Artículo de investigación

Concordancia interobservador de hallazgos cardiopulmonares en la radiografía de tórax entre radiólogos y médicos generales de un servicio de urgencias

Interobserver agreement of cardiothoracic diseases in chest radiography between radiologists and general practitioners

Felipe Aluja-Jaramillo[1] ✉, Martín Cañón-Muñoz[2], Rodolfo Mantilla-Espinosa[3] CvLAC, Héctor Mauricio Martínez-Orduz[3] CvLAC, Juan Mauricio Lozano-Barriga[4] CvLAC

Fecha correspondencia:
Recibido: agosto 21 de 2015.
Revisado: agosto 10 de 2016.
Aceptado: septiembre 8 de 2016.

Forma de citar
Aluja-Jaramillo F, Cañón-Muñoz M, Mantilla-Espinosa R, Martínez-Orduz HM, Lozano-Barriga JM. Concordancia interobservador de hallazgos cardiopulmonares en la radiografía de tórax entre radiólogos y médicos generales de un servicio de urgencias. Rev CES Med 2016; 30(2): 169-180.

Open access
© Copyright
Licencia creative commons
Ética de publicaciones
Revisión por pares
Gestión por Open Journal System
ISSN 0120-8705
e-ISSN 2215-9177

Resumen

Objetivo: estimar la concordancia inter-observador de hallazgos cardiopulmonares en la radiografía de tórax de adultos entre dos grupos independientes de radiólogos y médicos generales. **Materiales y métodos:** dos grupos de evaluadores, uno de radiólogos (n=2) y uno de médicos generales (n=5) valoraron 100 radiografías de tórax. Los ítems de evaluación fueron la calidad técnica radiológica, la normalidad de la radiografía y 27 hallazgos radiológicos comunes en la consulta de urgencias. Los evaluadores calificaron la posibilidad de encontrar los hallazgos específicos en la radiografía (cinco por cada placa) en un formato similar a un *script* de concordancia, con escala de respuesta tipo Likert. El cálculo de concordancia se realizó con el estadístico kappa por grupos de Vanbelle (κ_{2g}). **Resultados:** los grados de concordancia entre radiólogos y médicos generales fueron débiles para la identificación de hallazgos cardiopulmonares (κ_{2g} 0,46; IC 95 % 0,43 – 0,51), calidad de la imagen (κ_{2g} 0,44; IC 95 % 0,35 – 0,53) y determinación de normalidad (κ_{2g} 0,58; 0,44 – 0,72). Los índices de prevalencia fueron elevados (mín. – max.: 0,59 - 0,85) en la valoración de normalidad de las placas. **Conclusiones:** El grado de acuerdo en la determinación de normalidad puede estar subestimado por un alto índice de prevalencia. El poco tiempo de formación en radiología y de experiencia en el campo de los médicos generales podrían estar asociados al bajo grado de acuerdo entre los grupos.

Palabras clave: Radiografía, Tórax, Variaciones dependientes del observador, Radiología, Medicina general, Concordancia

Abstract

Objective: To estimate the inter-observer agreement between two independent groups of radiologists and general practitioners in the identification of cardiopulmonary findings via standard plain chest radiographs in adults. **Materials and Methods:** Two groups of independent raters (Radiologists, n=2; General Practitioners, n=5) analyzed 100 chest radiographs according to the technical quality, normality, and 5 specific findings. Cardiopulmonary

Sobre los autores:
1 Residente de Radiología e Imágenes Diagnósticas. Fundación Universitaria Sanitas.

2 Médico Familiar, Epidemiólogo Clínico. Universidad Icesi, Facultad de Ciencias de la Salud, Departamento de Salud Pública y Medicina Comunitaria, Cali, Colombia.

3 Médico Radiólogo. Departamento de Radiología, Clínica Universitaria Colombia. Profesor asociado de Radiología, Fundación Universitaria Sanitas.

4 Médico Radiólogo Intervencionista. Departamento de Radiología, Clínica Universitaria Colombia. Profesor asociado de Radiología, Fundación Universitaria Sanitas.

findings were registered via a script concordance-like test. We calculated agreement between groups with Vanbelle's kappa coefficient (κ_{2g}). **Results:** The concordance between the groups of radiologists and general practitioners in specific chest x-ray findings (κ_{2g} 0.46, 95 %CI 0.43 - 0.51), image technical quality (κ_{2g} 0.44; 95 %CI 0.35 – 0.53), and normality (κ_{2g} 0.58; 95 %CI 0.44 – 0.72) was weak. Prevalence indices were high in the analysis of chest x-ray normality (min. – max.: 0.59 – 0.85). **Conclusions:** Kappa coefficients in the determination of normality could have been biased downward due to high prevalence indices. Short time of training in radiology and experience in the field could account for low agreement between the groups.

Keywords: Radiography, Thorax, Reliability, Radiologist, General practice, Agreement

Introducción

La radiografía de tórax es uno de los métodos diagnósticos más solicitados en los servicios de urgencias (1) y continúa siendo un método vigente para el estudio de las enfermedades torácicas (2). Cambios sutiles en su interpretación pueden llevar a errores diagnósticos (3). Múltiples estudios realizan comparaciones entre la lectura de radiólogos contra varios tipos de especialistas del área clínica y encuentran que la concordancia es moderada o baja (4-7).

Varias razones explican la variabilidad inter-observador: formación o entrenamiento médico (8-10), la técnica de la imagen (11), tipo de imagen (digital o convencional) (10,12), historia clínica (2,4,11), falta de consenso sobre los criterios y definiciones (8,10) y acceso a imágenes previas (9,10).

Los médicos de atención primaria pueden cometer dos tipos de error en la lectura de radiografías: por *omisión*, aquellos en los que no se encuentran los hallazgos detectados por el radiólogo; por *inclusión*, se identifican imágenes no descritas por el radiólogo o que no existen (13). En la radiografía pueden visualizarse múltiples ítems y un solo error en su identificación pone en riesgo la validez del informe (14).

El objetivo de este estudio fue estimar la concordancia inter-observador de enfermedades cardiopulmonares en la radiografía de tórax de adultos entre dos grupos independientes de radiólogos y médicos generales.

Métodos

Estudio observacional de concordancia (consistencia) (15) en el servicio de urgencias de la *Clínica Universitaria Colombia* en Bogotá, Colombia, de cuarto nivel de complejidad, que incluyó 100 radiografías de tórax de pacientes con edades entre los 18 a 80 años, quienes consultaron al servicio de urgencias, con cualquier sintomatología, desde el 1 de abril de 2014. Las radiografías fueron aleatorizadas por uno de los investigadores previamente a la evaluación, con el programa Microsoft Excel.

Los evaluadores fueron cinco médicos generales del servicio de urgencias de la institución con al menos un año de experiencia, sin formación en programas de postgrado en áreas médico-quirúrgicas y dos médicos radiólogos con al menos 10 años de experiencia. Ambos grupos participaron voluntariamente y fueron cegados: no tuvieron acceso a la selección y aleatorización de las imágenes o participaron durante el análisis de los datos.

Las imágenes se organizaron en una presentación en Power Point® que se distribuyó por correo electrónico junto con un instrumento de recolección de datos.

Se valoró la calidad técnica radiológica y la determinación de normalidad de la imagen. La calidad técnica de la radiografía fue analizada de acuerdo a la cantidad de criterios que presentaba como: inspiración, penetración, rotación e inclusión de los ápices a las bases pulmonares. Las radiografías que cumplían con solo un criterio eran marcadas como malas; dos criterios, regulares; tres criterios, buenas, y cuatro criterios, excelentes.

Se valoró la calidad técnica radiológica y la determinación de normalidad de la imagen. La calidad técnica de la radiografía fue analizada de acuerdo a la cantidad de criterios que presentaba como: inspiración, penetración, rotación e inclusión de los ápices a las bases pulmonares.

Los datos de los hallazgos específicos se recogieron mediante un formato similar a un *script* de concordancia (SCT) (16). Se incluyeron cinco hallazgos, de 23 posibles, en cada radiografía, sin repetir las opciones, y tomados del glosario de términos de Fleishner Society (17).

A cada uno de los evaluadores se les preguntó si el hallazgo se encontraba o no en la radiografía con una escala tipo Likert de cinco opciones de respuesta: -2: el hallazgo no se encuentra en la imagen, -1: el hallazgo es poco probable que se encuentre en la imagen, 0: no se puede determinar si el hallazgo está o no en la imagen, +1: el hallazgo es probable que esté en la imagen y +2: el hallazgo se encuentra en la imagen.

Al final los evaluadores decidieron si la radiografía era normal o anormal. A su vez, se incluyó una casilla de comentarios para escribir hallazgos adicionales de consideración en las radiografías anormales, en cuyas opciones no apareciera dicho diagnóstico. Ninguno de los evaluadores utilizó este espacio (cuadro 1). Todos los evaluadores tuvieron 15 días para valorar las radiografías.

Cuadro 1. Ejemplo formato recolección de datos - script de concordancia

Característica	***Valoración***			
	Excelente	***Buena***	***Regular***	***Mala***
Calidad técnica de la radiografía				
Hallazgo radiográfico				
Cardiomegalia por crecimiento de cavidades derechas	-2	0	1	2
Calcificación pulmonar	-2	0	1	2
Opacidades intersticiales	-2	0	1	2
Consolidación	-2	0	1	2
Atelectasia	-2	0	1	2
¿Considera esta radiografía normal?	Si		No	

El tamaño de la muestra se calculó de acuerdo a las indicaciones de Donner (18) bajo los siguientes supuestos: clasificaciones positivas por radiólogo (radiografías anormales): 90 %; clasificaciones positivas para médicos generales: (radiografías anormales) 40 %; precisión de 10 %, y nivel de confianza de 95 %. La mínima muestra necesaria para cumplir con los supuestos fue de 99 radiografías. Se analizaron 100 imágenes.

Se realizó un análisis descriptivo de las variables demográficas con medias y frecuencias relativas y absolutas. El grado de acuerdo entre los grupos de especialistas en radiología y los médicos generales se realizó con el método descrito por Vanbelle (κ_{2g}) (16), con errores estándar tipo *jackknife*. La concordancia de la normalidad de las radiografías entre dos evaluadores se realizó con el coeficiente kappa de Cohen (19).

Para determinar la influencia de los efectos de prevalencia y sesgo se calcularon los índices respectivos, además de un coeficiente kappa ajustado (PABAK) (16). Se utilizó el estadístico de Fleiss para múltiples evaluadores (20) para determinar el acuerdo de normalidad dentro del grupo de médicos generales.

El tiempo de formación en Medicina fueron 12 semestres para tres médicos generales y 13 semestres para los dos restantes. Todos los médicos generales recibieron una formación en radiología en pregrado menor de 10 horas.

Cuando se trató del acuerdo de la calidad o los hallazgos específicos de la radiografía, entre dos evaluadores, se utilizó un coeficiente kappa de Cohen con ponderación lineal (κ_w) (19). El acuerdo dentro del grupo de médicos generales para estas dos variables se obtuvo con el alfa de Krippendorf (α_K) (21) para variables ordinales.

Los resultados de acuerdo para el estadístico kappa fueron clasificados así: 0 - 0,20 *inexistente*, 0,21 - 0,39 *mínimo*, 0,40 - 0,59 *débil*, 0,60 - 0,79 *moderado*, 0,80 - 0,90 fuerte y sobre 0,90 casi perfecto (21). Para el coeficiente alfa de Krippendorf se consideró un acuerdo inaceptable menor a 0,67; aceptable 0,67 a 0,80 y casi perfecto, mayor a 0,80.

Se obtuvieron intervalos de confianza estándar de aproximación normal de 95 % para los coeficientes kappa de Cohen y Fleiss. Para los coeficientes kappa por grupos, kappa de Cohen ponderado y alfa de Krippendorf se obtuvieron intervalos de confianza de 95 %, corregidos por sesgo, con 2000 iteraciones *bootstrap*.

Todos los análisis se realizaron con el programa estadístico R versión 3.0.2 (2013-09-25) (23) (paquetes "psy" (24), "psych" (25), "epicalc" (26) e "irr" (27)). El coeficiente kappa entre dos grupos y el error estándar se calcularon con los comandos kappa2g y jackvar, respectivamente.

El estudio fue aprobado por el Comité de Ética en Investigación de la Fundación Universitaria Sanitas (CEIFUS 1834–14).

Resultados

La experiencia promedio de los especialistas fue 14 años, de cinco para los médicos generales y 18 meses para los de servicios de urgencias. Sólo uno de los evaluadores fue de sexo femenino y pertenecía al grupo de médicos generales.

El tiempo de formación en Medicina fueron 12 semestres para tres médicos generales y 13 semestres para los dos restantes. Todos los médicos generales recibieron una formación en radiología en pregrado menor de 10 horas. Cuatro de estos médicos generales manifestaron interés en una formación futura en radiología.

La prevalencia de radiografías normales varió entre 7 y 22 %. Este intervalo lo delimitan los médicos generales 1 y 3, respectivamente. Los radiólogos obtuvieron valores similares de prevalencia (cuadro 2).

Cuadro 2. Porcentaje de normalidad y calidad de las radiografías de tórax por evaluador

	Normalidad (%)	***Calidad de la radiografía (%)***			
		Mala	***Regular***	***Buena***	***Excelente***
Radiólogo 1	16	1	22	75	2
Radiólogo 2	15	6	28	53	13
Médico general 1	7	11	25	44	20
Médico general 2	8	0	22	78	0
Médico general 3	22	10	25	57	8
Médico general 4	19	0	12	81	7
Médico general 5	17	6	26	53	15

La concordancia por grupos de la calidad de la imagen fue débil (κ_{2g} 0,44; error estándar -SE-: 0,047, IC 95 % 0,35–0,53). El mayor grado de concordancia alcanzado entre los evaluadores también fue débil. De los médicos generales, el valor más alto lo obtuvieron el 1 y 5 (κ_w 0,51; IC 95 % 0,39-0,63). Entre los médicos 2 y el 4 se obtuvo un grado de concordancia inexistente (κ_w 0,13; IC 95 % -0,02 - 0,33), el más bajo en este grupo de evaluadores. El menor grado de acuerdo fue entre el médico general 4 y el radiólogo 1 (κ_w 0,06; IC 95 % -0,09 - 0,22). El grado de acuerdo entre los radiólogos fue mínimo (κ_w 0,35; IC 95 % 0,22-0,48).

La concordancia para la determinación de normalidad fue débil entre los grupos de radiólogos y médicos generales (κ_{2c} 0,58; SE 0,071, IC 95 % 0,44–0,72). En la interpretación de normalidad de las radiografías, el acuerdo entre los radiólogos fue 0,66 (moderado). Entre los radiólogos y los médicos generales el máximo acuerdo fue 0,67 (moderado), y el mínimo 0,16 (inexistente). El mínimo y máximo entre médicos generales fueron 0,21 y 0,54, mínimo y débil, respectivamente (cuadro 3).

Cuadro 3. Concordancia ponderada interobservador para la calidad de la radiografía de tórax y los hallazgos específicos

	Calidad		***Hallazgos específicos***	
Evaluador	***Kappa $_w$***	***IC 95 %***	***Kappa $_w$***	***IC 95 %***
		Radiólogos		
1 vs. 2	0,35	0,219 - 0,484	0,46	0,391 - 0,571
		Radiólogo vs médico general		
1 vs. 1	0,32	0,219 - 0,431	0,39	0,337 - 0,451
1 vs. 2	0,42	0,205 - 0,610	0,37	0,309 - 0,435
1 vs. 3	0,28	0,147 - 0,428	0,37	0,294 - 0,444
1 vs. 4	0,06	-0,087 - 0,215	0,35	0,278 - 0,427
1 vs. 5	0,39	0,260 - 0,513	0,40	0,324 - 0,464
2 vs. 1	0,42	0,297 - 0,551	0,42	0,361 - 0,478
2 vs. 2	0,22	0,077 - 0,356	0,33	0,269 - 0,386
2 vs. 3	0,39	0,268 - 0,531	0,40	0,331 - 0,458
2 vs. 4	0,16	0,034 - 0,305	0,35	0,282 - 0,415
2 vs. 5	0,55	0,425 - 0,679	0,38	0,314 - 0,444
		Médicos generales		
1 vs. 2	0,22	0,113 - 0,330	0,32	0,258 - 0,379
1 vs. 3	0,40	0,286 - 0,522	0,38	0,321 - 0,440
1 vs. 4	0,17	0,059 – 0,293	0,32	0,258 - 0,379
1 vs. 5	0,51	0,389 - 0,632	0,37	0,312 - 0,428
2 vs. 3	0,25	0,108 - 0,379	0,37	0,305 - 0,435
2 vs. 4	0,13	-0,018 - 0,326	0,27	0,209 - 0,333
2 vs. 5	0,29	0,160 - 0,426	0,27	0,199 - 0,329
3 vs. 4	0,17	0,047 - 0,299	0,41	0,336 - 0,482
3 vs. 5	0,48	0,349 - 0,592	0,38	0,301 - 0,455
4 vs. 5	0,10	-0,006 - 0,243	0,42	0,343 - 0,494

Kappa w: Kappa con ponderación lineal

Al ajustar el coeficiente kappa por los índices de prevalencia (mín. – max.: 0,59 - 0,85) y sesgo (mín. – max.: 0,01 - 0,15) se encontraron grados más altos de acuerdo entre los evaluadores (cuadro 4).

La concordancia entre el grupo de radiólogos y el de médicos generales, para la identificación de hallazgos cardiopulmonares en la radiografía de tórax, fue débil (κ_{2g} 0,46; SE: 0,021, IC 95 % 0,43-0,51). El grado de acuerdo dentro de los grupos de evaluadores fue débil en el grupo de radiólogos (κ_w 0,46 IC 95 % 0,39-0,57) e inaceptable en los médicos generales (α_K 0,41; IC 95 % 0,36-0,46).

Cuadro 4. Grado de concordancia interobservador en la determinación de normalidad de las radiografías de tórax

Evaluador	*Índice de prevalencia*	*Índice de sesgo*	*PABAK*	*Kappa*	*IC 95 %*
		Radiólogos			
1 vs. 2	0,69	0,01	0,38	0,66	0,45 - 0,86
		Radiólogo vs médico general			
1 vs. 1	0,77	0,09	0,54	0,57	0,32 - 0,81
1 vs. 2	0,76	0,08	0,52	0,16	-0,08 - 0,40
1 vs. 3	0,62	0,06	0,24	0,61	0,42 - 0,81
1 vs. 4	0,65	0,03	0,30	0,55	0,33 - 0,77
1 vs. 5	0,67	0,01	0,34	0,67	0,48 - 0,87
2 vs. 1	0,78	0,08	0,56	0,50	0,24 - 0,76
2 vs. 2	0,77	0,07	0,54	0,37	0,10 - 0,64
2 vs. 3	0,63	0,07	0,26	0,57	0,37 - 0,78
2 vs. 4	0,66	0,04	0,32	0,43	0,20 - 0,67
2 vs. 5	0,68	0,02	0,36	0,55	0,33 - 0,78
		Médicos generales			
1 vs. 2	0,85	0,01	0,70	0,35	0,02 - 0,68
1 vs. 3	0,71	0,15	0,42	0,42	0,20 - 0,64
1 vs. 4	0,74	0,12	0,48	0,31	0,08 - 0,55
1 vs. 5	0,76	0,10	0,52	0,44	0,19 - 0,70
2 vs. 3	0,70	0,14	0,40	0,24	0,03 - 0,46
2 vs. 4	0,73	0,11	0,46	0,21	-0,02 - 0,44
2 vs. 5	0,75	0,09	0,50	0,24	-0.01 - 0,48
3 vs. 4	0,59	0,03	0,18	0,54	0,34 - 0,74
3 vs. 5	0,61	0,05	0,22	0,52	0,31 - 0,73
4 vs. 5	0,64	0,02	0,28	0,46	0,23 - 0,68

En este ítem el intervalo de acuerdo entre un médico general y un radiólogo estuvo entre mínimo y débil (κ_w 0,42-0,33). Los coeficientes de concordancia entre dos médicos generales no superaron un grado mayor al débil; lo obtuvieron los participantes 4 y 5 (κ_w 0,42; IC 95 % 0,35-0,45). El médico general No. 2 obtuvo los grados de acuerdo más bajos, clasificados como mínimos, con los médicos generales 4 y 5 (2 vs. 4: κ_w 0,27; IC 95 % 0,21-0,33; 2 vs. 5: κ_w 0,27; IC 95 % 0,20-0,33).

Los resultados indican que la concordancia para la identificación de hallazgos cardiopulmonares en la radiografía de tórax entre ambos grupos es débil, similar al obtenido por otras series en la literatura.

Discusión

Los resultados indican que la concordancia para la identificación de hallazgos cardiopulmonares en la radiografía de tórax entre ambos grupos es débil, similar al obtenido por otras series en la literatura. La concordancia interobservador obtenida en estudios similares, realizados con médicos especialistas de diferentes áreas, contra radiólogos, ha sido de baja a moderada (2,6,28), más aún cuando sólo se incluyen radiografías de tórax (13).

El grado de concordancia para los hallazgos específicos entre radiólogos fue débil Este resultado es más bajo que en otros que reportaron coeficientes de concordancia superiores a 0,80 (2,11,29). Este hallazgo es similar en el grupo de médicos generales.

El uso de opciones intermedias de la escala Likert ofrece la posibilidad de realizar un análisis más completo de cada imagen y así evaluar competencias individuales sin pretender encontrar una única respuesta correcta. La baja concordancia en la calidad de las imágenes pudo condicionar una baja concordancia en los hallazgos específicos en el grupo de radiólogos. Esto, sumado a que no se contó con radiólogos dedicados exclusivamente a la radiología de tórax, son factores que condicionan una baja concordancia (30).

La diferencia en la formación específica en el área de tórax de los radiólogos, su experiencia personal y el tiempo que dedicaron para la lectura de estas imágenes pudo tener un efecto sobre los valores de concordancia.

La heterogeneidad de los grados de concordancia para el grupo de médicos generales puede explicarse por el tiempo de formación (9,10,31) y su interés por la especialidad (32): un bajo tiempo de formación en radiología durante el pregrado tendría como consecuencia que haya muchos conceptos que puedan estar errados. La formación durante el pregrado a veces es insuficiente (33). Podría esperarse que estos valores estén relacionados con discrepancias en los conceptos entre los observadores, ya descrito por Markus *et al.* (8).

Aunque la literatura reporta que los errores por omisión son más frecuentes que los de inclusión, en este estudio ambos tipos de errores se presentaron en una distribución similar.

La prevalencia de normalidad tuvo en los extremos a los médicos generales y en la media de estos, a los radiólogos. Esto sugiere que la experiencia permite una mayor exactitud en la determinación de la normalidad o anormalidad. Los porcentajes más altos de normalidad se relacionan con errores por omisión (de hallazgos positivos que aumentan la normalidad de radiografías), mientras que los más bajos se relacionan con errores por inclusión (de hallazgos negativos que disminuyen el número de imágenes normales).

Aunque la literatura reporta que los errores por omisión son más frecuentes que los de inclusión (13), en este estudio ambos tipos de errores se presentaron en una distribución similar. Cualquiera que sea el tipo de error va a causar un aumento en los costos hospitalarios, por el aumento en las complicaciones o necesidad de estudios adicionales innecesarios.

Otra variable con un acuerdo débil fue la calidad técnica de la radiografía, que pudo haber ocurrido por la poca claridad en los criterios de calidad en el grupo de médicos generales. A medida que aumenta la experiencia, como en el grupo de radiólogos, la calidad es valorada de manera menos estricta. Venera *et al.* (33) informan que a mayor experiencia de los evaluadores, menor importancia le dan a la técnica radiológica.

La mayoría de valores se encontraron en las casillas *buena* y *regular*, con una baja cantidad de radiografías de excelente y mala calidad. Los radiólogos presentaron valores similares en cuanto a cantidad de radiografías de mala y excelente calidad, mientras que los médicos generales obtuvieron porcentajes muy variados. Es posible que estos no tengan claro el concepto de calidad técnica de la radiografía y su análisis sea muy subjetivo.

Una de las variables con menos concordancia fue la determinación de normalidad. Aquí se encontró un acuerdo débil entre los grupos. Así mismo, el grupo de radiólogos obtuvo un acuerdo moderado, menos de lo informado en estudios previos (29).

Los altos índices de prevalencia en este ítem pueden explicar la subestimación de los coeficientes kappa que se observan cuando se ajusta con el PABAK. Aun así, el efecto de la prevalencia en los valores de kappa sigue siendo materia de discusión (34-36).

Esta variabilidad también puede estar relacionada con el concepto de normalidad de cada evaluador. Algunos de los hallazgos específicos de las radiografías estaban asociados al envejecimiento o no asociados a procesos de enfermedad, por lo que cada evaluador realizó una interpretación global de los hallazgos para definir si era

normal o anormal. Aquí no encontramos evidencia empírica que soporte esta explicación, posiblemente debido a que se utilizaron radiografías con hallazgos que pudieron no considerarse patológicos, mientras que otros estudios utilizaron radiografías con hallazgos claramente patológicos como neuotórax o edema pulmonar (37).

El médico general No.4 tuvo la menor concordancia en las variables calidad y hallazgos específicos, con el grupo de radiólogos. Esto indica que si la concordancia en la calidad técnica es baja, lo más probable es que la concordancia de hallazgos específicos sea baja también. Esta correlación no es similar cuando el grado de acuerdo es alto en alguno de los factores. Por tanto, es importante que la evaluación incluya los tres factores (calidad, normalidad y hallazgos específicos) con un análisis independiente de los mismos.

Este estudio es de concordancia inter-observador, no de exactitud diagnóstica, por lo que no hubo un "patrón de oro" para verificar si las respuestas de cada grupo eran correctas.

La evaluación de las imágenes diagnósticas se realizó con un *script* de concordancia, herramienta útil para la medición de datos clínicos en casos poco claros o que puedan generar dudas clínicas (39). La herramienta es útil para la realización de este tipo de estudios porque el SCT permite que los participantes valoren e interpreten las imágenes con la posibilidad de análisis individual. Intuitivamente, aumenta la probabilidad de obtener un menor acuerdo por la presencia de respuestas intermedias. Sin embargo, la ponderación de los coeficientes de concordancia ajusta este tipo de error.

Otra fortaleza es la aplicación de técnicas estadísticas novedosas para encontrar el acuerdo entre grupos. El método de Vanbelle permitió la estimación del grado de acuerdo entre dos grupos independientes y el cálculo robusto de errores estándar, y fue útil para comparar grupos independientes con la heterogeneidad propia de cada uno y determinar el acuerdo global.

Este estudio es de concordancia inter-observador, no de exactitud diagnóstica, por lo que no hubo un "patrón de oro" para verificar si las respuestas de cada grupo eran correctas.

La falta de la proyección lateral de las radiografías es una de las limitaciones más importantes. Su uso permite una identificación más clara del hallazgo radiológico encontrado en la proyección posteroanterior (2). Esto puede ser aún más importante en el análisis de la concordancia entre el grupo de radiólogos, quienes están acostumbrados a contar con las dos proyecciones. En la práctica común, la radiografía lateral no siempre es solicitada por el médico general, por lo que, al no tenerla para lectura, probablemente no vio afectada su interpretación.

Otra de las limitaciones fue la falta de información clínica. Esto pudo haber contribuido a la baja concordancia entre grupos. Sin embargo, excluir los datos clínicos nos permitió evaluar los hallazgos semiológicos de la radiografía y no su correlación con un dato clínico que sesgaría los hallazgos radiológicos.

El uso de pantallas de baja resolución fue otra limitante para ambos grupos, aunque esto se planificó con la intención de volverlos equitativos para que ninguno de evaluadores tuviera ventajas. Este factor pudo influir sobre todo en el grupo de radiólogos quienes usualmente valoran las radiografías en pantallas de alta resolución.

Ambos grupos sabían que estaban siendo evaluados, lo que podría incrementar su desempeño. Sin embargo, el acuerdo fue débil, aún con tiempo para consultar. Dado este grado de acuerdo, la probabilidad de que hubiesen consultado fue baja.

Ninguno de los radiólogos que valoraron las radiografías se dedicaba exclusivamente a la radiología de tórax, motivo por el cual puede haber mayor discrepancia en los conceptos.

Es conveniente que se implementen estrategias para mejorar la calidad de la educación en radiología en el pregrado. La radiología es una especialidad que ha evolucionado rápidamente y los programas de pregrado no se han adaptado a estos cambios (32). Sería importante implementar un programa de reentrenamiento en radiología de tórax para los radiólogos que no se dedican exclusivamente a este campo.

La implementación del SCT como técnica de evaluación de competencias de personal de salud en formación podría ser útil en el área de imágenes diagnósticas y en otras donde surjan dudas clínicas y se requiera comparar contra lo que haría un médico experimentado, en una misma situación.

Es conveniente que se implementen estrategias para mejorar la calidad de la educación en radiología en el pregrado. La radiología es una especialidad que ha evolucionado rápidamente y los programas de pregrado no se han adaptado a estos cambios.

Conclusiones

El grado de acuerdo entre los dos grupos evaluados fue débil. El poco tiempo de formación en radiología y la poca experiencia en el campo podrían explicar estos resultados. Es recomendable la lectura por un radiólogo para disminuir los errores de interpretación, ya sean de omisión o inclusión. Esto mejoraría la aproximación diagnóstica de los pacientes.

Agradecimientos

A una persona anónima por su valiosa orientación en el cálculo de los coeficientes kappa por grupos.

Conflicto de interés

Los autores declaran no tener conflicto de intereses.

Financiación

Fue asumida por los autores.

Bibliografía

1. Preston CA, Marr J, Amaraneni KK, Suthar BS. Reduction of "Callbacks" to the ED due to discrepancies in plain radiograph interpretation. Am J Emerg Med. 1998; 16(2):160 – 162. http://dx.doi.org/10.1016/S0735-6757(98)90036-5.

2. Simó Miñana J, Riquelme Miralles DA. Variabilidad en la interpretación de la radiografía de tórax entre una comunidad médica de atención primaria y sus radiólogos de referencia. Aten Primaria. 1998; 21:599 – 606. http://www.elsevier.es/es-revista-atencion-primaria-27-articulo-variabilidad-interpretacion-radiografia-torax-entre-15129

3. Cherian T, Mulholland EK, Carlin JB, Ostensen H, Amin R, de Campo M, et al. Standarized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. Bulletin of the World Health Organization. 2005;83:353 – 359. http://dx.doi.org/10.1590/S0042-96862005000500011

4. Fleisher G, Ludwig S, McSorley M. Interpretation of Pediatric X-Ray Films by Emergency Department Pediatricians. Ann Emerg Med. 1983;12(3):153 – 158. http://dx.doi.org/10.1016/S0196-0644(83)80557-5

5. Formento Tirado JA, Domínguez Gabas JL, Arenas Abad A, Lorente Aznar T, Vásquez Pueyor R, Isanta Pomar C. Grado de acuerdo en la interpretación radiológica de crecimiento de cavidades cardíacas izquierdas entre radiólogo, médico de familia y residentes de MFYC. Atención Primaria. 1993;11(5): 243 – 245.

6. Gatt ME, Spectre G, Paltiel O, Hiller N, Stalnikowicz R. Chest radiographs in the emergency department. Is the radiologist really necessary? Postgrad Med J. 2003;79: 214 – 217. doi:10.1136/pmj.79.930.214

7. Nesterova GV, Leftridge CA, Natarajan AR, Appel HJ, Bautista MV, Hauser GJ. Discordance in interpretation of chest radiographs between pediatric intensivists and a radiologist: Impact on patient management. J Crit Care. 2010; 25:179 – 183. doi:10.1016/j.jcrc.2009.05.016

8. Markus JB, Somers S, Franic SE, Moola C, Stevenson GW. Intraobserver Variation in the Interpretation of Abdominal Radiographs. Radiology. 1989; 171:69 – 71. DOI: http://dx.doi.org/10.1148/radiology.171.1.2928547

9. Potchen EJ, Cooper TG, Sierra AE, Aben GR, Potchen MJ, Potter MG, et al. Measuring Performance in Chest Radiography. Radiology. 2000; 217: 456 – 459. doi:10.1148/radiology.217.2.r00nv14456.

10. Robinson PJA. Radiologist`s achilles` heel: error and variation in the interpretation of the Röntgen image. Br J Radiol. 1997; 70: 1085 – 1098. http://dx.doi.org/10.1259/bjr.70.839.9536897

11. Tudor GR, Finlay D, Taub N. An assessment of inter-observer agreement and acurracy when reporting plain radiographs. Clin Radiol. 1997; 52:235 – 238. DOI: 10.1016/S0009-9260(97)80280-2

12. Eng J, Mysko WK, Weller GER, Renard R, Gitlin JN, Bluemke DA et al. Interpretation of emergency department radiographs: A Comparison of emergency medicine physicians with radiologists, residents with faculty, and film with digital display. AJR Am J Roentgenol. 2000; 175:1233 – 1238. https://www.ncbi.nlm.nih.gov/pubmed/11044013

13. Kuritzky L, Haddy RI, Curry RW. Interpretation of chest roentgenograms in primary care physicians. South Med J. 1987; 80(11):1347 – 1351. https://www.ncbi.nlm.nih.gov/pubmed/3686134

14. Herman PG, Gerson DE, Hessel SJ, Mayer BS, Watnick M, Blesser B, Ozonoff D. Disagreements in chest roentgen interpretation. Chest. 1975; 68(3): 278 – 282. DOI: 10.1378/chest.68.3.278

15. Kramer MS, Feinstein AR. The biostatistics of concordance. Clin Pharmacol Ther. 1981; 29(1):111-123. DOI: 10.1038/clpt.1981.18

16. Vanbelle S, Albert A. Agreement between two independent group of raters. Psychometrika. 2009;74(3):477-491. DOI:10.1007/s11336-009-9116-1

17. Hansell DM, Bankier AA, MacMahon H, McLoud T, Müller NL, Remy J. Fleischner Society: Glossary of Terms of Thoracic Imaging. Radiology. 2008; 246(3):697 – 722. DOI: 10.1148/radiol.2462070712

18. Donner A, Rotondi MA. Sample size requirements for interval estimation of the Kappa Statistic for interobserver agreement studies with a binary outcome and multiple raters. Int J Biostat. 2010;6(1):Article31. DOI: 10.2202/1557-4679.1275,

19. Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. PhysTher. 2005; 85: 257 – 268. http://ptjournal.apta.org/content/85/3/257

20. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76(5):378-382. http://dx.doi.org/10.1037/h0031619

21. Krippendorff, K. Computing Krippendorff 's Alpha-Reliability. Retrieved from http://repository.upenn.edu/asc_papers/43

22. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica. 2012; 22(3):276-282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

23. R Core Team 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

24. Bruno Falissard. Various procedures used in psychometry. R package version 1.1. Available at: http://CRAN.R-project.org/package=psy

25. Revelle, W. Procedures for personality and psychological research, Northwestern University, Evanston, Illinois, USA, Available at: http://CRAN.Rproject.org/package=psych Version = 1.4.5.

26. Chongsuvivatwong V. Epicalc: Epidemiological calculator. R package version 2.15.1.0. Available at https://CRAN.R-project.org/package=irr

27. Gamer M, Lemon J, Fellows Puspendra. Gamer M, Lemon J, Fellows I, Singh P (2013) IRR: Various coefficients of interrater reliability and agreement. R package version 0.84. CRAN: http://www.r-project.org

28. Various coefficients of interrater reliability and agreement. R package version 0.84. Available at: http://CRAN.R-project.org/package=irr

29. Novak V, Avnon LS, Smolyakov A, Barnea R, Jotkowitz A, Schlaeffer F. Disagreement in the interpretation of chest radiographs among specialists and clinical outcomes of patients hospitalized with suspected pneumonia. Eur J Intern Med. 2006; 17:43 – 47. http://dx.doi.org/10.1016/j.ejim.2005.07.008

30. Campbell SG, Murray DD, Hawass A, Urquhart D, Ackroyd-Stolarz S, Maxwell D. Agreement between emergency physician diagnosis and radiologist reports in patients discharged from an emergency department with community-acquired pneumonia. Emerg Radiol. 2005; 11: 242 – 246. DOI: 10.1007/s10140-005-0413-4.

31. Rhea JT, Potsaid MS, DeLuca SA. Errors of interpretation as elicited by a quality audit of an emergency radiology facility. Radiology. 1979; 132:277-280. http://dx.doi.org/10.1148/132.2.277

32. Jefferey DR, Goodard PR, Callaway MP, Greenwood R. Chest Radiograph Interpretation by Medical Students. Clin Radiol. 2003; 58:478 – 481. http://dx.doi.org/10.1016/S0009-9260(03)00113-2.

33. Del Cura Rodríguez JL, Martínez Noguera A, Sendra Portero F, Rodríguez González R, Alguersuari Cabiscol A. La enseñanza de la Radiología en los estudios de la licenciatura de Medicina en España. Informe de la Comisión de Formación de la SERAM. Radiología. 2008; 50: 177 – 182. doi: 10.1016/S0033-8338(08)71963-5

34. Venera A, Rincón DA, Torres LI, Arango M. Concordancia interobservador en los hallazgos de radiografía de tórax pediátrica. Rev Fac Med Univ Nac Colomb. 2004; 52(3):192 – 198. http://www.bdigital.unal.edu.co/39113/1/43420-201827-1-PB.pdf

35. Vach W. The dependence of Cohen's kappa on the prevalence does not matter. J Clin Epidemiol. 2005; 58(7):655-61. doi:10.1016/j.jclinepi.2004.02.021

36. Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. J Clin Epidemiol. 2000; 53(5):499-503. http://dx.doi.org/10.1016/S0895-4356(99)00174-2

37. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993; 46(5):423-9. http://dx.doi.org/10.1016/0895-4356(93)90018-V

38. Kaufman B, Dhar P, O´Neill DK, Leitman B, Fermon CM, Wahlander SB, Sutin KM. Chest radiograph interpretation skills of Anesthesiologists. J Cardiothorac Vasc Anesth. 2001; 15(6):680 – 683. DOI: http://dx.doi.org/10.1053/jcan.2001.28307

39. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. Medical Education. 2012; 46:552-563. DOI: 10.1111/j.1365-2923.2011.04211