Supplementary 2: The sensitivity and specificity of two readers without and with deep learning assistance in the detection of CT rib fractures

In the main manuscript, we used the sensitivity and false-positive per scan (FPS) as the measures to evaluate the performance of the rib fracture detection. In this supplementary, we provide an analysis of the specificity instead of the FPS.

We calculated the sensitivity and specificity of 102 patients with rib fractures. To calculate the specificity, the rib fractures were binned into the segments. As the schematic drawing showed in Fig.2 in the main manuscript, the 2nd to 10th ribs were divided into the anterior, lateral, and posterior segments using two ancillary lines at the level of the anterior axillary line and the inferior scapula point. The 1st, 11th, and 12th rib were not divided, because of their short length. Therefore, 30 segments on the left and right, and 60 segments in total in each patient were obtained. On a segment, if one or more fractures were identified, the segment was annotated as 'fractured'. Otherwise, the segment was annotated as 'intact'. The sensitivity and specificity were calculated by:

$$\text{Sensitivity} = \frac{\text{Detected true fractured}}{\text{All true fractured segments}}$$

$$\text{Specificity} = \frac{\text{Detected true intact}}{\text{All true intact segments}}$$

In 6120 segments of 102 patients, there were 829 fractured and 5291 intact segments. It was shown in Table S2.1 that reader 1 (R1) found 693, 734, and 732 true fractured segments in the reading session 1, 2, and 3 (S1, S2, and S3) resulting in a sensitivity of 83.6%, 88.5%, and 88.3%. The improvement of R1's sensitivity was significant in S2 and S3 compared to S1 (both $p < 0.05$) using a pairwise Chi-square test with Holm correction, but there was no difference between S2 and S3 ($p = 0.94$). Similar to the result of R1, the sensitivity of reader 2 (R2) was also significantly improved in S2 and S3 compared to S1 (both $p < 0.05$). The specificity of R1 and R2 was 99.3% - 99.7% in S1, S2, and S3. The difference of specificity between S1, S2, and S3 was not significant in both R1 and R2 ($p = 0.31$ and 0.47).

The deep learning (DL) technique identified 658 true fractured segments and 5260 true intact segments, which translates to a sensitivity of 79.4% and a specificity of 99.4%. The sensitivity of DL was lower than R1 or R2 in S1 (both $p < 0.05$). The sensitivity between R1 and R2 in S1 was not significantly different ($p = 0.89$). The specificity of DL was not significantly different from R1 or R2 in S1 ($p = 0.52$).

The precision, recall, and F1-score were often used as performance measures for the deep learning classification algorithm. They are calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where TP, FP, and FN are the count for true positive fractures, false positives, and false negatives.

The precision, recall, and F1-score of DL were 95.5%, 79.4%, and 0.867.

In conclusion, the sensitivity of CT rib fracture detection was improved with DL's assistance. The specificity of the human readers and DL was excellent (>99%), and no significant difference was found between three reading sessions.

Table S2.1. The sensitivity and specificity of R1 and R2 in S1, S2, and S3.

	S1	S2	S3
R1			
Sensitivity	83.6% (693/829)	88.5% (734/829)	88.3% (732/829)
Specificity	99.5% (5264/5291)	99.3% (5256/5291)	99.3% (5255/5291)
R2			
Sensitivity	84.0% (696/829)	88.4% (733/829)	89.3% (740/829)
Specificity	99.3% (5255/5291)	99.5% (5267/5291)	99.7% (5275/5291)

* The sensitivity was presented as percentage (detected /all true fractured segment).

* The specificity was presented as percentage (detect / all true intact segment).