor K en la etiqueta central, y el porcentaje de mejora (+) o empeoramiento (−) o ningún cambio (/) de la IA en comparación con el valor más alto de la primera y la segunda evaluación en la etiqueta superior; no se ha considerado ningún cambio cuando la IA generó un valor intermedio entre la primera y la segunda evaluación del lector.
IA: inteligencia artificial; EM4-1: estudiante de medicina de cuarto año número 1; EM4-2: estudiante de medicina de cuarto año número 2; EM5-1: estudiante de medicina de quinto año número 1; EM5-2: estudiante de medicina de quinto año número 2; EM6-1: estudiante de medicina de sexto año número 1; EM6-2: estudiante de medicina de sexto año número 2; RR2: residente de radiología de segundo año; RR3: residente de radiología de tercer año; RE1: radiólogo experto número 1; RE2: radiólogo experto número 2.

[30%]) y una proporción aún menor de concordancias moderadas (4/60 [6,66%]) o débiles (1/60 [1,66%]).

Concordancias con el estándar de referencia en la primera, segunda y tercera evaluaciones

Para «probabilidad de neumonía COVID-19» con radiografía (fig. 6) se obtuvieron 20 valores K tras agrupar el valor K de cada lector para la primera y la segunda evaluación. Como en el caso anterior, las concordancias de todos los lectores con el estándar de referencia en la primera y la segunda evaluaciones sin IA fueron predominantemente buenas (14/20 [70%]) frente a concordancias excelentes (2/20 [10%]) y moderadas (4/20 [20%]). Con la IA, las concordancias fueron buenas (9/10 [90%]) excepto una moderada (1/10 [10%]).

Respecto a la «probabilidad de afectación de cada zona pulmonar» (tablas suplementarias 1a-1f), se obtuvieron 120 valores K tras agrupar el valor K de cada lector para la primera y la segunda evaluaciones. En general, la concordancia con el estándar de referencia en la primera y la segunda evaluación sin utilizar IA en todos los participantes fue mayoritariamente buena (66/120 [55%]), con proporciones menores de concordancias moderadas (27/120 [22,5%]) y excelentes (25/120 [20,83%]), y proporciones anecdóticas de concordancias débiles (2/120 [1,66%]).

Con la IA, las concordancias con el estándar de referencia en todos los participantes fueron en su mayoría buenas (30/60 [50%]), con proporciones menores de concordancias excelentes (20/60 [33,3%]) y moderadas (7/60 [11,66%]), y proporciones anecdóticas de concordancias débiles (2/60 [3,33%]) y pobres (1/60 [1,66%]).

En cuanto al efecto en el grupo de estudiantes, la IA aumentó las concordancias con el estándar de referencia en un 50% (18/36), las redujo en un 19,44% (7/36) y fue indiferente para el 30,55% (11/36). En el grupo de residentes y radiólogos, la IA aumentó las concordancias en un 37,5% (9/24), las disminuyó en un 45,83% (11/24) y fue indiferente para el 16,66% (4/24; columna «Porcentaje de ±» en la tabla suplementaria 1).

La concordancia con el estándar de referencia para la «probabilidad de afectación de cada zona pulmonar» fue significativamente inferior para el grupo de estudiantes en comparación con los residentes y el grupo de radiólogos ($p < 0,001$). No se encontraron diferencias significativas en las demás comparaciones. La concordancia con el estándar de referencia para «probabilidad de neumonía COVID-19» mejoró para todos los estudiantes después de la evaluación de la IA en comparación con la primera y la segunda evaluaciones, excepto para los estudiantes 4-1 y 6-2 (fig. 6), sin diferencias significativas.

Rendimiento diagnóstico

Para «probabilidad de neumonía COVID-19» en la primera, segunda y tercera evaluación

Diferencias para cada lector (tabla 3):

- En los estudiantes sin emplear las imágenes procesadas con IA, el AUC osciló entre 0,737 y 0,881. Utilizando las imágenes procesadas con IA, el AUC osciló entre 0,775 y 0,890. En dos de los estudiantes (estudiante 4-2; estudiante 6-1) la IA mejoró el rendimiento diagnóstico de la

Tabla 3 Rendimiento diagnóstico de cada lector en su primera evaluación, segunda evaluación y evaluación conjunta con inteligencia artificial

	p			AUC (IC 95%)		
	1.ª-2.ª	1.ª-IA	2.ª-IA	1.ª	2.ª	RXT+IA
Estudiante de 4.°-1	0,2494	0,221	0,8884	0,833 (0,750-0,897)	0,804 (0,714-0,871)	0,801 (0,715-0,870)
Estudiante 4.°-2	0,0707	0,005	0,0742	0,744 (0,652-0,823)	0,822 (0,738-0,889)	0,866 (0,788-0,923)
Estudiante de 5.°-1	0,0904	0,1421	1	0,759 (0,669-0,834)	0,821 (0,737-0,886)	0,817 (0,733-0,884)
Estudiante de 5.°-2	0,2248	0,0961	0,4909	0,837 (0,756-0,900)	0,881 (0,807-0,934)	0,89 (0,817-0,941)
Estudiante de 6.°-1	0,1525	0,011	0,2048	0,75 (0,660-0,827)	0,802 (0,716-0,871)	0,842 (0,761-0,904)
Estudiante de 6.°-2	0,2103	0,3494	0,6646	0,737 (0,645-0,815)	0,784 (0,697-0,856)	0,775 (0,686-0,849)
Residente de 2.°	0,5938	0,2578	0,2212	0,887 (0,814-0,939)	0,868 (0,791-0,924)	0,844 (0,764-0,906)
Residente de 3.°	0,5791	0,9285	0,3173	0,894 (0,822-0,944)	0,907 (0,837-0,953)	0,896 (0,825-0,946)
Radiólogo 1	0,3925	0,7031	1	0,866 (0,789-0,923)	0,901 (0,831-0,949)	0,88 (0,806-0,934)
Radiólogo 2	0,5108	0,3144	0,3173	0,864 (0,786-0,921)	0,886 (0,812-0,938)	0,896 (0,825-0,946)

IA: inteligencia artificial; 1.ª-2.ª: valor p de la primera y segunda evaluación; 1.ª-IA: valor p de la primera evaluación y de la evaluación con la IA; 2.ª-IA: valor p de la segunda evaluación y de la evaluación con la IA; AUC: área bajo la curva; IC-95%: intervalo de confianza del 95%; 1.ª: AUC de la primera evaluación; 2.ª: AUC de la segunda evaluación; RXT+IA: AUC de evaluación con IA.

primera evaluación. No hubo diferencias significativas en el resto de estudiantes.

- En residentes y radiólogos, sin emplear las imágenes procesadas con IA, el AUC osciló entre 0,864 y 0,907. Utilizando las imágenes procesadas con IA, el AUC osciló entre 0,844 y 0,896. No hubo diferencias estadísticamente significativas entre los rendimientos diagnósticos de las tres evaluaciones para cada uno de los lectores.

Diferencias entre grupos de lectores:

En general, hubo una disminución del número de diferencias estadísticamente significativas (DESRD) entre el grupo de estudiantes y el de residentes y radiólogos en la primera evaluación (21 DESRD [66,66%]) y la segunda (17 DESRD [58,33%]), así como en la evaluación con IA (15 DESRD [45,83%], fig. 7).

Cambios para «probabilidad de afectación de cada zona pulmonar» entre la segunda y la tercera evaluación

Teniendo en cuenta los 6 estudiantes, 2 residentes y 2 radiólogos y las 6 zonas pulmonares de los 113 pacientes, la evaluación consistiría en 6.780 observaciones de afectación de zonas pulmonares. En 331 zonas los lectores no dieron ningún veredicto de probabilidad de afectación, por lo que se evaluaron 6.449 zonas pulmonares para este análisis.

Tras evaluar la imagen IA, se modificó la probabilidad de afectación de la zona pulmonar en 213/6449 (3,3%) zonas pulmonares. La evaluación de la imagen IA influyó en la evaluación zonal de las radiografías patológicas en mayor medida que en las normales. Dentro del grupo de pacientes con neumonía COVID-19, la evaluación de la imagen IA modificó la evaluación de las zonas pulmonares en las neumonías leves (puntuación Brixia inferior o igual a 9 puntos) con mayor frecuencia que en las neumonías graves. Esta influencia fue siempre mayor en el grupo de estudiantes (> 3% frente a < 1,5% de las zonas pulmonares en toda la muestra; > 4% frente a < 2% en las neumonías leves). Sin embargo, la interpretación de la imagen IA no siempre mejoró la concordancia con el estándar de referencia, sino que la empeoró en una proporción similar (tabla 4).

Estudio de las frecuencias de opinión en la lectura conjunta de la radiografía con IA

En cuanto a la contribución global de la IA, por término medio, la IA: a) resolvió dudas en un 31,30%, un 14,45% y un 10,05%; b) generó dudas en un 8,32%, un 5,70% y un 6,15%, y c) no aportó nada en un 59,22%, un 78,95% y un 82,90% para estudiantes, residentes y radiólogos, respectivamente (tabla 5).

Con respecto a «cambiar la primera impresión por la probabilidad de afectación de cada zona pulmonar», la opinión de los tres grupos de lectores fue que, cuando el software de IA influía en su confianza, resolvía sus dudas con más frecuencia de la que las generaba, especialmente en los estudiantes, seguidos de los residentes y, por último, de los radiólogos. La IA no modificó la primera impresión de los lec-

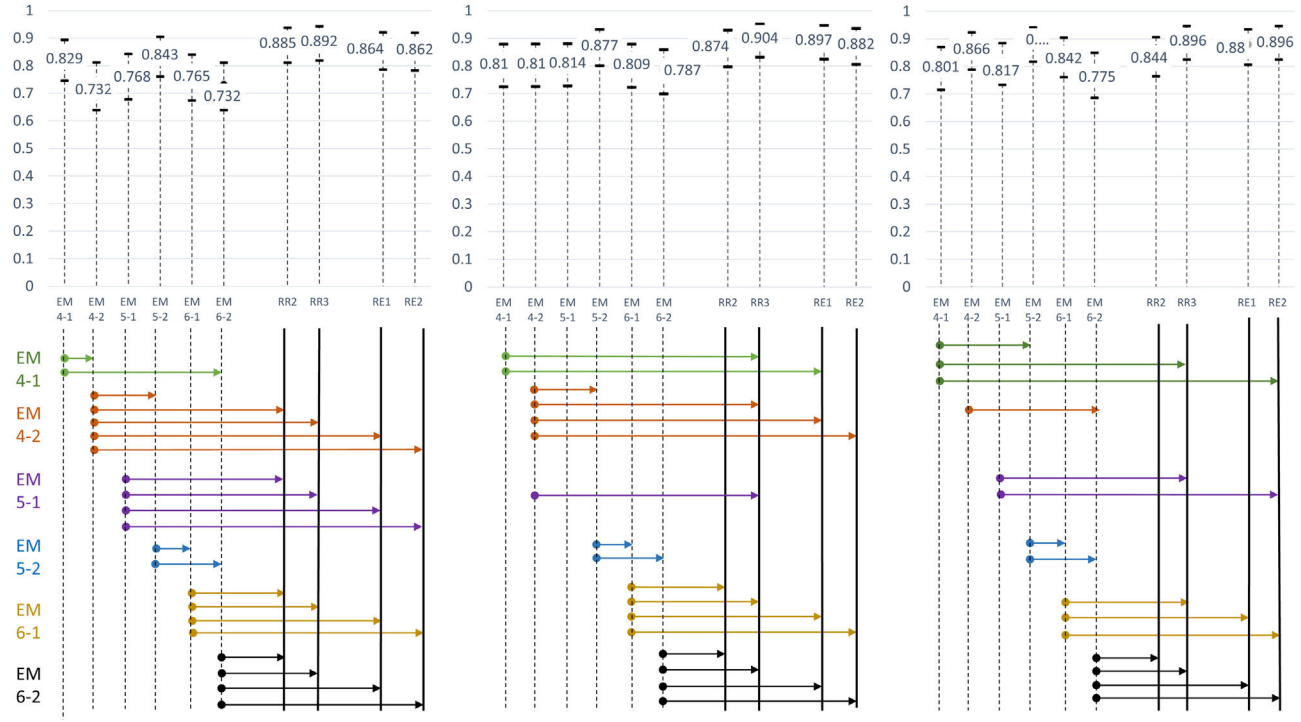


Figura 7 Comparación del rendimiento diagnóstico para la primera y la segunda evaluación, y para la evaluación con inteligencia artificial (IA).

Para cada evaluación, el área bajo la curva (AUC) corresponde al centro de la barra y sus intervalos de confianza del 95% a los límites de la barra. Las flechas representan las diferencias significativas según la prueba de DeLong de cada estudiante de medicina con el resto de los lectores. El número de diferencias significativas disminuye entre la primera y la segunda evaluación, y entre la segunda y la evaluación con IA. No hubo diferencias significativas entre radiólogos residentes, radiólogos expertos o entre residentes y expertos en las evaluaciones primera, segunda y con IA.

IA: inteligencia artificial; EM4-1: estudiante de medicina de cuarto año número 1; EM4-2: estudiante de medicina de cuarto año número 2; EM5-1: estudiante de medicina de quinto año número 1; EM5-2: estudiante de medicina de quinto año número 2; EM6-1: estudiante de medicina de sexto año número 1; EM6-2: estudiante de medicina de sexto año número 2; RR2: residente de radiología de segundo año; RR3: residente de radiología de tercer año; RE1: radiólogo experto número 1; RE2: radiólogo experto número 2.

Tabla 4 Cambios en la afectación de la zona pulmonar tras la interpretación de la imagen de IA, y efecto en la concordancia con el estándar de referencia. Se sumaron todas las zonas pulmonares. Los lectores se agruparon como estudiantes frente a residentes más radiólogos. Los datos se muestran en frecuencias absolutas y relativas

	Concordancia con el estándar de referencia tras la evaluación de la imagen IA	Radiografías normales		Radiografías patológicas		Radiografías patológicas Puntuación de Brixia $\leq$ 9		Radiografías patológicas Puntuación de Brixia $\leq$ 9	
Todos los lectores	Mejora	62/3,995	(1,55%)	58/2,454	(2,36%)	53/1,586	(3,34%)	5/868	(0,58%)
	Empeora	37/3,995	(0,93%)	56/2,454	(2,28%)	43/1,586	(2,71%)	13/868	(1,50%)
Estudiantes	Mejora	58/2,352	(2,47%)	45/1,433	(3,14%)	42/919	(4,57%)	3/514	(0,58%)
	Empeora	31/2,352	(1,32%)	47/1,433	(3,28%)	37/919	(4,03%)	10/514	(1,95%)
Residentes más radiólogos	Mejora	4/1,643	(0,24%)	13/1,021	(1,27%)	11/667	(1,65%)	2/354	(0,56%)
	Empeora	6/1,643	(0,37%)	9/1,021	(0,88%)	6/667	(0,90%)	3/354	(0,85%)

Tabla 5 Frecuencia en porcentajes de la opinión sobre «lo que la inteligencia artificial me ha aportado en general» en la interpretación de la radiografía de tórax de pacientes con sospecha de neumonía COVID-19

	?	D+	D−	Nada	Total
Estudiante 4-1	1,8	17,5	14,9	65,8	100
Estudiante 4-2	0,9	8,8	70,2	20,2	100
Estudiante 5-1	0,9	2,6	15,8	80,7	100
Estudiante 5-2	0,9	7	19,3	72,8	100
Estudiante 6-1	1,8	9,6	43,9	44,7	100
Estudiante 6-2	0,9	4,4	23,7	71,1	100
Media de estudiantes	1,2	8,32	31,30	59,22	100
Residente 2	0,9	9,60	25,40	64,00	100
Residente 3	0,9	1,80	3,50	93,90	100
Media de residentes	0,9	5,70	14,45	78,95	100
Radiólogo 1	1,8	10,50	14,00	73,70	100
Radiólogo 2	0	1,80	6,10	92,10	100
Media de radiólogos	0,9	6,15	10,05	82,90	100

IA: inteligencia artificial; ?: porcentaje de casos fallidos; D+: porcentaje de casos en los que la IA me ha planteado dudas; D−: porcentaje de casos en los que la IA ha resuelto mis dudas; Nada: porcentaje de casos en los que la IA no ha contribuido en absoluto.

tores en más del 89,03% de los casos (tablas suplementarias 2a-2f).

Discusión

Los principales resultados de este estudio pueden resumirse en los siguientes puntos: 1) En la mayoría de los casos los lectores afirman que la IA no cambió su opinión, pero resolvió más dudas de las que generó. 2) La evaluación de la IA facilitó a los estudiantes la valoración de la «probabilidad de neumonía COVID-19», mejorando la concordancia interobservador con el estándar de referencia para 4/6 estudiantes, y reduciendo el número de diferencias significativas en el rendimiento diagnóstico con el grupo de residentes y radiólogos. 3) El algoritmo de IA influyó en la interpretación de los lectores con mayor frecuencia en las radiografías con neumonías leves que en las graves o ausentes, especialmente en el caso de los estudiantes, aunque este cambio en la interpretación no siempre mejora la concordancia con el estándar de referencia. 4) Por último, aunque existen diferentes análisis que apoyan estas contribuciones, no se encontraron diferencias significativas en la mayoría de las comparaciones.

Nuestra hipótesis asumía el efecto beneficioso del software comercial de IA para todos los lectores, pero nuestros resultados solo apoyan algunos beneficios para los estudiantes de medicina. Creemos que las razones de las discrepancias con los estudios publicados[8-13] residen en el software de IA. En nuestro caso, a diferencia de los modelos de aprendizaje profundo de dichos estudios de investigación, que sí muestran un rendimiento prometedor de la IA por sí sola[11,12] o de la IA como apoyo de los radiólogos[8-10,13] frente a los radiólogos por sí solos, hemos utilizado un software de aplicación comercial no entrenado para establecer la probabilidad de COVID-19, sino para la probabilidad de lesiones pulmonares básicas, como nódulos o consolidaciones pulmonares. No obstante, el grupo de estudiantes sí se vio afectado: aunque consideraron que usar el programa informático no había cambiado su opinión en la mayoría de casos, en las ocasiones en las que sí les influyó les permitió resolver dudas en lugar de generarlas. Con su uso la mayoría de ellos aumentaron las concordancias con el estándar de referencia, dos de ellos aumentaron su rendimiento diagnóstico y disminuyeron el número de DESRD de estudiantes, residentes y radiólogos. Respecto a la influencia del algoritmo en función de la gravedad de la neumonía: en las radiografías normales (muchas de las cuales detectadas como normales por la imagen procesada con IA) el cambio en la interpretación por parte del lector será menor. En las neumonías graves, el mayor número de zonas afectadas y el mayor grado de afectación facilitan la interpretación del lector, por lo que es más probable que tampoco ayude disponer de una imagen procesada por IA. En las neumonías leves habrá una afectación más sutil y dudosa, por lo que la imagen procesada por IA puede tener más influencia en la interpretación, sobre todo en los lectores menos experimentados. En la actualidad, el software comercial de IA se ha entrenado con pequeñas parcelas de conocimiento. Como pone de relieve nuestro trabajo, estos programas informáticos pueden ser útiles en entornos específicos, pero siempre previa validación y sin dar por hecha su utilidad en todas las circunstancias, por ejemplo, tanto para descartar como para confirmar una enfermedad, o en cualquier entorno de trabajo, independientemente del nivel de experiencia del usuario. Este trabajo abre la puerta al desarrollo de herramientas aplicables a gran escala en la formación de usuarios con poca experiencia en la lectura de radiografías de tórax, como la amplia comunidad de estudiantes de medicina o médicos de familia. Sería conveniente crear equipos de trabajo formados por participantes de campos específicos para focalizar los esfuerzos y recursos para desarrollar herramientas de IA hacia soluciones prácticas específicas y útiles.

Nuestro estudio tiene algunas limitaciones. El tamaño final de la muestra fue reducido, y el número de lectores no fue lo suficientemente grande como para obtener resultados que nos permitieran disponer de datos sólidos de estos grupos. Dada la elevada concordancia intraobserva-

dor e interobservador y el elevado rendimiento diagnóstico mediante evaluación radiográfica sin IA de los estudiantes incluidos en este grupo, es lógico no encontrar diferencias estadísticamente significativas en la mayoría de las comparaciones. Es posible que estos estudiantes no representen a la población estudiantil de esta universidad, por lo que no se descarta que el programa informático pueda tener mejores resultados en lectores menos experimentados que los que ha mostrado este estudio. Por otro lado, habría sido óptimo incluir a otros especialistas, ya que en la sobrecarga de trabajo de los departamentos de radiología a menudo se encuentran sin apoyo en la lectura de las radiografías de tórax. La utilización de la tomosíntesis torácica como parte del estándar de referencia podría considerarse una limitación debido a su uso limitado y no probado. Aunque la tomosíntesis no es equivalente a la TC, se ha utilizado en este estudio debido a la mejor concordancia interobservador lograda para la neumonía COVID-19 en comparación con la radiografía[14]. Por último, la herramienta de IA utilizada no está entrenada para leer radiografías de tórax con neumonía COVID-19, pero sigue siendo una herramienta disponible, a diferencia de la mayoría de las experimentales.

En conclusión, esta herramienta comercial de IA no entrenada para COVID-19 no ha mostrado ningún impacto en la lectura de radiografías de tórax de pacientes con sospecha de neumonía COVID-19 en lectores expertos ni en aquellos con menor experiencia Emplear el programa de IA facilita interpretar la probabilidad de neumonía COVID-19 a lectores inexpertos y los casos con neumonía leve.

Financiación

Esta investigación no ha recibido apoyo específico de organismos del sector público, del sector comercial ni de organizaciones sin ánimo de lucro.

Autoría

1. Responsable de la integridad del estudio: JMPM, JMGS.
2. Concepción del estudio: JMPM, JMGS.
3. 3.Diseño del estudio: JMPM, JMGS.
4. Obtención de los datos: MPL, JMPM, MSC, CJP, RRM, LME, CGH, DGG, PHM, LCC, EGG, MNPM, SMR, JGM.
5. Análisis e interpretación de los datos: MPL, JMPM, JMGS.
6. Tratamiento estadístico: JMPM.
7. Búsqueda bibliográfica: JMPM, MPL.
8. Redacción del trabajo: MPL, JMPM, JMGS.
9. Revisión crítica del manuscrito con aportaciones intelectualmente.
10. relevantes: todos los autores.
11. Aprobación de la versión final: todos los autores.

Conflicto de intereses

Juana María Plasencia Martínez y José María García Santos han sido contratados por General Electric Healthcare Company para una investigación clínica destinada a explorar la utilidad en COVID-19 de la herramienta de inteligencia artificial Thoracic Care Suite para la radiografía de tórax. No se ha recibido ninguna otra financiación o apoyo de la empresa para llevar a cabo esta investigación.

El resto de los autores declaran no tener ningún conflicto de intereses.

Anexo. Material adicional

Se puede consultar material adicional a este artículo en su versión electrónica disponible en doi:10.1016/j.rx.2024.01.007.

Bibliografía

1. Wong HYF, Lam HYS, Fong AH, Leung ST, Chin TW, Lo CSY, et al. Frequency and distribution of chest radiographic findings in patients positive for COVID-19. Radiology. 2020;296:E72–8, http://dx.doi.org/10.1148/radiol.2020201160.
2. Mollura DJ, Culp MP, Lungren MP. Radiology in Global Health: Strategies, Implementation, and Applications. 2nd rev. Springer Cham; 2019. Crossref, Google Scholar.
3. Oren O, Gersh BJ, Bhatt DL. Artificial intelligence in medical imaging: Switching from radiographic pathological data to clinically meaningful endpoints. Lancet Digit Health. 2020;2:e486–8, http://dx.doi.org/10.1016/S2589-7500(20.30160-6.
4. Chassagnon G, Vakalopoulou M, Paragios N, Revel MP. Artificial intelligence applications for thoracic imaging. Eur J Radiol. 2020;123:108774, http://dx.doi.org/10.1016/j.ejrad.2019.108774.
5. The British Society of Thoracic Imaging. United Kingdom: The British Society of Thoracic Imaging; 2020 [actualizado 22 May 2020]. COVID-19 BSTI Reporting templates [about 4 screens]. Disponible en: https://www.bsti.org.uk/covid-19-resources/covid-19-bsti-reporting-templates/
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
7. Borghesi A, Maroldi R. COVID-19 outbreak in Italy: Experimental chest X-ray scoring system for quantifying and monitoring disease progression. Radiol Med. 2020;125:509–13, http://dx.doi.org/10.1007/s11547-020-01200-3.
8. Rangarajan K, Muku S, Garg AK, Gabra P, Shankar SH, Nischal N, et al. Artificial intelligence-assisted chest X-ray assessment scheme for COVID-19. Eur Radiol. 2021;31:6039–48, http://dx.doi.org/10.1007/s00330-020-07628-5.
9. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. Radiology. 2020;296:E156–65, http://dx.doi.org/10.1148/radiol.2020201491.
10. Yang Y, Lure FYM, Miao H, Zhang Z, Jaeger S, Liu J, et al. Using artificial intelligence to assist radiologists in distinguishing COVID-19 from other pulmonary infections. J Xray Sci Technol. 2021;29:1–17, http://dx.doi.org/10.3233/XST-200735.
11. Murphy K, Smits H, Knoops AJG, Korst MBJM, Samson T, Scholten ET, et al. COVID-19 on chest radiographs: A multireader evaluation of an artificial intelligence system. Radiology. 2020;296:E166–72, http://dx.doi.org/10.1148/radiol.2020201874.
12. Ghaderzadeh M, Aria M, Asadi F. X-ray equipped with artificial intelligence: Changing the COVID-19 diagnostic paradigm during the pandemic. Biomed Res Int. 2021;2021:9942873, http://dx.doi.org/10.1155/2021/9942873.
13. Li MD, Little BP, Alkasab TK, Mendoza DP, Succi MD, Shepard JO, et al. Multi-radiologist user study for artificial intelligence-guided grading of COVID-19 lung disease

severity on chest radiographs. Acad Radiol. 2021;28:572–6, http://dx.doi.org/10.1016/j.acra.2021.01.016.
14. Plasencia-Martínez JM, Moreno-Pastor A, Lozano-Ros M, Jiménez-Pulido C, Herves-Escobedo I, Pérez-Hernández G, et al. Digital tomosynthesis improves chest radiograph accuracy and reduces microbiological false negatives in COVID-19 diagnosis. Emerg Radiol. 2023;30:465–74, http://dx.doi.org/10.1007/s10140-023-02153-6.