

How does ChatGPT-4 match radiologists in detecting pulmonary congestion on chest X-ray?

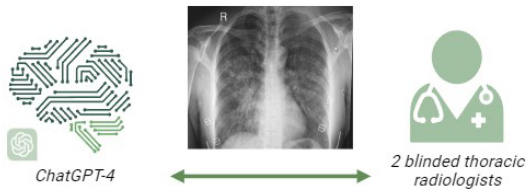
ChatGPT-4 (OpenAI) is a popular language model with the possibility of radiology imaging analysis.

Can ChatGPT-4 accurately provide physicians an estimate of pulmonary congestion on chest X-ray?

Population: 50 selected patients with acute dyspnea, whom all had a chest X-ray performed within 4 hours from admission.

ChatGPT-4 was asked to evaluate the likelihood of pulmonary congestion on a 5-point Likert scale, where 1=unlikely and 5=definite.

First, ChatGPT-4 was provided **age and sex** of the patient, and later also medical history, clinical examination, vital parameters, and ECG.



Match: a discrepancy of ≤ 1 points between the average assessments of the radiologists and ChatGPT-4.
Mismatch: a discrepancy of >1 point.

Conclusion:

ChatGPT-4 could match radiologist assessment of pulmonary congestion in only 54% of the chest X-rays. ChatGPT-4 was wrong in 46% of the assessments, and physicians should not use ChatGPT as a radiologist stand-in.

	ChatGPT -4	Radiologist 1	Radiologist 2	Match	Match after clinical information
1	1	1	1	0	-2
2	1	1	1	0	0
3	1	1	1	0	0
4	4	4	4	0	0
5	1	1	1	0	0
6	1	1	1	0	0
7	1	1	1	0	0
8	1	1	1	0	-1
9	1	1	1	0	-1
10	5	5	5	0	1
11	1	1	1	0	-1
12	1	1	1	0	-1
13	4	5	4	0,5	0,5
14	4	5	4	0,5	-0,5
15	4	5	4	0,5	2,5
16	4	5	4	0,5	1,5
17	4	5	4	0,5	0,5
18	4	5	5	1	0
19	4	5	5	1	0
20	4	5	5	1	2
21	2	1	1	-1	-1
22	3	4	4	1	-1
23	2	1	1	-1	-2
24	2	1	1	-1	0
25	4	5	5	1	3
26	2	1	1	-1	-1
27	2	1	1	-1	-2
28	3	5	4	1,5	0,5
29	3	5	4	1,5	0,5
30	3	1	1	-2	-3
31	3	1	1	-2	-3
32	4	1	3	-2	-2
33	3	1	1	-2	-1
34	3	1	1	-2	-1
35	3	5	5	2	0
36	2	4	4	2	0
37	2	5	4	2,5	0,5
38	2	5	5	3	3
39	4	1	1	-3	-1
40	2	5	5	3	3
41	4	1	1	-3	-1
42	4	1	1	-3	-2
43	4	1	1	-3	-2
44	4	1	1	-3	-2
45	1	4	4	3	2
46	2	5	5	3	2
47	2	5	5	3	2
48	1	5	4	3,5	0,5
49	1	5	5	4	3
50	1	5	5	4	2

Figure S1 Visual abstract. Created with Biorender.com. ECG, electrocardiographic.