Performance and Adjustment of an AI-Assisted Fracture Detection Tool in a Real-World Post-Clinical Deployment Setting

Sameed Khan, MD, Researcher, Cleveland Clinic Lerner Research Institute
Sarang Ingole, MBBS, DMRD, DNB, Charit Tippareddy, MD; Amy Zhou; Kacey Pagano; Mia Zivkovic; Orlando Martinez; Navid Faraji, MD

Introduction

AI fracture detection tools high performance in research settings is established but their effectiveness after deployment in clinical practice is understudied. Post-deployment analysis offers the opportunity to study the tool before and after on-the-fly modifications to operational parameters based on observed performance "in the wild."

Hypothesis

Fracture detection tool sensitivity and specificity will be similar to levels assessed in preclinical studies (86.5% and 82.6% respectively). Corrective modifications will improve performance.

Methods

This study analyzed 2378 appendicular trauma studies referred to a tertiary ED following deployment of the fracture detection tool. After corrective modification to increase threshold points A (low-suspicion detection) and B (high-suspicion detection), a further 2023 patients were analyzed. All radiograph reports were reviewed by two board-certified radiologists. Performance was evaluated before and after modifications to the AI tool. Significance testing was performed via chi-square test for proportions.

Results

Initially, the AI tool achieved a sensitivity of 89.5%, specificity of 76.0%, accuracy of 79.6%, and a negative predictive value (NPV) of 95.0%. The precision was 0.58, and the F1 score was 0.71. Notably, 93.2% of false positives came from the low-suspicion group. After refining the tool, sensitivity increased significantly to 94% (p=0.008), specificity to 87% (p=1.9 x 10-15), and accuracy to 88.7% (p=3.4 x 10-16), with NPV of 97.7% (p=0.004). Precision improved significantly to 0.71 (p=8.72 x 10-8), and the F1 score to 0.81. Common false negatives involved fractures near complex joints or in cases of severe osteoporosis, while false positives were often associated with misidentified sesamoid bones, artifacts, and external hardware.

Conclusion

Initially, the AI tool demonstrated high sensitivity but a high rate of false positives. Refining the tool to adjust thresholds led to improved sensitivity, specificity, and overall performance, highlighting the importance of ongoing AI tool refinement for clinical deployment.

Figure(s)

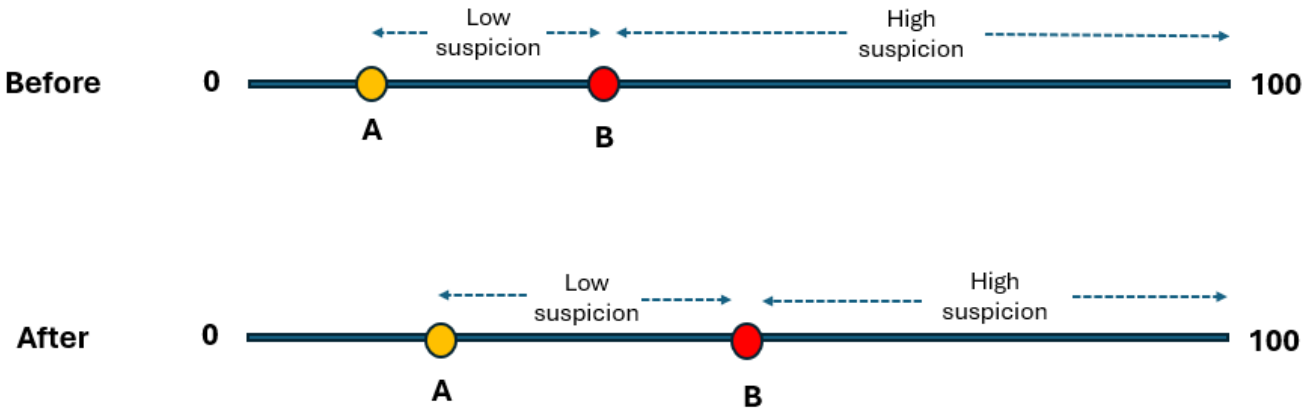

Figure 1. Diagramatic representation of the adjustment of threshold values

	Before	After
Total Cases	2378	2023
True Positive	577	486
False Positive	416	197
True Negative	1317	1309
False Negative	68	31
Sensitivity	89.5	94
Specificity	76	87
Precision	0.58	0.71
Negative Predictive Value	95	97.7
Accuracy	79.6	88.7
F1 Score	0.71	0.81

Table 1. Performance of AI fracture detection tool before and after adjustment of thresholds

Keywords

Applications; Artificial Intelligence/Machine Learning; Clinical Workflow & Productivity; Imaging Research; Quality Improvement & Quality Assurance