Post-Market Surveillance of Artificial Intelligence (AI) Fracture Detection in Live Clinical Operational Use in the UK

Dr. Amanjeet Braich, Dr. Daniel Fascia, Dr. Gui Tran
Harrogate and District NHS Foundation Trust

For queries, please contact
amanjeet.braich1@nhs.net

PURPOSE

Artificial Intelligence (AI) fracture detection was implemented into live operational use for plain film radiography in a district general hospital.

The ongoing performance of this software was evaluated to determine concordance between AI fracture detection algorithms and human reports.

METHODS AND MATERIALS

All musculoskeletal x-rays were processed by AI over an 18-month period. 1000 sampled cases from August 2024 were analysed to assess rate of concordance and discrepancy between the algorithm and human reports.

For each radiograph, the age, procedure name, number of AI-detected fractures, number of human-reported fractures and concordance was recorded.

RESULTS

Statistical analysis generated rates of: complete agreement, false positives and false negatives. Partial agreement was also measured, representing cases with multiple findings containing at least one discrepancy.

- Complete Agreement 91.1% (n=911)
- False Positive 4.6% (n=46)
- Partial Agreement 2.3% (n=23)
- False Negative 2% (n=20)

Incidence of false positives (FP) was highest for foot, hip and ankle imaging.
Incidence of false negatives (FN) was highest for wrist and pelvic imaging.

Foot	Hip	Ankle	Fingers	Wrist	Pelvis	Knee
FP 0.7% / FN 0.1%	FP 0.7% / FN 0.2%	FP 0.7% / FN 0.1%	FP 0.5% / FN 0.3%	FP 0.3% / FN 0.4%	FP 0.2% / FN 0.4%	FP 0.2% / FN 0.2%

Paediatric imaging was most likely to result in false positives at 6.3% incidence, whilst elderly imaging (age>75) was more likely to result in partial agreement and false negatives (3.2% and 2.5% respectively).

Ankle and wrist imaging had increased incidence of partial agreement, mostly due to over-detection of fractures.

Combination of Concordance	True Fracture(s) and False Positive(s)	False Positive(s) and False Negative(s)	True Fracture(s) and False Negative(s)	True Fracture(s) and False Positive(s) and False Negative(s)
% of 'Partial Agreement' Cases	82.6% (n=19)	8.7% (n=2)	4.3% (n=1)	4.3% (n=1)

Table 1. Percentage of 'Partial' Agreement Cases attributable to each combination of concordance.

The algorithm performed well in cases of multiple true fractures, with a 93.9% rate of complete agreement, however had poorer performance in cases which included four or more fractures.

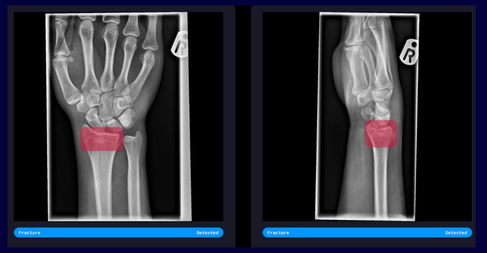


Figure 1. Adult distal radial fracture detected by Radiobotics RBfracture™

CONCLUSION

AI-fracture detection had <10% rate of overall disagreement with human reports, a false positive rate of 4.6% and false negative rate of 2%.

Weakest performance for AI occurred for paediatric and elderly cases, alongside cases with four or more true fractures.