FRACT-AI Research Protocol

FRACT-AI: Evaluating the Impact of Artificial Intelligence-Enhanced Image Analysis on the Diagnostic Accuracy of Frontline Clinicians in the Detection of Fractures on Plain X-Ray

Authors

Prof Alex Novak, Oxford University Hospitals NHS Foundation Trust, alex.novak@ouh.nhs.uk

Dr Sarim Ather, Oxford University Hospitals NHS Foundation Trust, sarim.ather@ouh.hns.uk

Dr Max Hollowday, Oxford University Hospitals NHS Foundation Trust, max.hollowday@nhs.net

Dr Immanuel Paul, Oxford University Hospitals NHS Foundation Trust, immanuel.paul@ouh.nhs.uk

Keywords

AI (artificial intelligence), diagnostic x-ray radiology, bone fractures, emergency departments

Abstract

Introduction

Missed fractures radiographs form the most frequent type of diagnostic error attributed to clinicians in UK Emergency Departments (ED) (Duron, 2021; Guly, 2001). Noting this, artificial Intelligence (AI)-enhanced algorithms have been developed to support clinicians in the detection of fractures on plain radiographs. Previous research has shown these algorithms to have promising diagnostic performance (Bousson et al., Guermazi et al., 2022), but the impact this has on added diagnostic accuracy has not yet been evaluated in a UK healthcare setting.

Methods and analysis

A dataset of 500 plain radiographs will be collated to include all bones excluding the skull, facial bones and cervical spine. The dataset will be split between radiographs with a fracture and those without. The reference *ground truth* for each image will be established through independent review by two senior musculoskeletal radiologists. This dataset will be analysed by commercially available algorithm BoneView© (Gleamer, Paris, France), and its accuracy for detecting fractures determined with reference to the ground truth. We will undertake a Multiple-Reader Multiple-Case study in which clinicians interpret all images without AI support and after a four-week washout period the same images again with access to AI support. Eighteen clinicians will be recruited as readers from six distinct clinical groups, each with three levels of seniority. Changes in the accuracy, confidence and speed of reporting will be compared with and without AI support. Results will be reported for all readers as well as for clinical role, level of seniority, pathological finding, and difficulty of image.

Ethics and Dissemination

The study has been approved by the UK Healthcare Research Authority (IRAS 310995, approved 13/12/2022). The use of anonymised retrospective radiographs has been authorised by Oxford University Hospitals NHS Foundation Trust. The results will be presented at relevant conferences and published in a peer-reviewed journal.

Strengths and limitations of study

This study will evaluate the impact of using an AI-assisted fracture detection algorithm to support UK clinicians in identifying fractures on plain radiographs.

This is the first detailed study of using an AI-assisted fracture detection algorithm in a UK clinical setting with an NHS-derived dataset. A range of professional backgrounds and levels of experience will be represented amongst the clinical participants, although the numbers of readers and levels of experience will be fairly small, reducing statistical power for any comparison between professional groups.

Introduction

A 2001 (Guly) study of missed diagnoses in UK emergency departments showed that missed fractures represented 79.7% of them, of these 77.8% were through the misreading of radiographs. The study mentions low levels of experience of clinicians as a leading factor in misdiagnosis, with cases only being picked up in the presence of a senior review. A more recent study (Shelmerdine et al. 2022) noted up to 11% of acute paediatric fractures are missed in the ED.

Recent advances in computer vision and machine learning have been used to augment interpretation of medical imaging, several artificial intelligence (AI) algorithms being developed for detection of fractures on radiographs.

AI fracture detection technology by Gleamer has been the subject of academic scrutiny already, Duron et al's 2021 reader study showing an increase in sensitivity and specificity amongst clinicians, without an increase in reading time, when supported by AI. Further reader studies (Guermazi et al., 2022) showed significant reduced reading time and improved sensitivity amongst AI-assisted clinicians. Gleamer's BoneView software has seen specific investigation in paediatric radiology on appendicular fractures, noting a low false positive rate (0.11%) and high sensitivity for all but avulsion fractures (Hayashi et al, 2022). Further reader studies (Bousson et al, 2023) looked at BoneView in comparison with commercial competitors, showing BoneView to have consistently high specificity and sensitvity, but without sacrificing accuracy.

This algorithm has received both CE marking and FDA approvals. This algorithm was chosen as it has the highest volume of peer reviewed evidence and is the mostly widely used fracture detection algorithms in the NHS (currently deployed in 5 NHS trusts) as well as worldwide (>800 sites in 30 countries). The algorithm estimates the likelihood of fracture, joint dislocation, effusions and bone lesions being present on a radiograph on a scale of 1-100 along. If the likelihood has been estimated to be above a designated cut-off value, the area of abnormality is highlighted as a region of interest on a secondary image which is made available to clinicians via their picture archive and communication system (PACS). If no abnormality is detected, this is also stated on the secondary image. Prior studies have demonstrated that the algorithm is highly accurate at detecting abnormalities, and it is already in use in a number of European centres, having received regulatory approval for use to support clinicians interpreting plain radiographs. Recent studies have reported that the use of AI software for detecting bone fractures can decrease the rate of missed fractures. However, this software has not yet been fully tested in a UK setting using a locally-derived dataset, and it is unknown how such systems would affect the diagnostic performance of staff groups specific to the NHS, such as reporting radiographers and Emergency Nurse Practitioners (ENPs).

Automation bias (the propensity for humans to favour suggestions from automated decision-making systems) is a known source of error in human-machine interaction (Challen et al., 2019), and has been one of a number of causes for concern regarding the increasing usage of AI in radiology (Neri et al, 2020). A recent reader study in mammography (Dratsch et al. 2023), suggested significant automation bias presence across all levels of experience, noting too that it was only the high-

experienced reporters that consistently picked up on AI error. During our study, we will assess the impact of incorrect advice given by the algorithm on the clinical end users.

This study would address this gap in the current evidence base, which is consistent with the NICE Evidence Standards Framework (ESF) for Digital Health Technologies (DHTs) that recommends retrospective and prospective evaluations of algorithms to assess their performance within UK healthcare setting (National Institute for Health and Care Excellence, 2023). This study will illustrate the impacts of BoneView™ on the diagnostic performance of the full range of clinicians that are tasked with detecting fractures in the NHS.

Study aims

1. Evaluate the impact of AI-enhanced imaging on the diagnostic performance, efficiency and confidence of clinicians in detecting fractures on plain radiographs (primary).

2. To determine the stand-alone diagnostic accuracy of the BoneView© AI tool with respect to the reference standard and compare it with the stand-alone diagnostic accuracy of clinicians (primary).

3. To determine associations between professional background and level of experience when determining the impact of AI support on clinician fracture detection (secondary).

4. To explore which imaging factors influence clinicians' reporting accuracy and efficiency, and algorithm performance, e.g. category of abnormality, size of abnormality, image quality, presence of multiple abnormalities (secondary).

5. Analyse whether clinicians are more likely to make a mistake when AI provides an incorrect diagnosis (secondary).

Methods and analysis

Study design

A fully crossed Multiple-Reader Multiple-Case (MRMC) study.

Case selection and composition

The image dataset will include anonymised radiographs of adult patients (≥ 18 years) who presented to the Emergency Department of Oxford University Hospitals NHS Foundation Trust (OUH) with a suspicion of fracture after injury to the limbs, pelvis, or thoraco-lumbar spine. As CT is the investigation of choice for skull and many cervical spine injuries, these will be excluded from the study. Paediatric patients will be excluded from the dataset as their fracture types differ from those in adults, and there is an on-going study evaluating this aspect (FRACTURE study).

To constitute the dataset, radiology reports will be screened from the radiology information system (RIS) to develop an enriched dataset of the 500 standard clinical examinations to reach the power calculation (see statistical section below) with prevalence of 50% normal and 50% abnormal cases. The ratio of radiographs from each anatomical location has been informed by the proportion of missed fractures mentioned in the NHS resolution report. (Table 1)

Table 1

Body part	**Number of cases in the dataset**

Spine	42
Shoulder	20
Elbow	20
Wrist/hand	150
Hip/Pelvis	130
Knee	42
Foot/Ankle	96

The dataset of each anatomical location will include at least one of the following high 20% of high clinical impact fractures including:

- Thoracic and lumbar spine: compression fracture of the vertebrae
- Pelvis: fracture of the acetabulum or pelvic ring
- Upper limb: fracture of the proximal humerus, radial head, distal radius and scaphoid
- Lower limb: fracture of the neck of femur, femoral condyle, tibial plateau, fibula head, talus, Lisfranc fracture.

Consecutive scans will be reviewed and all scans fitting the inclusion and exclusion criteria will be included until the case number requirements have been met. To ensure a like-for-like comparison, case finding for abnormal cases will be performed first. The normal scans will be age and sex matched per body part.

We will aim to include significant representation of the different image views, system type (mobile or fixed), system vendors, and patient demographics (e.g. age, sex) without any predefined quota.

The anonymised dataset will then be anonymised and uploaded to the Report and Image Quality Control (RAIQC) platform under an existing data governance approval from the Oxford University Hospitals NHS Foundation Trust Caldicott guardian.

Case inclusion and exclusion summary

Inclusion
-Plain radiographs of adult patients (age ≥18 years) presenting to the OUH Emergency Department with a suspected fracture.

Exclusion

- Plain skull radiographs
- Plain cervical spine radiographs
- Follow-up radiographs for known fracture
- Paediatric radiographs (age <18)
- Obvious fractures defined as:
  - Displacement > 5mm
  - Shortening > 5mm
  - Angulation > 5°

Setting

The reads will be performed using a secure web-based DICOM viewer (www.raiqc.com). The platform allows readers to view radiographs, where they can identify the site of an abnormality with a mouse click. The images will be viewable through a web browser on desktop or laptop devices,

reflecting standard real-world hospital practice in which radiographs are typically interpreted by clinicians without dedicated high-resolution viewing stations.

Prior to commencing each phase of the study, the readers will be asked to review 10 practice cases (not part of the 500 case dataset) to familiarise themselves with the use of the study platform and the output of the BoneView© tool.

Participants

Readers (n=18, six specialities with three readers from each):

- Emergency Medicine
- Trauma and Orthopaedic Surgery
- Emergency Nurse Practitioners
- Physiotherapy
- General Radiology
- Radiographers

Readers Experience:

- Consultant/Senior/Equivalent - >10yrs experience
- Middle Grade/Registrar/Equivalent – 5-10yrs experience
- Junior Grade/Senior House Officer/Equivalent - <5yrs experience

Each speciality reader group will include one reader at each level of experience. These speciality groups have been selected to show professions most likely to be involved in adult acute radiograph interpretation in or peripheral to the ED setting.

Readers will be recruited from across five NHS organisations that comprise the Thames Valley Emergency Medicine Research Network (www.TaVERNresearch.org):

- Oxford University Hospitals NHS Foundation Trust
- Royal Berkshire NHS Foundation Trust
- Buckinghamshire Healthcare NHS Trust
- Frimley Health NHS Foundation Trust
- Milton Keynes University Hospital NHS Foundation Trust

Ground Truthing

Once the dataset has been uploaded to the reading platform, every radiograph in the dataset will be independently reviewed and annotated for the presence or absence of abnormalities by two musculoskeletal radiologists

They will draw bounding boxes around each fracture they detect, and will grade the images based on the image quality and degree of difficulty of abnormality detection on a scale of 1-5.

In case of discrepancy between the findings of the two ground truth radiologists, a third senior musculoskeletal radiologist will review the images and arbitrate as to the presence or absence of abnormalities.

Inferencing the Image Dataset

The entire dataset of images will then be analysed using BoneView©, creating a duplicate dataset of radiographs with alerts and regions of interest indicated. It will also provide a linked output with the

estimated probability of a fracture as well as the threshold cut-off score for the presence of an abnormality.

Radiographic interpretation

All readers will review all 500 radiographs individually across two reporting rounds

In the first round, they will interpret the images as per clinical practice without any AI assistance. After a washout period of a month to mitigate the effects of recall bias, they will review the same 500 radiographs a second time with the assistance of BoneView©, which will contribute its suggestions as to abnormality presence and location. In both sessions clinicians will be blinded to the ground truth established by the MSK radiologists.

Clinician readers will be asked to identify the presence or absence of fracture by placing a marker on the image at the location of the fracture (if present) and to rank their confidence for fracture identification. Confidence rating will take the form of a Likert scale from 1 to 10, 1 being least confident, 10 being very confident)

The data collection phase is projected to finish by March 2024.

Outcome measures

Reader and AI performance will be evaluated using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Area Under Receiver Operating Characteristic Curve (AUC). Reader performance will be evaluated with and without AI assistance.

Reader speed will be evaluated as the mean review time per scan, with and without AI assistance.

Reader confidence will be evaluated as self-reported diagnostic confidence on a 10 point visual analogue scale, with and without AI assistance.

Data de-identification and management

Scans selected for the study will be anonymised in accordance with Oxford University Hospitals NHS Foundation Trust information governance protocol and uploaded to the secure image viewing platform (www.raiqc.com). Access to the radiographs will be controlled via the study platform using separate user accounts for each reader.

All study data will be entered into a password-protected and secure database. Individual reader accuracy scores will be anonymised, and the study team will not have access to the identifying link between the participants' personal details and the data. Data about the participants' seniority level and professional group will be retained to allow group comparisons.

Sample size and power calculation

500 images (250 normal, 250 abnormal).

The Multi-Reader Sample Size Sample Size Program for Diagnostic Studies was used to estimate power for the number of images in our study (https://perception.lab.uiowa.edu/power-sample-size-estimation). Parameter values for the error variances and the covariances were taken from our last multi-reader, multi-case study on detecting pneumothoraces 18 readers, reading 500 cases yields 85%

power to detect a difference in accuracy of 10% with a type 1 error of 5% (See output from software below). This is similar to previously published studies using the BoneView algorithm.

Statistical analyses

The performance of BoneView© will be compared with the ground truth generated by the musculoskeletal radiologist panel. The continuous probability score from the algorithm will be utilised for the AUC analyses, while binary classification results with three different operating cut-offs would be utilised for evaluation of sensitivity, specificity, PPV, and NPV.

Difference in AUC: sensitivity and specificity of readers with and without AI will be tested based on the Obuchowski-Rockette model for MRMC analysis which will model the data using a 2-way mixed effects ANOVA model treating readers and cases (images) as random effects and effect of AI as a fixed effect with recommended adjustment to degrees of freedom by Hillis et. al.

The main analysis will be performed as a single pool including all groups and sites.

Subgroup analyses will be performed for the following:

- Professional group (radiologist vs EM clinician vs radiographer)
- Senior vs middle vs junior
- Pathological finding
- Difficulty of image

Ethics and dissemination

The study has been approved by the UK Health Research Authority (IRAS number 310995, approved 13/12/2022). The use of anonymised retrospective radiographs has been authorised by the Caldicott Guardian and information governance team at Oxford University Hospitals NHS Foundation Trust. Readers will provide written informed consent and will be able to withdraw at any time.

The study is registered at Clinicaltrials.gov NCT06130397 and the ISRCTN registry (approval pending reference 44612). The results of the study will be presented at relevant conferences and published in peer-reviewed journals. The detailed study protocol will be freely available upon request to the corresponding author. Further dissemination strategy will be strongly guided by our PPIE activities. This will be based on co-productions between patient partners and academics and will involve media pieces (mainstream and social media) as well as communication through charity partners.

Authors' contributions

All authors contributed to the writing of the protocol and reviewed the manuscript, with Alex Novak and Sarim Ather leading the overall design. xx wrote the manuscript. Xxx are the ground-truthers.

Patient and public involvement

This protocol has been reviewed by the Oxford ACUTECare PPI group. They supported the study and its aims, influenced design and data management, and dissemination strategies. Lois Greenhalgh serves as PPI lead.

Funding statement

This work was supported by the NIHR Research for Patient Benefit in Health and Care Award NIHR204982.

Competing interests statement

Jeanne Ventre of the Steering Committee is an employee of Gleamer.

Sarim Ather is a shareholder of RAIQC Ltd.

Acknowledgments

FRACT-AI steering committee:

David Metcalfe, Susan Shelmerdine, Jason Oke, Fergus Gleeson, Micholas Woznitza, Sarah Wilson, Simon Triscott, Ravi Shashikala, Divyansh Guilati, Matthew Costa, Jeanne Ventre, Nick Welch, Daniel Jones

PPI Lead Lois Greenhalgh

Author Statement

Prof Alex Novak – Study design, protocol review, NIHR grant application, co-CI, specialist emergency medicine input.

Dr Sarim Ather – Study design, protocol review, co-CI, specialist radiology input.

Dr Max Hollowday – Protocol drafting, study registration, recruitment.

Dr Immanuel Paul – Protocol drafting, study registration, recruitment.

Summary

This study will help determine the likely impact of Gleamer BoneView on clinician accuracy, time, and confidence in detecting fractures. The findings will help determine whether this technology should be incorporated into routine clinical practice within NHS EDs. It has been established that there are errors in reporting of fractures on plain film radiographs by ED emergency department clinicians, and this may be reduced by augmentation with AI. This therefore offers a potential aid not just to diagnostic efficacy, but patient flow and clinician time. AI can assist a clinician by providing confirmation where the clinician may have less confidence in their conclusion, or similarly express a dissenting opinion that may compel earlier discussion with a senior colleague or specialist, benefiting patient care. It may also prove to reduce reliance on senior colleagues and specialists for simpler radiographs, providing a saving in time and resources.

Word Count: 3,310

References

Bousson, V., Attané, G., Benoist, N., Perronne, L., Diallo, A., Hadid-Beurrier, L., Martin, E., Hamzi, L., Depil Duval, A., Revue, E., Vicaut, E., Salvat, C., 2023. Artificial Intelligence for Detecting Acute Fractures in Patients Admitted to an Emergency Department: Real-Life Performance of Three Commercial Algorithms. Academic Radiology 30, 2118–2139. https://doi.org/10.1016/j.acra.2023.06.016

Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, Krasimira Tsaneva-Atanasova, 2019. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 28, 231. https://doi.org/10.1136/bmjqs-2018-008370

Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., Pinto dos Santos, D., 2023. Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. Radiology 307, e222176. https://doi.org/10.1148/radiol.222176

Duron, L., Ducarouge, A., Gillibert, A., Lain\uc0\u233{}, J., Allouche, C., Cherel, N., Zhang, Z., Nitche, N., Lacave, E., Pourchot, A., Felter, A., Lassalle, L., Regnard, N.-E., Feydy, A., 2021. Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study. Radiology 300, 120\uc0\u8211{}129. https://doi.org/10.1148/radiol.2021203886\

Guermazi, A., Tannoury, C., Kompel, A.J., Murakami, A.M., Ducarouge, A., Gillibert, A., Li, X., Tournier, A., Lahoud, Y., Jarraya, M., Lacave, E., Rahimi, H., Pourchot, A., Parisien, R.L., Merritt, A.C., Comeau, D., Regnard, N.-E., Hayashi, D., 2022. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. Radiology 302, 627\uc0\u8211{}636. https://doi.org/10.1148/radiol.210937\

H R Guly, 2001. Diagnostic errors in an accident and emergency department. Emerg Med J 18, 263. https://doi.org/10.1136/emj.18.4.263\

Hayashi, D., Kompel, A.J., Ventre, J., Ducarouge, A., Nguyen, T., Regnard, N.-E., Guermazi, A., 2022. Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning. Skeletal Radiology 51, 2129\uc0\u8211{}2139. https://doi.org/10.1007/s00256-022-04070-0\

Hillis SL, Soh BP. Obuchowski-Rockette analysis for multi-reader multi-case (MRMC) readers-nested-in-test study design with unequal numbers of readers. Proc SPIE Int Soc Opt Eng. 2023 Feb;12467:124670F. doi: 10.1117/12.2655190. Epub 2023 Apr 3. PMID: 37736244; PMCID: PMC10512791.

National Institute for Health and Care Excellence. (2023). Evidence standards framework (ESF) for digital health technologies. [Online]. www.nice.org.uk. Available at: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-hea [Accessed 21 November 2023].

Neri, E., Coppola, F., Miele, V., Bibbolino, C., Grassi, R., 2020. Artificial intelligence: Who is responsible for the diagnosis? La radiologia medica 125, 517–521. https://doi.org/10.1007/s11547-020-01135-9

Shelmerdine, S.C., White, R.D., Liu, H., Arthurs, O.J., Sebire, N.J., 2022. Artificial intelligence for radiological paediatric fracture assessment: a systematic review. Insights into Imaging 13, 94. https://doi.org/10.1186/s13244-022-01234-3\