HIGHLY CONFIDENTIAL

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE

Diagnostics Advisory Committee – Wednesday 18 September 2024

Artificial intelligence software to help detect fractures in on X-rays in urgent care

The following documents are made available to the Committee:

1. **Overview**

2. **Organisation submission from:** The Society of Radiographers

3. **Updated External Assessment Report (EAR) – prepared by** Peninsula Technology Assessment Group (PenTAG)
   Note, this report is an updated version to the one issued to stakeholders on 30 August 2024. The updates are listed on page 3a of the report.
   The update report also includes an addendum on impact of potential implementation costs

4. **EAR Consultation comments from Stakeholders and EAG responses.**

Artificial intelligence software to help detect fractures in on X-rays in urgent care

GID-HTE10044

Early Value Assessment

Assessment Report Overview

NICE National Institute for Health and Care Excellence

Background

NICE National Institute for Health and Care Excellence

Clinical background

Plain film radiography or X-ray is the most common medical imaging approach used to detect fractures in urgent care settings. Missed fractures are reported to be the most common diagnostic error in the ED[1].

Missed or delayed diagnosis of fractures on radiographs is reported to occur in around 3% to 10% of cases[2].

Missed fractures can lead to poor patient outcomes and further harms including[3]:

- pain and suffering
- loss of function
- need for further or prolonged treatments
- cosmetic deformity
- nerve damage
- prolonged recovery
- death.

Missed and delayed fracture diagnoses can also have an impact on service delivery, for example:

- increased waiting times
- delays in people being discharged
- people being recalled
- additional medical appointments
- surgical procedures and physiotherapy.

NICE

1. Hussain et al. 2019; 2. Kuo et al. 2022;
3. NHS resolution report on missed fractures

Unmet need

- The radiology get it right first time programme national speciality report highlights the increasing demand on radiology services that is not matched by growth in NHS radiology capacity. As a result, following interpretation in urgent care, a definitive diagnosis by a radiology specialists is often delayed.
- X-rays are initially interpreted in the urgent care setting by healthcare professionals who are not radiology specialists and may be inexperienced at interpreting X-rays, potentially leading to missed fractures or unnecessary referrals to fracture clinics prior to a definitive radiology report.
- Other factors that may contribute to missed or delayed diagnosis include busy work environments and frequent distractions, suboptimal image visualisation facilities, and interpretation outside normal working hours.

NICE

1. Hussain et al. 2019; 2. Kuo et al. 2022

Purpose of the technology

Artificial intelligence (AI) technologies that can help detect fractures and support healthcare professional interpretation of X-ray images could improve the accuracy of X-ray fracture diagnoses in urgent care settings. This could help reduce:

- the number fractures that are missed before a radiologist or reporting radiographer reviews the X-rays.
- the number of people being recalled to hospital following radiology review
- the risk of further injury or harm to people during the time between interpretation and initial treatment decision in the ED and the radiology report.
- the burden of unnecessary referrals to virtual fracture clinics.

Ionising radiation (medical exposure) regulations (IRMER)[1] state that clinical evaluation of X-rays requires a trained person. Therefore, AI technologies currently can't be used autonomously without human interpretation.

1. IR(ME)R: Implications for clinical practice in diagnostic imaging, interventional radiology and diagnostic nuclear medicine (2020)

Target condition and current practice

- Fracture assessment and diagnosis typically involves triage where an ED nurse, Advanced Clinical Practitioner (ACP) or ED doctor will carry out an initial assessment before requesting imaging.
- X-rays are usually the first line imaging approach for non-complex fractures and are performed by a diagnostic radiographer.
- Multiple treatment options are available for fractures including surgical and non-surgical approaches depending on the type of fracture.

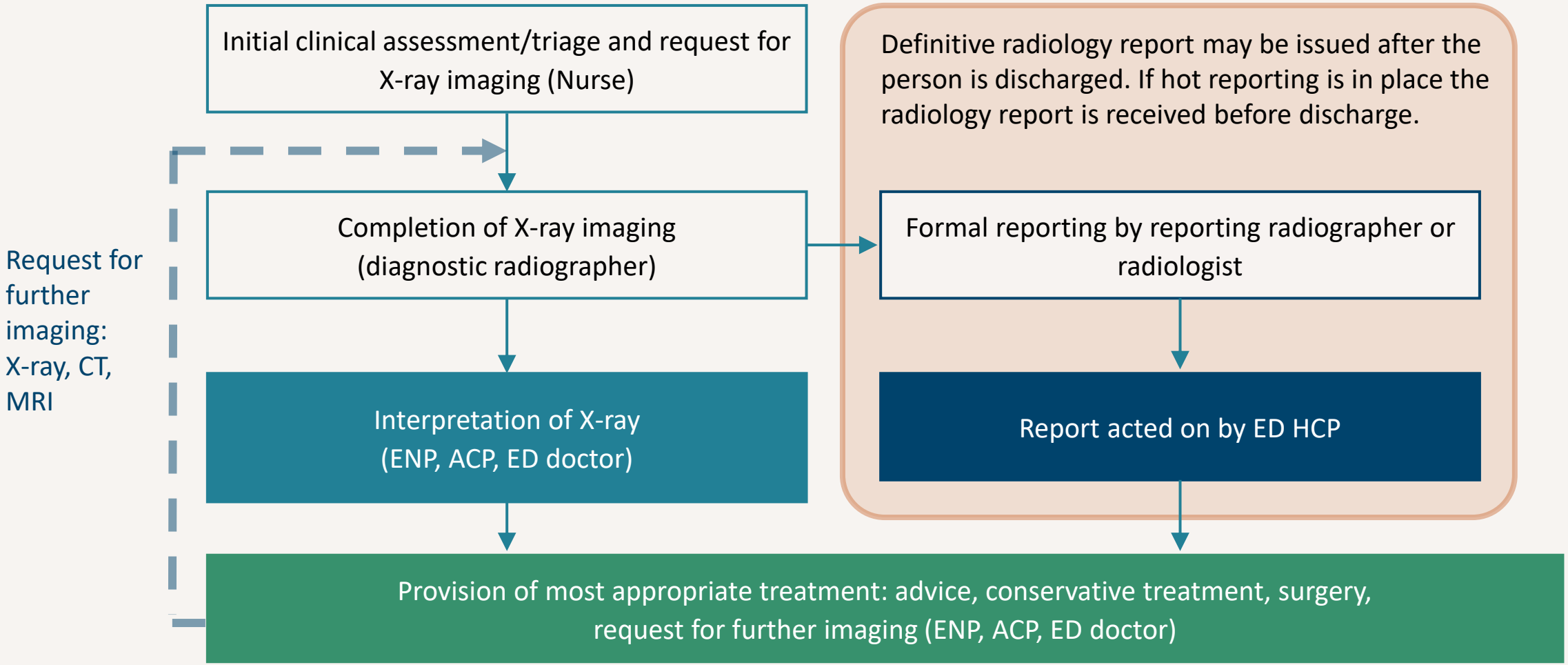
See the final scope for further details.

Variation in practice

- NICE guideline on non-complex fractures (NG38)[1]) recommends that a radiologist, radiographer or other trained reporter should review X-rays and provide a definitive report before the injured person is discharged (hot reporting). Clinical experts explained that in practice this is not always possible and reporting delays can occur ranging from days to weeks.
- Clinical experts said that X-rays are not usually prioritised for radiology reporting, with most centres operating a first-in, first-out system.
- There may be variations in the staff groups that would be involved in diagnosing fractures for people that attend an urgent care settings out of hours
- Different imaging types may be used for some suspected fractures, depending on centre resources and capacity:
    - NG38[1] recommends that MRI should be considered for first-line imaging for suspected scaphoid fractures
    - CG124[2] recommends offering MRI or CT if a hip fracture is suspected despite no fracture being detected on X-ray

1. NICE guideline 38 Fractures (non-complex): assessment and management, 2. NICE clinical guideline 124 Hip fracture: management

Overview of medical imaging pathway for non-complex fractures



NICE

ENP: emergency nurse practitioner, ACP: Advanced clinical practitioner, ED: emergency department, HCP: healthcare professional

Interventions

AI Technology (manufacturer)	CE marking	Regions covered	Population	Other
BoneView (Gleamer)	Class IIa	Appendicular skeleton, ribs and thoracic-lumbar spine	2 years and over	The software identifies fractures, dislocations, effusions and bone lesions
qMSK (Qure.ai)	Class IIb	Appendicular skeleton and ribs	Adults	
Rayvolve (AZmed)	Class IIa	Appendicular skeleton and ribs	Adults	Detects dislocations, joint effusions and chest pathologies (pneumothoraces, cardiomegaly, pleural effusions, pulmonary oedema, consolidation, nodules)
RBfracture (Radiobotics)	Class IIa	Appendicular skeleton	Approved for use in people above 2 years of age	Detects effusion of the knee and elbow, lipohaemarthrosis of the knee, rib fractures, and periprosthetic fractures
TechCare Alert (Milvue)	Class IIa	Appendicular skeleton and ribs	No age limit	Detects dislocations, elbow joint effusion, pleural effusion, pulmonary opacity, pulmonary nodules and pneumothorax

Decision problem (1)

- Does the use of software with artificial intelligence (AI) derived algorithms for analysing X-ray images to detect suspected fractures have the potential to be clinically and cost-effective to the NHS?
- Does the software have the potential to address an unmet need in the NHS?

PICO	
Populations	People presenting to the ED, UTC or MIU with a suspected fracture.
Potential subgroups	• Children and young people (0 to 16 years of age) • Older people or people with frailty • People with conditions affecting bone health (for example, osteoporosis and osteogenesis imperfecta) • Hip • Hand (including wrist), foot (including ankle) • Fractures including the growth plate (Salter-Harris) in children • Fractures of the elbow in children
Interventions	AI used as a decision aid for X-ray image interpretation and fracture assessment prior to radiology review
Comparator	ED clinician or healthcare professional interpretation of X-ray radiograph without AI assistance.

For further details see the final scope and the EAG's final protocol.
ED, emergency department; UTC, urgent treatment centre; MIU, minor injuries unit

Decision problem (2)

Outcomes Intermediate measures for consideration may include:	• Measures of diagnostic accuracy to detect fractures • Accuracy when used by different healthcare professionals • Diagnostic confidence • Healthcare professional X-ray reading time • Time to diagnosis or time to X-ray definitive radiology report • Time spent in the emergency department, urgent treatment centre or minor injuries unit • Time to treatment • Proportion of people that need further imaging • Number of missed fractures • Rate of missed fracture-related further injury • Number of people recalled following radiology review • Number of treatments • Number of hospital appointment/visits including referrals to fracture clinics and orthopaedics • Number of hospital admissions and length of stay in hospital • Number of further imaging events required • Failure rate or rate of inconclusive AI reports • Healthcare professional user acceptability of AI tools for detecting fractures

For further details see the final scope and the EAG's final protocol

Clinical effectiveness

NICE National Institute for Health and Care Excellence

Clinical effectiveness: evidence base

- 16 studies were identified that met the inclusion criteria for the clinical effectiveness review.
    - 8 studies evaluated BoneView
    - 5 studies evaluated RBfracture
    - 1 study each for Rayvolve and TechCare Alert
    - No studies were found for qMSK
    - One study (Bousson et al. 2023) was a head-to-head comparison of assisted reading using 3 technologies: BoneView, Rayvolve and TechCare Alert.

Full details of the included studies are in table 2, pages 26 to 29 in the EAR.

Evidence base: outcomes

- Sensitivity, specificity, and contingency tables were reported or calculable for all studies.
- Diagnostic accuracy that was reported per patient (rather than per fracture or per scan) was prioritised by the EAG for inclusion in the review. This is because most studies reported data in this way and because these data were most relevant to the economic analysis.
- PPV and NPV were either not reported or were not extracted for case-control studies.
- Full details of outcomes reported in the included studies are presented in table 5 (page 50) in the EAR.

PPV, positive predictive value; NPV, negative predictive value

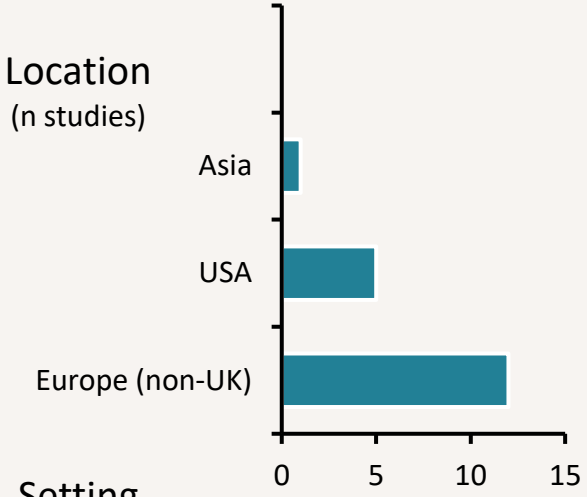
Evidence base: impact of study design

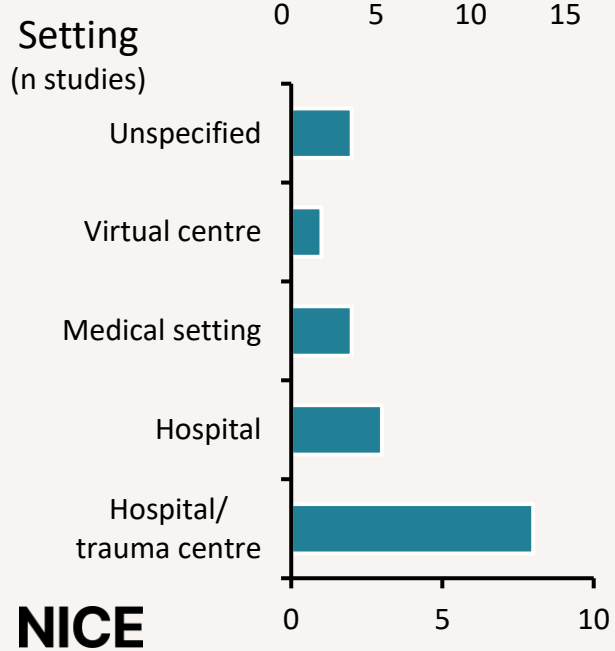
- The EAG said that consecutive and random sampling study designs are generally more robust for diagnostic evaluation, as they more closely represent the prevalence of the target condition that would be seen in clinical practice.
- Six studies[1–6] included consecutive cases presenting to participating centres during the study period.
- One study[7] included a random selection of cases from a database of patients who presented with a suspected fracture.
- Nine studies[8–16] used a case-control design.
- Most studies were retrospective with only 2[3,5] using a prospective design.

See section 4.2.1 (page 30) of the EAR for further details.

1. Cohen et al. (2023); 2. Canoni-Meynet et al. (2022); 3. Oppenheimer et al. (2023); 4. Bousson et al. (2023); 5. Dell-Aria et al. (2024); 6. Ruitenbeek et al. (2024); 7. Fu et al. (2024); 8. Radiobotics (2021); 9. Meetschen et al. (2024); 10. Duron et al. (2021); 11. Guermazi et al. (2022); 12. Bachmann et al. (2024); 13. Jørgensen et al. (2023); 14. Yogendra et al.; 15. Suite (2020); 16. Nguyen et al. (2022)

Evidence base limitations





- None of the included studies were set in the UK.
  - The EAG stated that it is uncertain how applicable data from other countries is to the target settings
- Diagnostic accuracy of the index test and technologies would be expected to vary according to both case mix and reader experience.
- Most studies did not specify the version of the technology or training received
- Washout period between reading ranged between no washout to 3 months
- Where the reference standard was based on limited access to information about the patient and injury, the EAG considered there to be an increased risk of incorrect judgements
- Reference standard used in 3 studies[1-3] included the results of the AI technology, and in one study[4] it was unclear whether this was the case.

1. Canoni-Meynet et al. 2022; 2. Bousson et al. 2023; 3. Yogendra [unpublished]; 4. Oppenheimer et al. 2023. For more detail see section 4.2.1 in EAR

NICE

Quality appraisal of included studies

- Three studies[1-3] were considered to be the most appropriate for sensitivity and specificity estimates.
- Only 1 study[4] was considered to be appropriate for estimates of prevalence, NPV and PPV. However, this study only included wrist fractures.
- None of the included evidence was considered to be robust for all diagnostic outcomes.
- The EAG did not do a formal quality assessment of the included studies. An overview of how quality considerations influenced the interpretation of diagnostic evidence and the selection of evidence to inform the economic analysis is shown in Table 6 (pages 51 to 52) in the EAR.

See slides 20, 23, 25 and 27 for further details on the quality assessment of the key studies.

1. Duron et al. 2021; 2. Nguyen et al. 2022; 3. Bachmann et al. 2024;
4. Cohen at el. 2023

Evidence base: key studies

The EAG investigated whether it was possible to conduct a meta-analysis of data from the included studies:

- A meta-analysis to identify a pooled estimate of sensitivity and specificity for a particular technology was not feasible due to unexplained heterogeneity in the results of the studies.

Due to the evidence limitations and lack of meta-analysis, the EAG identified key studies that provided the best quality evidence available to inform the economic analysis.

The EAG prioritised studies which:

- did not include the AI reports in the reference standard
- reported results for both AI assisted and unassisted readers
- had relatively large sample sizes
- were peer-reviewed.

The key studies for each technology were:

- BoneView: Duron et al. 2021 (for adults) and Nguyen et al. 2022 (for children and young adults)
- RBfracture: Bachmann et al. 2024 (adults and children)
- Rayvolve: Fu et al. 2024 and Bousson et al. 2023
- TechCare Alert: Suite 2020* and Bousson et al. 2023.

*Suite 2020 was not presented in a peer-reviewed publication. Evidence for this technology was therefore of poorer quality than the studies listed above.

Diagnostic accuracy: issues

- The EAG grouped results according to the description of reader experience and seniority described in the publications. As all included studies were based outside of the UK, it was unclear how relevant the staff grades were to the intended staff groups in the NHS. Three reader groupings were used:
    - Less experienced,
    - mixed or unclear staff level, and
    - senior and highly experienced staff.
- In general, readers with greater seniority and expertise at reading X-rays were associated with more accurate unassisted diagnosis estimates, but this was not always the case. The EAG noted that, as this lacks face validity, these results should be interpreted with caution when pooled for the evidence synthesis.

Key studies: Boneview

Study design feature	Duron et al. (2021)	Nguyen et al. (2022)
Location	France	USA/France
Intervention and comparator	BoneView assisted vs unassisted	
Design	Retrospective, case-control with random selection	
Reference standard	Consensus between 2 skeletal imaging radiologists with >9 years of experience (disagreements resolved by another radiologist). Timing: NR.	Consensus between 2 radiologists with >8 years of experience in MSK imaging. Paediatric radiologist (with 5 years of specialisation) reviewed all radiographs after ground truth was determined to classify fractures by type. Timing: NR.
Reader details	6 radiologists, 6 emergency physicians (including residents and experts). No consultant support.	5 radiology residents and 3 expert paediatric radiologists (at least 7 years of experience, including >3 years specialising in paediatric radiology (No consultant support).
Population, fracture sites and prevalence	600 adults, 600 images, included shoulder, arm, hand, pelvis, leg, foot. N with fractures 300 (50%): Foot 44 (7.4%); Hand 44 (7.3%)	300 children and young adults (2 to 21 years), 300 images. Appendicular. N with fractures 150 (50%): foot/ankle 30 (10%); hand/wrist 30 (10%); elbow/arm in children 60 (20%); Salter-Harris, Salter II 21 (7%), Salter IV: 3 (1%)
Quality assessment	Sensitivity and specificity: Green Prevalence, NPV and PPV: Red - Case-control design	

NICE

See tables 2, 3 and 4 in the EAR for full details. NPV, negative predictive value; PPV, positive predictive value

Diagnostic accuracy: Boneview (1)

Mixed fracture and age groups

Study	Staff experience group	Sensitivity (assisted)	Sensitivity (unassisted)	Specificity (assisted)	Specificity (unassisted)
Duron et al. (2021)	Less experienced readers (emergency physicians)	74.3	61.3	96.6	90.6
Nguyen et al. (2022)	Mixed or unclear	82.7	73.2	90.3	89.6

Hand and wrist

Study	Staff experience group	Sensitivity (assisted)	Sensitivity (unassisted)	Specificity (assisted)	Specificity (unassisted)
Duron et al. (2021)	Mixed or unclear	80.2	59.6	91.0	84.7
Nguyen et al. (2022)	Mixed or unclear	87.1	68.8	88.3	87.9

Further details including results for other staff experience groups are shown in tables 7 and 9 in the EAR.

Diagnostic accuracy: Boneview (2)

Foot and ankle

Study	Staff experience group	Sensitivity (assisted)	Sensitivity (unassisted)	Specificity (assisted)	Specificity (unassisted)
Duron et al. (2021)	Mixed or unclear	86.9	71.8	92.9	88.0
Nguyen et al. (2022)	Mixed or unclear	70.8	53.8	86.3	85.8

Salter-Harris

Study	Staff experience group	Sensitivity (assisted)	Sensitivity (unassisted)	Specificity (assisted)	Specificity (unassisted)
Nguyen et al. (2022)	Mixed or unclear	92.3	81.0	NR	NR

Further details including results for other staff experience groups are shown in tables 7 and 9 in the EAR.

Key studies: RBfracture

Study design feature	Bachmann et al. (2024)
Location	US/Denmark
Intervention and comparator	RBfracture assisted vs unassisted
Design	Retrospective, case-control with random selection
Reference standard	Consensus between 2 consultant radiologists with 1 and 10 years post-specialty experience. Each had access to clinical referral notes and radiology reports. Timing: NR. N = 340 (6 didn't receive reference standard).
Reader details	2 advanced trauma nurses, 3 diagnostic radiographers, 4 A&E trainees, 3 orthopaedic specialty registrars, 3 radiology specialty registrars. Provided with written instructions and 5 training cases.
Population, fracture sites and prevalence	340 adults over 21 years and children over 2 years, 340 images. Appendicular skeleton, excluded ribs and spine. N with fractures 164 (49.1%): pelvis/hip 19 (5.7%); foot/ankle 30 (9.0%); hand/wrist 30 (9.0%); elbow in children 9 (2.7%)
Quality assessment	Sensitivity and specificity: Green Prevalence, NPV and PPV: Red - Case-control design

23 See tables 2, 3 and 4 in the EAR for full details. NPV, negative predictive value; PPV, positive predictive value

Diagnostic accuracy: RBfracture

Mixed fracture and age groups

Study	Staff experience group	Sensitivity (assisted)	Sensitivity (unassisted)	Specificity (assisted)	Specificity (unassisted)
Bachmann et al. (2024)	Less experienced readers (A&E trainees)	83	74	90	87
	Less experienced readers (Trauma-care nurses)	70	58	67	60

Further details including results for other staff experience groups are shown in tables 7 and 9 in the EAR.

Key studies: Rayvolve

Study design feature	Fu et al. 2024	Bousson et al. 2023
Location	USA	France
Intervention and comparator	Rayvolve assisted vs unassisted	BoneView/Rayvolve/TechCare Alert head-to-head
Design	Retrospective, case-control with random selection	Retrospective, consecutive sampling
Reference standard	Consensus between at least 2 of 3 musculoskeletal radiologists with 7 to 16 years of experience.	Consensus between 4 musculoskeletal radiologists, 3 fellows and 1 senior radiologist. Combination of radiology reports and AI results. Timing: 2 months later
Reader details	8 each of emergency physicians, non-MSK radiologists, and MSK radiologists.	6 radiology residents (4 years of residency)
Population, fracture sites and prevalence	Adults over 22 years, sample size NR but 186 exams. Ankle, clavicle, elbow, forearm, humerus, hip, knee, pelvis, shoulder, tibia/fibula, wrist, hand, foot	1,210 adults and adolescents (15 years or older), 1,500 images. Appendicular skeleton. N of fractures 326 (21.7%); Pelvis/hip 50 (3.3%); Ankle 232 (15.5%); Foot 186 (12.4%); hand/wrist 314 (20.9%)
Quality assessment	Sensitivity and specificity: Amber: small sample size Prevalence, NPV and PPV: Amber: random sampling, but limited due to small sample size.	Sensitivity and specificity: Amber: reference standard includes AI results Prevalence, NPV and PPV: Amber: consecutive sampling, but limited due to reference standard including AI results

See tables 2, 3 and 4 in the EAR for full details. NPV, negative predictive value; PPV, positive predictive value

Diagnostic accuracy: Rayvolve

Mixed fracture and age groups

Study	Staff experience group	Sensitivity (assisted)	Sensitivity (unassisted)	Specificity (assisted)	Specificity (unassisted)
Fu et al. 2024	Less experienced readers (emergency physicians)	93.8	79.2	85.3	85.2
Bousson et al. 2023	Mixed or unclear	92.6	NR	70.4	NR

Hand and wrist

Study	Staff experience group	Sensitivity (assisted)	Sensitivity (unassisted)	Specificity (assisted)	Specificity (unassisted)
Bousson et al. 2023	Mixed or unclear	97.8	NR	74.6	NR

Foot and ankle

Study	Staff experience group	Sensitivity (assisted)	Sensitivity (unassisted)	Specificity (assisted)	Specificity (unassisted)
Bousson et al. 2023	Mixed or unclear	Foot: 90.8 Ankle: 92.1	NR	67 (EAG calculation)	NR

Further details including results for other staff experience groups are shown in tables 7 and 9 in the EAR.

NICE

Key studies: TechCare Alert

Study design feature	Suite 2020	Bousson et al. 2023
Location	France	France
Intervention and comparator	TechCare Alert assisted vs unassisted	BoneView/Rayvolve/TechCare Alert head-to-head
Design	Retrospective, case-control with random selection	Retrospective, consecutive sampling
Reference standard	Original radiology report produced by a radiologist	Consensus between 4 musculoskeletal radiologists, 3 fellows and 1 senior radiologist. Combination of radiology reports and AI results. Timing: 2 months later
Reader details	4 junior and 4 senior radiologists.	6 radiology residents (4 years of residency)
Population, fracture sites and prevalence	N with fractures 253 (40.8%); dislocation 28 (4.5%); effusion 25 (36.2%): hip 67 (10.8%); foot/ankle 144 (23.2%); hand/wrist 134 (21.6%); elbow in children 30 (9.4%)	1,210 adults and adolescents (15 years or older), 1,500 images. Appendicular skeleton. N with fractures 326* (21.7%): Pelvis/hip 50 (3.3%); Ankle 232 (15.5%); Foot 186 (12.4%); hand/wrist 314 (20.9%);
Quality assessment	Sensitivity and specificity: Amber: not peer-reviewed and reference standard decision by a single radiologist. Prevalence, NPV and PPV: Red: Case-control design	Sensitivity and specificity: Amber: reference standard includes AI results Prevalence, NPV and PPV: Amber: consecutive sampling, but limited due to reference standard including AI results

See tables 2, 3 and 4 in the EAR for full details. NPV, negative predictive value; PPV, positive predictive value

Diagnostic accuracy: TechCare Alert

Mixed fracture and age groups

Study	Staff experience group	Sensitivity (assisted)	Sensitivity (unassisted)	Specificity (assisted)	Specificity (unassisted)
Suite 2020	Less experienced readers	95	92	98	97
Bousson et al. 2023	Mixed or unclear	90.2	NR	92.5	NR

Hand and wrist

Study	Staff experience group	Sensitivity (assisted)	Sensitivity (unassisted)	Specificity (assisted)	Specificity (unassisted)
Bousson et al. 2023	Mixed or unclear	93.6	NR	91.7	NR

Foot and ankle

Study	Staff experience group	Sensitivity (assisted)	Sensitivity (unassisted)	Specificity (assisted)	Specificity (unassisted)
Bousson et al. 2023	Mixed or unclear	Foot: 85.4 Ankle: 89.9	NR	91 (EAG calculation)	NR

Further details including results for other staff experience groups are shown in tables 7 and 9 in the EAR.

NICE

Diagnostic accuracy: Paediatric subgroup

Two of the key studies[1,2] reported diagnostic accuracy data in children and young people. One evaluated Boneview and was in children and young people only[1] and 1 study evaluated RBfracture and reported paediatric subgroup data[2].

Study	Staff experience group	Sensitivity (assisted)	Sensitivity (unassisted)	Specificity (assisted)	Specificity (unassisted)
Nguyen et al. (2022)	Mixed or unclear	82.7	73.2	90.3	89.6
Bachmann et al. (2024)	Mixed or unclear	89.0	78.0	80.0	77.0

Further details including results for other staff experience groups are shown in table 8 in the EAR.

NICE

1. Nguyen et al. (2022); 2. Bachmann et al. (2024)

Evidence synthesis (1)

The EAG used 2 approaches to synthesise the evidence base:

1. Data from included studies within each grouping was summarised using median and ranges (see slides 31 to 37)
2. Conducted a narrative synthesis to identify patterns in the data that could be used to inform an understanding about the potential value of the technology for assisting in the diagnosis of fractures.

- Synthesised data from the included studies was split by fracture type (all fractures and specific fracture sites).
- The EAG said that the results provide an insight into potential patterns across the dataset, rather than precise diagnostic accuracy data for the technologies.

For further details see section 5.2 in the EAR

Evidence synthesis (2): unassisted diagnostic accuracy

Staff group (n studies) all fractures	Median sensitivity (range)	Median specificity (range)	Median % missed fractures (range)	Median % over diagnosis (range)
Any staff (11)	72 (31, 93)	89 (60, 100)	28.1% (7.1, 42.1)	13.3% (1.4, 40.0)
Less experienced staff (7)	70 (58, 92)	87 (60, 97)	30.4% (7.9, 42.1)	12.9% (3.0, 40.0)
Mixed or unclear staff (9)	73 (58, 87)	90 (77, 97)	26.7% (13.5, 42.0)	10.7% (2.9, 23.0)
Senior and expert staff (3)	[redacted]	[redacted]	[redacted]	[redacted]

- The EAG noted that the rate of missed fractures for clinicians reading X-rays without AI assistance was high across studies, even for senior and expert readers.
- Sensitivity and specificity each varied significantly across studies though, in general, unassisted readers had higher specificity, resulting in a high median rate of missed fractures.
- Accuracy of unassisted readers for detecting hip fractures was high.
- Sensitivity for detecting hand/wrist and foot/ankle fractures was lower than the mixed fracture analyses, and there was variability in specificity for detecting hand/wrist fractures across studies.
- There was poorer sensitivity for identifying non-obvious fractures across all readers.

See table 11 in the EAR for full details

Boneview accuracy across studies (all fractures)

Staff group (n studies)	Median sensitivity (range)	Median specificity (range)	Median % missed fractures (range)	Median % over diagnosis (range)
Any staff (7)	83 (75, 91)	93 (65, 99)	17.3% (8.6, 25.0)	7.4% (0.6, 35.0)
Less experienced staff (3)	79 (74, 89)	97 (93, 98)	21.3% (11.2, 25.7)	3.3% (1.7, 7.4)
Mixed or unclear staff (6)	81 (75, 91)	90 (65, 97)	19.0% (8.6, 25.0)	9.8% (2.6, 35.0)
Senior and expert staff (1)	88 (NA)	99 (NA)	11.7% (NA)	0.6% (NA)

The EAG noted that BoneView showed high sensitivity and specificity, irrespective of the reader group. However, median numbers of missed fractures (all fracture analyses) exceeded 15% for all readers except the senior and expert reader group. In general, BoneView had improved specificity relative to sensitivity, with fewer false positives than missed fractures.

Boneview accuracy across studies (by fracture site)

Staff group and fracture site (n studies)	Median sensitivity (range)	Median specificity (range)	Median % missed fractures (range)	Median % over diagnosis (range)
Mixed or unclear staff, hand/wrist (4)	89 (80, 96)	92 (88, 95)	13.3% (8.3, 20.5)	9.4% (7.8, 13.3)
Senior and expert staff, hand/wrist (2)	89 (88, 90)	92 (NA)	11.7% (NA)	7.4% (NA)
Mixed or unclear staff, foot/ankle (4)	89 (71, 98)	93 (80, 96)	11.6% (1.8, 30.0)	10.4% (6.8, 20.8)
Senior and expert staff, foot/ankle (1)	83 (NA)	NR	NR	NR
Mixed or unclear staff, hip (1)	93 (NA)	99 (NA)	NR	NR
Less experienced staff, non-obvious fractures (1)	56 (NA)	79 (NA)	43.8% (NA)	21.1% (NA)
Mixed or unclear staff, non-obvious fractures (1)	83 (NA)	NR	16.7%	100%
Senior or expert staff, non-obvious fractures (1)	81 (NA)	89 (NA)	18.8% (NA)	10.5% (NA)

Sensitivity for non-obvious fractures was improved compared to the results for unassisted, although the rate of missed fractures and false positives was still high in the less experienced staff group (43.8% and 21.1%).

RBfracture accuracy across studies

Staff group and fracture site (n studies)	Median sensitivity (range)	Median specificity (range)	Median % missed fractures (range)	Median % over diagnosis (range)
Any staff, all fractures (3)	[redacted]	[redacted]	[redacted]	[redacted]
Less experienced staff, all fractures (2)	[redacted]	[redacted]	[redacted]	[redacted]
Mixed or unclear staff, all fractures (3)	[redacted]	[redacted]	[redacted]	[redacted]
Senior and expert staff, all fractures (1)	[redacted]	[redacted]	[redacted]	[redacted]
Junior, hip (1)	[redacted]	[redacted]	[redacted]	[redacted]
Mixed or unclear, hip (2)	[redacted]	[redacted]	[redacted]	[redacted]

The EAG noted that RBfracture showed good sensitivity and specificity across all studies, however rates of false positives across reader experience levels were similar to unassisted readers.

NICE

Rayvolve accuracy across studies

Staff group and fracture site (n studies)	Median sensitivity (range)	Median specificity (range)	Median % missed fractures (range)	Median % over diagnosis (range)
Any staff, all fractures (2)	94 (93, 96)	83 (70, 85)	6.1% (4.4, 7.4)	16.9% (14.7, 29.6)
Junior staff, all fractures (1)	94 (NA)	85 (NA)	6.1% (NA)	14.7% (NA)
Mixed or unclear staff, all fractures (2)	94 (93, 96)	77 (70, 83)	5.9% (4.4, 7.4)	23.3% (16.9, 29.6)
Mixed or unclear staff, hand/wrist (1)	98 (NA)	75 (NA)	2.1% (NA)	25.4%
Mixed or unclear staff, foot/ankle (1)	91 (91, 92)	67 (63, 72)	8.0% (7.1, 8.9)	32.8% (27.9, 37.7)

Two studies evaluated Rayvolve, both of which reported high sensitivity but poor specificity, particularly for hand/wrist and foot/ankle fractures. The EAG considered this was a feature of the technology algorithm, to prioritise missed fractures over false positives. Accordingly, specificity was comparable with unassisted diagnosis, while sensitivity was generally improved.

TechCare Alert accuracy across studies

Staff group and fracture site (n studies)	Median sensitivity (range)	Median specificity (range)	Median % missed fractures (range)	Median % over diagnosis (range)
Any staff, all fractures (2)	93 (90, 95)	98 (93, 98)	7.1% (5.1, 9.8)	1.9% (1.9, 7.5)
Junior staff, all fractures (1)	95 (NA)	98 (NA)	5.1% (NA)	1.9% (NA)
Mixed or unclear staff, all fractures (1)	90 (NA)	93 (NA)	5.1% (NA)	1.9% (NA)
Senior and expert staff, all fractures (1)	93 (NA)	98 (NA)	7.1%	1.9%
Mixed or unclear staff, hand/wrist (1)	94 (NA)	92 (NA)	6.2% (NA)	8.3% (NA)
Mixed or unclear staff, foot/ankle (1)	88 (85, 90)	91 (90, 92)	11.9% (9.5, 14.3)	8.6% (7.9, 9.2)

Two studies evaluated TechCare Alert, with no crossover in the reader groupings. Both reported high sensitivity and specificity estimates.

Diagnostic accuracy across studies in paediatric participants (all fractures)

Staff group and intervention (n studies)	Median sensitivity (range)	Median specificity (range)	Median % missed fractures (range)	Median % over diagnosis (range)
Mixed or unclear staff, unassisted (4)	78 (73, 100)	91 (77, 95)	22.6% (0.0, 26.7)	9.3% (5.6, 22.5)
Mixed or unclear staff, BoneView (2)	91 (83, 100)	91 (30, 92)	8.7% (0, 17.3)	9.0% (8.0, 10.0)
Mixed or unclear staff, RBFracture (2)	[redacted]	[redacted]	[redacted]	[redacted]
Senior and expert staff, unassisted (1)	[redacted]	[redacted]	[redacted]	[redacted]
Senior and expert staff, RBFracture (1)	[redacted]	[redacted]	[redacted]	[redacted]

No diagnostic accuracy data in children and young people was available in a less experienced reader group only. In mixed or unclear experience readers, BoneView and RBfracture improved median sensitivity for detecting fractures, though made no clear difference to specificity.

NICE

Subgroup evidence availability

- Five[1-5] studies included a mix of adults, children and young people, with 2[3,4] reporting subgroup data specifically in children and young people.
- Two[6,7] studies were conducted only in children and young people
- No studies reported frailty measures for participants and no studies reported information on the number of participants with diseases that affect bone health.

1.Meetschen et al. 2024; 2. Canoni-Meynet et al. 2022;
3. Oppenheimer et al. 2023; 4. Bachmann et al. 2024;
5. Bousson et al. 2023; 6. Nguyen et al. 2022; 7. Yogendra [unpublished]

AIC

X-ray reading time

Data on X-ray reading time with and without AI assistance was available for 3 technologies: BoneView (4 studies), RBfracture (2 studies) and Rayvolve (1 study).

- There were no noticeable differences in reading time across the staff groupings.
- BoneView and Rayvolve were both associated with a reduction in X-ray reading times across all staff groups:
    - BoneView 2.6 to 13 seconds per X-ray
    - Rayvolve 7 seconds per X-ray
- One study reported that RBfracture was associated with [redacted]
- There were large standard deviations around reading time in all studies, which may be due in part to the reading time varying widely by type and complexity of the fracture.
- The EAG was also concerned about the reliability of how reading time would be measured in studies, and potential differences in the way this was defined and recorded between studies.

NICE

39 Reading times for all fracture types by staff experience group are shown in table 10 (page 77) in the EAR

Summary of clinical effectiveness evidence

- There was a trend (across technologies and reader groups) for the AI technologies to improve sensitivity with little improvement in specificity
- Differences in accuracy between the technologies are uncertain due to limited evidence and variation in study designs
- Very few studies are specific to emergency care settings and all were associated with limitations due to risk of bias or uncertain generalisability.
- Fractures were still missed with AI assisted interpretation in all reader groups. Reported rates of missed fractures across all studies and fracture types ranged from 1.8% to 43.8%
- In children and young people, 2 key studies reported improved sensitivity but little improvement to specificity
- No evidence for people who are frail or with conditions that affect bone health and long-term recovery

Summary of key diagnostic accuracy results

Boneview:

- Key studies reported improvements in sensitivity and specificity, across all fracture types.
- Improved sensitivity and to a lesser extent specificity, irrespective of the reader group
- In general, BoneView had improved specificity relative to sensitivity, with fewer false positives than missed fractures.

RBfracture:

- Key study reported improved sensitivity and specificity in the less experienced reader group when using RBfracture to help diagnose mixed fractures
- Showed high sensitivity and specificity across all studies, however rates of false positives across reader experience levels were similar to unassisted readers

Rayvolve:

- Key study reported improved sensitivity in less experienced and mixed reader groups. Specificity (where reported) was similar with or without AI assistance.
- Across all studies poor specificity, particularly for hand/wrist and foot/ankle fractures. Specificity was comparable with unassisted diagnosis, while sensitivity was generally improved

Techcare Alert:

- In key studies and across all studies, Techcare Alert showed high sensitivity and specificity

Ongoing studies and evidence

Study	AI technology	Primary outcomes	Completion/end date
AI assisted detection of fractures on X-rays (FRACT-AI) study (NCT06130397)	Boneview	Diagnostic accuracy (sensitivity and specificity) of the AI algorithm alone, diagnostic accuracy of readers with and without AI assistance, and reader speed with and without AI	June 2025 (estimated)
Paediatric Fracture Study (NIHR301322)	Boneview	Diagnostic accuracy (sensitivity and specificity) of the AI algorithm alone, diagnostic accuracy of readers with and without AI assistance.	August 2026
[redacted]	[redacted]	[redacted]	[redacted]
5 NHS based real world data collection studies	RBfracture	Increase in productivity through time-saving; reductions in missed fractures (ED), CT scans, inappropriate referrals to fracture clinic, equivocal findings. Post-market surveillance data and standalone performance	Late 2024 to late 2025
AutoRayValid-RBfracture	RBfracture	Multi-national, retrospective generalisability study. Aims to assess AI impact on diagnostic thinking by analyzing consecutive cases with clinical data, providing insights into fracture detection and clinical decision-making	Not reported

Economic evaluation

NICE National Institute for Health and Care Excellence

Review of the economic literature (1)

See section 4.1 of the EAR for details of the evidence searches used to identify relevant economic studies.

- No economic evaluations of AI to detect fractures were identified
- The EAG identified 4 studies that were used to inform health state costs and utilities:
    1. **Rua et al. (2020)** used to inform modelling of hand/wrist fractures
    2. **Nwankwo et al. (2022)** used to inform modelling of foot/ankle fractures
    3. **Low et al. (2021)** used to inform modelling of hip fractures
    4. Judge et al. (2016) used to inform modelling of hip fractures

Economic model (1)

- The EAG developed a de novo model to explore the potential cost-effectiveness of AI-assisted diagnosis compared with unassisted diagnosis of fractures in an urgent care setting from the perspective of the NHS and Personal Social Services.
- The EAG divided the analysis into 3 separate fracture sites, focussing on fractures of the
    - wrist and hand
    - ankle and foot
    - hip
- These 3 sites were considered to potentially gain the greatest benefit from AI-assisted diagnosis.
- Separate models were used because costs and consequences of these fractures differed substantially

Outputs of these were then weighted based on the fracture case mix of a typical urgent care setting to estimate the overall cost-effectiveness of AI-assisted diagnosis.

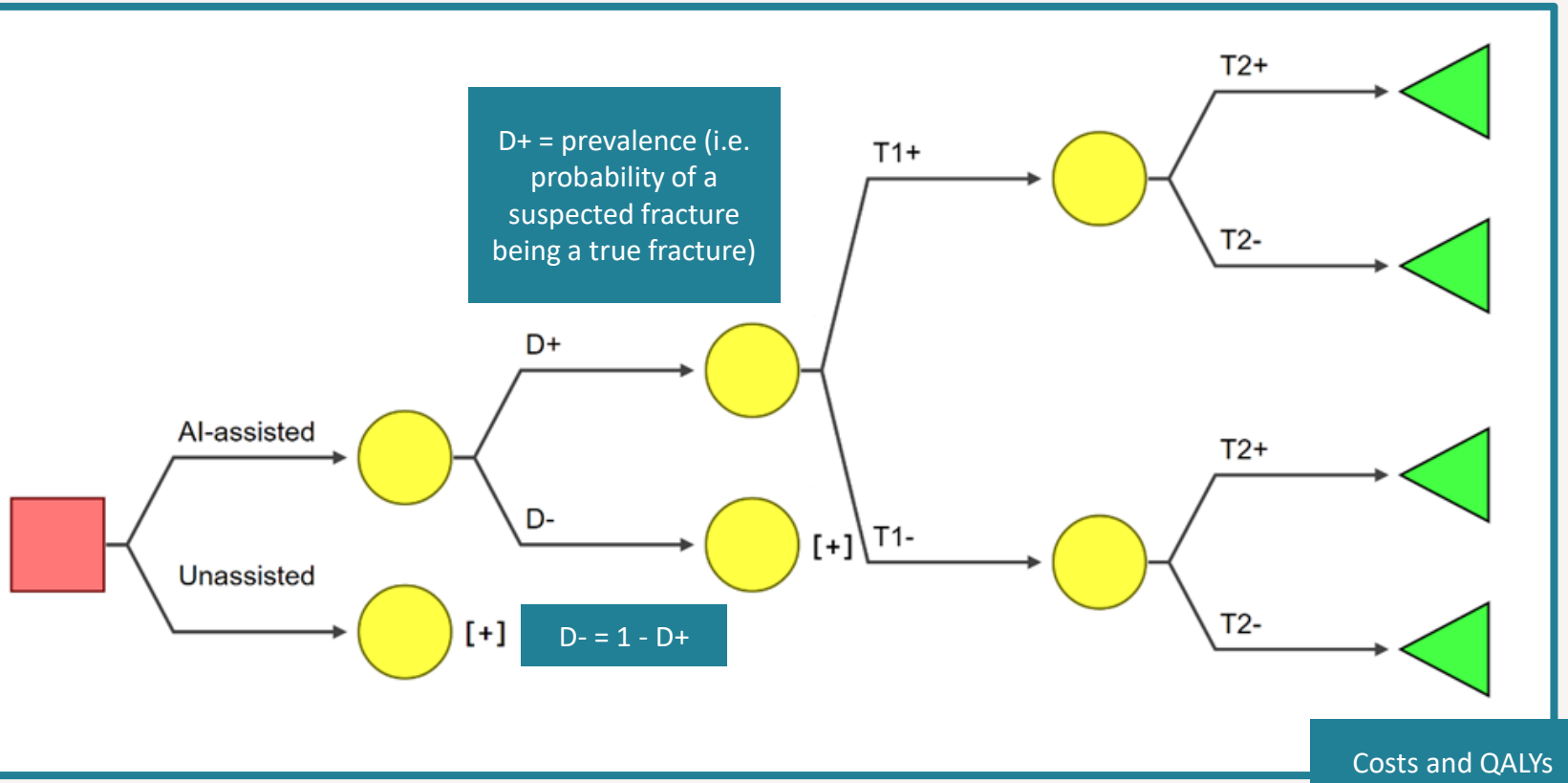
Economic model (2)

- The EAG said that the model described represents a rapid overview of the likely costs and consequences associated with use of AI algorithms to assist in the diagnosis of fracture, and unassisted diagnosis, in an urgent care setting.
- The purpose of this analysis was to explore whether there was a plausible case for any of the technologies to represent value for money for the NHS / taxpayer and to identify where further evidence generation may improve the certainty of the results.

Economic model structure

T+ = conditional probability of a positive result from review of X-ray (in branch shown this is sensitivity). Model allows for 1 or 2 reviews.

- The model structure comprised a decision tree incorporating the prevalence, sensitivity and specificity and cost per diagnosis of a strategy
- Base case assumed that patients present in an emergency department with a suspected fracture: ankle or foot, wrist or hand, and hip.
- Costs of ED attendance and X-ray common to all comparators and so was excluded from the analysis.
- Difference in cost between techs limited to cost per scan.



Square = decision node; Circle = chance node; Triangle = terminal node.

For further details see section 8.3.1 in the EAR

Base case prevalence, sensitivity and specificity

Two of the studies[1,2] identified in the review were selected as sources for prevalence and sensitivity and specificity of each diagnostic strategy:

Bousson et al. (2023) provided a directly comparison between BoneView, Rayvolve and TechCare Alert. Data were disaggregated by foot, ankle and hand (but not hip). For the base case, the EAG estimated a mean sensitivity and specificity for foot and ankle, assumed hand applied equally to wrist, and assumed the sensitivity and specificity of hip fracture diagnosis was equal to that for 'all fractures'.

Bachmann et al. (2024) compared RBfracture assisted to unassisted diagnosis in a wide range of fracture types, reporting results by 'mixed' staff types, ED trainees and trauma care nurses. This was used as the source study for RBfracture and unassisted diagnosis. As data were not disaggregated by fracture type, the base case assumed the same sensitivity and specificity for all fracture types for RBfracture and unassisted diagnosis.

1. Bousson et al. 2023; 2. Bachmann et al. 2024

Base case prevalence, sensitivity and specificity

Fracture site	Prevalence	Source
Ankle/foot	24.1%	Bousson et al. 2023
Hand/wrist	30.9%	
Hip	21.7%	

Technology	Sensitivity (foot/ankle)	Specificity (foot/ankle)	Sensitivity (hand/wrist)	Specificity (hand/wrist)	Sensitivity (hip)	Specificity (hip)	Source/notes
BoneView	94	86	92	92	91	91	Bousson et al. 2023
Rayvolve	91	67	98	75	93	70	Bousson et al. 2023
TechCare Alert	88	91	94	92	90	93	Bousson et al. 2023
RBfracture	83	90	83	90	83	90	Bachmann et al. 2024 (A&E trainee)
Unassisted	74	87	74	87	74	87	Bachmann et al. 2024 (A&E trainee)

Modelling key assumptions

Assumption	EAG's rationale
AI-assisted diagnosis was limited to ankle/foot, wrist/hand and hip fractures	3 sites were considered to potentially gain the greatest benefit from AI-assisted diagnosis
Sensitivity and specificity for hand applied equally to wrist	EAG assumption
Single read per scan	Source studies estimating the sensitivity and specificity of diagnosis assumed a single read of a scan
Sensitivity and specificity of the diagnosis depends on the grade of staff reading the scan, rather than the setting where it took place	EAG assumption
Prevalence by fracture type was 12.5% ankle, 7.5% wrist and 12.5% hip. EAG assumed that these proportions represented the base case distribution	Clinical opinion

Model specific assumptions are described separately (see slides 51 to 53)

Foot and ankle fractures: Costs and QALYs

The time horizon for the foot and ankle model was 12 months.

Health state	Mean QALYs	Mean costs (2022/23 prices)	Model specific assumptions
True positives	0.722 (weighted average)	£1,837	50% treated with a brace and 50% with a cast. Ankle fractures would be healed within 12 months.
True negatives	0.796	No cost	Only incur the cost of an ED attendance (excluded) Endure the fracture health state utility (0.225) for 2 weeks, then revert to 'healed' health state utility (0.818) for the rest of the year
False positives	0.796	£1,837	Incur the same cost as true positives and the same QALYs as true negatives.
False negatives	0.697	£1,986	Re-present in the ED after 2 weeks, where additional investigations and correct diagnosis is made. Incur the same costs as true positives, plus cost of additional examinations.
Source	Nwankwo et al. (2023)		

The utility associated with ankle/foot fracture (0.225) was considered to lack face validity for a soft tissue injury and so was explored in a scenario analysis (see slide 56)

NICE

Hand and wrist fractures: Costs and QALYs

The time horizon for the hand and wrist model was 6 months.

Health state	Mean QALYs	Mean costs (2022/23 prices)	Assumptions
True positives	0.346	£1,166	Utilities for the first 3 months are equal to 1 standard deviation below those reported in Rua et al. 2020.
True negatives	0.393	£773	Utilities equal to those reported in Rua et al. 2020
False positives	0.393	£850	Resource use associated to this health state is 10% higher than that of true negatives
False negatives	0.329	£1,056	Patients return to urgent care 2 weeks after initial presentation and receive correct diagnosis. Utility from Rua et al. 2020 at baseline represents the utility for those 2 weeks. No disutilities assumed in that period.
Source	Rua et al. 2020		

Hip fractures: Costs and QALYs

The hip model had a lifetime time horizon.

Health state	Duration (weeks)	Mean utility	QALYs (discounted)	Mean costs	Assumptions
TP: immediate post fracture	2	0.42	9.275	£57,471	1 index surgery followed by a second surgery
TP: first year	50	0.59			
TP: beyond first year	884	0.69			
TN: immediate post-injury	2	0.42	10.751	No cost	Incur only cost of ED attendance (excluded).
TN: baseline utility for age 65-74 years	934	0.79			
FP: immediate post-injury	4	0.42	10.751	£57,471	Assumed to have surgery. Incur same cost as true positives and same QALYs as true negatives.
FP: first year	48	0.59			
FP: baseline utility for age 65-74 years	884	0.79			
FN: immediate post fracture	4	0.42	9.268	£57,961	Incur same costs as true positives with additional costs of an ED attendance and additional investigations (1 CT assumed).
FN: first year	48	0.59			
FN: beyond first year	884	0.69			
Source		Low et al. 2021		Judge et al. 2016	

TP, true positive; TN, true negative; FP, false positive; FN, false negative.

Overall impact of AI-assisted diagnosis in an urgent care setting

- Clinical advice to the EAG was that for 2022-23, there were approximately 25.3 million ED attendances in England and that fractures typically account for 2-4%
- Clinical opinion on the prevalence by fracture type was that around 12.5% are ankle, 7.5% wrist and 12.5% hip
- EAG used these proportions to estimate the overall impact of diagnosis in an urgent care setting with 350–400 daily attendances, which results in 1,334 attendances for fracture a year

Base case distribution of fracture types (annual)

Fracture type	Base case proportions	Number of attendances
Ankle/Foot	38.5%	513
Wrist/Hand	23.0%	308
Hip	38.5%	513
Total		1,334

NICE

Costs

- In the base case the cost per scan was based on 1,334 scans per year
- The cost of the index presentation and X-ray was excluded from the analysis as it is common to all arms
- For health states in which additional presentations occur (false negatives), a mean cost of £149.04 was used
- Cost per scan for TechCare Alert [redacted]

Technology	Cost per Scan	Notes
BoneView	£1.00	Notional cost
Rayvolve	£1.00	No data received. Notional cost
RBfracture	[redacted]	[redacted]
TechCare Alert	[redacted]	[redacted]
Unassisted	£0.00	By definition

NICE

Scenario analyses

Scenario	Change	Scenario value(s)
1	Sensitivity and specificity	Optimistic scenario assumed the lowest sensitivity and specificity for unassisted diagnosis and highest for each technology, based on a review of all source studies. See table 38 in the EAR.
2		Pessimistic scenario assumed the highest sensitivity and specificity for unassisted diagnosis and lowest for each technology, based on a review of all source studies. See table 38 in the EAR.
3	Volume-based cost per scan	Low volume, high cost for those technologies with pricing based on volume. See table 39 in the EAR
4		High volume, low cost for those technologies with pricing based on volume. See table 39 in the EAR
5	Reading time based on Registrar grade reader	10 second reduction and 13.9p saving per X-ray. See table 40 in the EAR
6	Reading time based on Consultant reader	10 second reduction and 30.3p saving per X-ray. See table 40 in the EAR
7	Utility values for true negative ankle and foot fractures	0.727. Equivalent to EQ5D utility for a person with some mobility problems and some pain (see pages 112 to 113 in the EAR).
8	Use case – all fractures	Pessimistic scenario, additional cost for fractures in other places but no additional benefit. See table 41 in the EAR.
9	Number of reads	Second read of all X-rays rather than single read as used in the base case. See page 113 in the EAR

Base case results: Overall

Intervention	Cost (95%CI)	QALYs (95%CI)	INHB20k (95%CI)
BoneView	£6,901 (£5,099, £8,868)	4.41 (2.344, 6.187)	0.032 (-0.003, 0.072)
Rayvolve	£10,486 (£7,848, £13,306)	4.41 (2.344, 6.187)	-0.148 (-0.207, -0.096)
RBfracture	[redacted]	4.41 (2.344, 6.187)	[redacted]
TechCare Alert	[redacted]	4.41 (2.344, 6.187)	[redacted]
Unassisted	£7,515 (£5,534, £9,676)	4.41 (2.343, 6.186)	-

- Overall, with the exception of Rayvolve, the AI-assisted diagnostic algorithms were associated with a positive incremental net health benefit compared with unassisted diagnosis at £20,000 and £30,000 thresholds.
- 95% confidence intervals in most cases crossed zero, both for all separate fracture sites/types and when considered together.
- Due to data limitations, the EAG advised against direct comparisons between different AI algorithms

NICE

See table 45 in the EAR for further details including INHB30k results. Abbreviations: CI, confidence interval; INHB20k, incremental net health benefit at willingness to pay threshold of £20k;
QALY, Quality Adjusted Life Year

Base case results: Ankle and foot

Intervention	Cost (95%CI)	QALYs (95%CI)	INHB20k (95%CI)
BoneView	£638 (£518, £766)	0.786 (0.773, 0.798)	0.001 (-0.003, 0.006)
Rayvolve	£903 (£773, £1,043)	0.786 (0.773, 0.798)	-0.012 (-0.018, -0.007)
RBfracture	[redacted]	0.785 (0.772, 0.797)	[redacted]
TechCare Alert	[redacted]	0.785 (0.773, 0.798)	[redacted]
Unassisted	£634 (£519, £758)	0.784 (0.772, 0.797)	-

- There was minimal difference in QALYs between the different technologies and unassisted diagnosis.
- Costs varied more than QALYs for ankle and foot fractures
- Only Rayvolve had a significantly different cost to unassisted reads

NICE

See table 42 in the EAR for further details including INHB30k results. Abbreviations: CI, confidence interval; INHB20k, incremental net health benefit at willingness to pay threshold of £20k;
QALY, Quality Adjusted Life Year

Base case results: Wrist and hand

Intervention	Cost (95%CI)	QALYs (95%CI)	INHB20k (95%CI)
BoneView	£897 (£807, £989)	0.398 (0.386, 0.409)	0.000 (-0.001, 0.002)
Rayvolve	£908 (£827, £990)	0.398 (0.386, 0.409)	0.000 (-0.002, 0.002)
RBfracture	[redacted]	0.397 (0.386, 0.409)	[redacted]
TechCare Alert	[redacted]	0.398 (0.386, 0.409)	[redacted]
Unassisted	£893 (£808, £979)	0.397 (0.386, 0.408)	-

- There was minimal difference in costs and QALYs between the different technologies and unassisted reads.

NICE

See table 43 in the EAR for further details including INHB30k results. Abbreviations: CI, confidence interval; INHB20k, incremental net health benefit at willingness to pay threshold of £20k;
QALY, Quality Adjusted Life Year

Base case results: Hip

Intervention	Cost (95%CI)	QALYs (95%CI)	INHB20k (95%CI)
BoneView	£16,762 (£14,993, £18,640)	10.431 (5.660, 13.075)	0.080 (-0.010, 0.179)
Rayvolve	£25,806 (£23,810, £27,845)	10.431 (5.660, 13.075)	-0.372 (-0.481, -0.259)
RBfracture	[redacted]	10.431 (5.659, 13.075)	[redacted]
TechCare Alert	[redacted]	10.431 (5.660, 13.075)	[redacted]
Unassisted	£18,363 (£16,179, £20,612)	10.431 (5.659, 13.075)	-

- There was very little difference in QALYs between assisted and unassisted reads for hip fracture.
- Rayvolve had a significantly higher cost than unassisted reads.

NICE

See table 44 in the EAR for further details including INHB30k results. Abbreviations: CI, confidence interval; INHB20k, incremental net health benefit at willingness to pay threshold of £20k;
QALY, Quality Adjusted Life Year

Base case results: maximum economically justified price

EJP threshold	BoneView	Rayvolve	RBfracture	TechCare Alert	Unassisted
£20k/QALY	£632	-£2,956	[redacted]	[redacted]	0
£30k/QALY	£640	-£2,948	[redacted]	[redacted]	0

- The EAG also calculated the maximum economically justifiable price for each of the technologies. This is the maximum cost per scan that would still result in a technology being cost effective with an ICER of £20k or £30k per QALY gained.
- The minimum economically justified prices were generally higher than the per-scan prices proposed by the companies and used in the base case model (see slide 55).

NICE

See table 47 in the EAR for further details including INHB30k results. Abbreviations: EJP, economically justifiable price; ICER, incremental cost effectiveness ratio

Scenarios results

- Only scenarios adjusting diagnostic accuracy had a large impact on model results.
- Other scenarios including low and high volume based pricing, reduction in reading times, higher utility values for true negative ankle/foot fractures, use across all fractures, and adding a second read did not affect the model results.
- Full scenario analysis results are presented in table 48 (pages 117 to 120) in the EAR.

Optimistic diagnostic accuracy

Intervention	INHB20k	INHB20k 95%CI
BoneView	[redacted]	[redacted]
Rayvolve	[redacted]	[redacted]
RBfracture	[redacted]	[redacted]
TechCare Alert	[redacted]	[redacted]
Unassisted	-	-

Pessimistic diagnostic accuracy

Intervention	INHB20k	INHB20k 95%CI
BoneView	[redacted]	[redacted]
Rayvolve	[redacted]	[redacted]
RBfracture	[redacted]	[redacted]
TechCare Alert	[redacted]	[redacted]
Unassisted	-	-

CI, confidence interval; INHB20k, incremental net health benefit at willingness to pay threshold of £20k; QALY, Quality Adjusted Life Year

Summary of economic evidence

- Early economic modelling suggests that most of the AI technologies considered have the potential to be cost effective.
- Most of the AI technologies had a positive INHB at £20k and £30k thresholds although 95% confidence intervals in most cases crossed zero
- The lower specificity of Rayvolve leads to higher costs dues to more people being referred for further investigation
- EAG noted that the potential cost-effectiveness appeared to be driven by reductions in costs rather than a gain in QALYs
- EAG cautions against using this analysis to compare one AI algorithm against another due to data limitations

Evidence gaps and research recommendations

Key evidence gaps identified by the EAG included:

- Lack of prospective, consecutively sampled, comparative studies based in clinical settings comparable to the NHS urgent care setting, with staff/reader groups that would typically perform the initial interpretation.
- Studies designed to explore changes in outcomes according to key factors that would inform use of the technology, such as reader experience, fracture case mix, and determinants of patient outcomes, such as patient age, frailty, and prevalence of health conditions affecting bone health.
- Longer term costs and consequences of missed fracture diagnoses
- System level outcomes such as number of referrals to virtual fracture clinics or time spent in ED

NICE National Institute for Health and Care Excellence

Health Tech Programme

Artificial intelligence software to help detect fractures in the emergency department (provisional title)

Professional organisation submission

Thank you for agreeing to give us your organisation's views on this technology and its possible use in the NHS.

You can provide a unique perspective on the technology in the context of current clinical practice that is not typically available from the published literature.

To help you give your views, please use this questionnaire. You do not have to answer every question – they are prompts to guide you. The text boxes will expand as you type.

Information on completing this submission

- Please do not embed documents (such as a PDF) in a submission because this may lead to the information being mislaid or make the submission unreadable
- We are committed to meeting the requirements of copyright legislation. If you intend to include **journal articles** in your submission you must have copyright clearance for these articles. We can accept journal articles in NICE Docs.
- Your response should not be longer than 13 pages.

Any confidential information provided should be underlined and highlighted. Please underline all confidential information, and separately highlight information that is 'commercial in confidence' in blue and all that is 'academic in confidence' in yellow.

About you **1. Your name**	Tracy O'Regan
2. Name of organisation	The Society of Radiographers
3. Job title or position	Professional officer clinical imaging and research
4. Are you (please select Yes or No):	An employee or representative of a healthcare professional organisation that represents clinicians? Yes
5a. Brief description of the organisation (including who funds it).	For over 100 years, the Society of Radiographers (SoR) has advocated for radiography professionals at the heart of patient care. SoR is both a trade union and the UK professional body for the diagnostic imaging and radiotherapy workforces including diagnostic and therapeutic radiographers, sonographers, support workers, assistant practitioners, and pre-registration students.
5b. Has the organisation received any funding from any company with a technology included in the evaluation in the last 12 months? [Please refer to the final scope for a full list of technologies included. The final scope is due to be published on 2 July 2024]. **If so, please state the name of company, amount, and purpose of funding.**	No.

5c. Do you have any direct or indirect links with, or funding from, the tobacco industry?	No

The aim of treatment for this condition

6. What is the main aim of this technology? (For example, initial diagnosis, clinical monitoring, treatment triage assessing stages of disease progression or risk stratification.)	Clinical decision support to diagnostic radiographers and healthcare professional colleagues practicing in urgent care (preliminary/initial diagnosis).
7. In your view, is there an unmet need for patients and healthcare professionals in this condition?	Yes, in most healthcare settings – primarily where there is an absence of 24hour availability of hot reporting (immediate definitive clinical report from radiology services – reporting radiographer, MSK sonographer, or radiologist).

What is the expected place of the technology in current practice?

8. How is the condition currently treated in the NHS?	The technology does not apply to a single 'condition', body part, or type of fracture. It potentially covers the whole range of MSK fractures to appendicular and axial skeleton across the life-course of patients' and possibly beyond, when a patient does not survive life threatening injuries. Treatment varies according to type of fracture, service, and patient preference.

9a. Are any relevant clinical guidelines we should be aware of, and if so, which?	Yes, the relevant NICE and GIRFT guidance has been listed in the online document, July 2024 NICE Artificial intelligence software to help detect fractures on X-rays in urgent care - Final scope.
9b. Is the pathway of care well defined? Does it vary or are there differences of opinion between professionals across the NHS? (Please state if your experience is from outside England.)	SoR represents professionals who work across the 4 nations of the UK. In all nations the pathways of care are dependent on local service provision, staffing, facilities, and need of local populations. Urgent care is provided in various ways; in the main that is not due to differences of opinion between professionals. Instead, there are different models due to system pressures and local population needs. Broadly those models include the same staffing groups mentioned in the final scope although we note that while some of our Allied Health Professional (AHP) colleagues – physiotherapists - are considered in the final scope and protocol for this project/programme there is lack of reference to treatment and care provided in acute urgent settings by other AHP colleagues. Particularly occupational therapists and potentially social workers who can assess patients with fracture for discharge planning.
9c. What impact would the technology have on the current pathway of care?	That is dependent on the performance of the technology in terms of accuracy and efficiency. Different AI products appear to have different levels (or claims) to accuracy and efficiency, provide varying services, for varying populations. This is clearly recognised and has been considered in detail by the NICE specialist committee members, at the NICE June 2024 scoping workshop, and by the external assessment group. SoR hope that the main impact to care will be reduction of inaccurate diagnosis resulting from lack of access to immediate definitive diagnosis from radiology services (reporting radiographer or radiologist). There is, however, the potential for bias, over-diagnosis of fracture, or incorrect diagnosis from poorly

	performing AI algorithm. The current requirements of Ionising Radiation (Medical Exposure) Regulations (IRMER) mitigate for this with the need for oversight from a healthcare professional who is trained and qualified to provide that oversight.
10a. Will the technology be used (or is it already used) in the same way as current care in NHS clinical practice?	Yes, although if proven reliable and trustworthy in real world settings, there will be further potential for safe, accurate, and efficient systems to provide clinical decision support that enables innovation in provision of services in the future. This has a potential impact on roles of diagnostic radiographers in the future and may allow for development of radiographer discharge from urgent care for which there is some evidence of positive impact for patient care, outcomes, and emergency department waiting times.
10b. How does healthcare resource use differ between the technology and current care?	It is too early to provide a definitive answer to this question because of the lack of real-world prospective trials for this technology. We note that several studies are ongoing, but this has been limited in scope and breadth of populations to date. We can make inferences from espoused potential, that the technology/software will reduce the inappropriate use of orthopaedic services/fracture liaison clinic and reduce the use of other imaging modalities/repeat imaging and recall to urgent care. It will not reduce the need for radiology services to provide a definitive clinical report (majority of which is reporting radiographer provision) in line with IR(ME)R and noting that AI algorithms listed in the programme do not assess for the full scope of pathologies that a reporter will highlight, including, for example, primary bone tumour, infection, foreign body, normal variants.

10c. In what clinical setting should the technology be used? (For example, primary or secondary care, specialist clinics.)	The final scope for the project notes that assessment is in line with urgent care. That is reasonable given the stages of development of current AI algorithms for AI fracture detection. It is likely that once proven, this type of service would also be useful to services in community diagnostic centres, for assessment of injury sustained on inpatient wards, and for patients attending radiology directly from primary care referral with suspected fracture. SoR consider that alongside those further clinical settings for use of the technology, the technology and patient pathway may also be tailored to enable direct referral of patients to fracture liaison clinic for those people at further risk due to bone health or lifestyle factors.
10d. What investment is needed to introduce the technology? (For example, for facilities, equipment, or training.)	That is a difficult question to answer given the variability of digital services and architecture across England and the devolved nations. English imaging networks are classed by NHS England at variable levels of digital maturity. That has an impact on the costs to set up, deploy, and monitor the implementation of AI for MSK fracture detection. Onward surveillance will also have additional cost implications with questions about frequency of software update and service agreements, funding for collection of data and analysis/evaluation etc. NHSE AI Deployment Fund (AIDF) 2023-2025 will provide a comparator to some extent. We are aware that NICE do have representation on AIDF implementation board for lung cancer AI as do SoR, National Institute for Health Research, and Royal College of Radiologists. This is important to note

	because this will allow the extrapolation and estimation of potential costs, capital, revenue, ongoing licensing etc. with input from various stakeholders and different lenses. NIHR / AIDF Rapid Service Evaluation Team (RSET) will also provide further evidence from the point of lung cancer AI tools (again, representation on the stakeholder panel from the NICE team noted). In addition, investment is needed for education and training of all staff who use the AI software. Investment is needed for thorough standardised research/evaluation of products, staff to lead development, and dissemination of findings. Investment is needed in literature and methods to inform patients about the use of AI in order that their fully informed consent to treatment is valid.
11a. Do you expect the technology to provide clinically meaningful benefits compared with current care?	Yes. SoR are aware that although 24-hour immediate definitive clinical diagnosis (hot report) is the gold standard of care that services are advised / would like to provide, this is rarely achieved. Clinically meaningful benefits will be in the form of reduction of missed fracture / delayed diagnosis in urgent care settings.
11b. Do you expect the technology to increase length of life more than current care?	Very rarely.

11c. Do you expect the technology to increase health-related quality of life more than current care?	Yes, the main benefits of which are likely to be the reduction of pain and avoidable consequences of missed fractures. There will also be a public health benefit in terms of reduction of repeat imaging / use of onward CT imaging results in reduced radiation dose to individuals. It is important to point out that the missed diagnosis or delayed diagnosis of fractures does have an affect on the level of trust that patients, carers, and wider public place in healthcare systems. This is extremely important in relation to the prediction and prevention of disease. Patients' level of trust in the healthcare system, in imaging diagnosis and services, is linked to their likelihood to attend for healthcare including screening services for example mammography, abdominal aortic aneurysm, DEXA/DXA etc. All relevant for quality of life with early diagnosis of disease.
12. Are there any groups of people for whom the technology would be more or less effective (or appropriate) than the general population?	Not in addition to those already listed in the final scope although we note that there is slight discrepancy, between final scope and final protocol for this project. While scope refers to older people the final protocol highlights frailty – SoR agree with this approach (final protocol) since not all older people are frail and not everyone with frailty is older. This is more inclusive of people with disabilities/immobility likely to affect bone health.

	In the case of AI products, we would also like to highlight the importance of knowledge of geographical variations and local knowledge of regional pathologies affecting populations. For example, the prevalence of Pagets disease affecting the appendicular and axial skeleton in Northern English male populations would be relevant for deployment of AI in that region.

The use of the technology

13. Will the technology be easier or more difficult to use for healthcare professionals than current care? Are there any practical implications for its use (for example, additional clinical requirements, factors affecting patient acceptability or ease of use or additional tests or monitoring needed.)	Because the AI products are only implemented in a small number of areas currently, it is difficult to be exact in this reply. It appears that when systems integrate fully with digital systems in use, when people understand how to use them and what the limitations are, they know what to do when there are errors or discrepancies and understand the governance of the products, when they trust the technology – then there are likely to be positive adaptations to practice. It appears that currently there is no reduction in workload stresses for staff, perhaps even increase in workload due to implementation, but this is common with all technological development. It may smooth the way for increased efficiencies, accuracy and innovation in the future.
14. Do you consider that the use of the technology will result in any substantial health-related benefits that are unlikely to be included in the quality-	Level of trust in healthcare services, radiology and urgent care which can have an impact on attendance rates to screening, diagnosis and treatment in the future.

adjusted life year (QALY) calculation?	
15. Do you consider the technology to be innovative in its potential to make a significant and substantial impact on health-related benefits and how might it improve the way that current need is met?	No, not innovative in what the technology provides for health-related benefits. If healthcare services were adequately funded and staffed to provide 24 hour immediate definitive report (hot reporting) then there would not be a requirement for the technology. The capability of AI products to determine MSK fracture on X-rays does not exceed the capability of people-reporting radiographers and radiologists-to diagnose fractures. Also, radiographers and radiologists still need to assess patient's images for a plethora of associated disease, pathology, and conditions that are visible on images but AI software does not assess for. This does not reduce their workload in the long term, rather it may speed up the initial diagnosis or ruling out of a fracture during the individual patient's attendance in urgent care.
16. Does the use of the technology address any particular unmet need of the patient population?	There are currently avoidable inequalities in access to definitive clinical diagnosis for individuals. That is due to either the services available to the person in the areas where they live or the day of week/weekend, or time of day that they attend urgent care. The technology may mitigate for that inequality to some extent.
17. Are there any side effects or adverse effects associated with the technology and how do they affect the patient's quality of life?	No change to imaging procedure currently (imaging acquisition). In current care there is a margin of error in fracture diagnosis, and it is recognised that occult fractures will not be determined until a later time. Overdiagnosis of fracture (false positive) may have a limited affect on quality of life for a short period, only rarely if ever would that have long term affect on quality of life.

Sources of evidence

17a. Do studies on use of the technology reflect current UK clinical practice?	Studies have been retrospective not proven in real-world practice. We recognise that there are studies in progress, but this is still limited in comparison to the range of populations and variation across regions/nations also the range of types of fracture or axial, appendicular skeletal body region.
17b. If not, how could the results be extrapolated to the UK setting?	Theories of implementation science and change management indicate that the intricacies of healthcare and importantly local culture (among healthcare professionals) in the UK setting must be considered for the successful deployment of new technology. To date, few UK based/published AI studies have considered beyond retrospective comparison of accuracy and sensitivity of AI in comparison with clinical reporters.
17c. What, in your view, are the most important outcomes, and were they measured in trials?	---
17d. If surrogate outcome measures were used, do they adequately predict long-term clinical outcomes?	---
17e. Are there any adverse effects that were not apparent in clinical trials but have come to light subsequently?	There are anecdotal accounts from reporting radiographers that the deployment if AI for fracture detection in real-world settings has resulted in increased workload for fracture clinics and orthopaedic teams, with AI software not able to recognise adapted technique positioning, normal variants, or pathologies including arthritis and infection and inexperience healthcare professionals not recognising

	those limitations. Out of hours, without a reporters guidance, referring staff who have trialled the software do not always recognise those errors and refer patients to clinics. More experienced staff (in terms of experience and confidence reading imaging) appear to be more confident to overrule the AI results.
18. Are you aware of any relevant evidence that might not be found by a systematic review of the trial evidence?	The experiences of reporting radiographers and radiologists who are using AI MSK fracture detection software have been reported in various methods of shared learning with colleagues. For example, continuing professional development study days/evening, shared learning sessions, poster and oral presentations at conferences, discussion among professional body advisory groups including SoR, British Institute of Radiology, and Royal College of Radiologists.
19. How do data on real-world experience compare with the available data? Are you aware of any ongoing studies?	Not beyond the ongoing studies noted in the programme scope.

Equality

20a. Are there any potential equality issues that should be taken into account when considering these technologies?	The consideration of equalities issues in both the scope and protocol documents is excellent. The scope and range of people who are included in this assessment is immense; although it would be possible to delve further into certain aspects of inequality, for example, for people with certain conditions related to bone health, alcoholism, eating disorders etc but that is not appropriate given the scope, protocol and timescale for the project
20b. Consider whether these issues are different from issues with current care and why.	The additional factors mentioned above are routinely considered in usual care, different from the abilities of AI software, because healthcare professionals and trained reporters correlate clinical history, mechanism of injury, clinical presentation in the urgent care setting, with the images that they are reading. Not all healthcare professionals viewing images have the breadth and depth of knowledge to do that. AI software is also not at a stage that it is able to do that. If that is possible in the future, that would be a further improvement to the output of AI as a clinical decision support tool to the referrer viewing an image in urgent care – more able to consider variability and personalise with tailored diagnosis. In terms of organisational and service level inequality, there is potential for the use of AI software to decrease current inequalities in access to diagnosis in centres where hot reporting / preliminary clinical evaluation is not running – either due to service design, day of week, or time of day along with issues related to the ability to accurately read the images across various staff groups or level of competence.

Key messages

21. In up to 5 bullet points, please summarise the key messages of your submission.	• SoR are supportive of the implementation, research, and evaluation of AI for MSK fracture detection software. • SoR highlight the importance of training and education of staff who will use or read the outcomes of the software, especially with respect to the limitations of the software. • SoR consider it imperative that resources are available to staff using the software in order that they may ensure patient consent is fully informed. • SoR view the development of AI systems within clinical imaging as a step toward further innovation of services for patients. This will involve the development of professionals to undertake new roles, for example, to lead in the implementation and surveillance of AI software, to develop that ways in which patient pathways of care can be improved, including earlier diagnosis and appropriate treatment, onward referral, or discharge of patients. • SoR additionally would like to highlight that the use of AI software for MSK fracture detection should also be regarded as a tool for public health, early diagnosis, prediction and prevention of further fracture for patients with development of links to fracture liaison clinics.

Thank you for your time.

Please log in to your NICE Docs account to upload your completed submission.

Your privacy

NICE National Institute for Health and Care Excellence

The information that you provide on this form will be used to contact you about the topic above.

Please select YES if you would like to receive information about other NICE topics - YES

For more information about how we process your personal data please see our privacy notice.

Artificial intelligence software to help detect fractures on X-rays in urgent care: An Early Value Assessment
EAG assessment report (Post-FAC)

Produced by	Peninsula Technology Assessment Group (PenTAG) University of Exeter Medical School
Authors	Caroline Farmer[1] Helen Coelho[1] Madhusubramanian Muthukumar[1] Sophie Robinson[1] Robert Meertens[2] Obioha C. Ukoumunne[1,3] Val Santo[1] Niamh Gale[2] Jonathan T Evans Jonathan Evans Jenny Lowe[1] G.J. Melendez-Torres[1] Edward C.F. Wilson[1] [1] Peninsula Technology Assessment Group (PenTAG), University of Exeter Medical School, Exeter [2] Department of Health and Care Professions, University of Exeter Medical School, Exeter [3] NIHR Applied Research Collaboration South-West Peninsula (PenARC), University of Exeter Medical School, Exeter 4 Department of Public Health and Sports Science, University of Exeter Medical School, Exeter.
Correspondence to	Caroline Farmer 3.09 South Cloisters, St Luke's Campus, Heavitree Road, Exeter, EX1 2LU; c.farmer@exeter.ac.uk
Date completed	11/09/2024

Produced by	Peninsula Technology Assessment Group (PenTAG) University of Exeter Medical School
Source of funding	This report was commissioned by the NIHR Evidence Synthesis Programme as project number NIHR136024
Declared competing interests of the authors	Dr Meertens has previously had contact with Qure.AI in his role as Director of Business Engagement and Innovation at the University of Exeter. Qure.AI visited the University of Exeter campus to deliver sessions to students and clinical radiographers. Dr Meertens has no financial interest or ongoing research relationship with Qure.Ai. No other conflicts.
Acknowledgments	The authors acknowledge the administrative support provided by Mrs Sue Whiffin (PenTAG) and data extraction completed by Mr Alex Allen.

Author contributions

Caroline Farmer	Clinical evidence lead and project manager. Input into the writing of assessment documents.
Helen Coelho	Literature screening, data extraction, and clinical and service evidence assessment. Writing of assessment documents.
Madhusubramanian Muthukumar	Literature screening, data extraction and health economic modelling. Writing of assessment documents.
Sophie Robinson	Developed and coordinated the evidence search strategy. Writing of assessment documents.
Robert Meertens	Literature screening and clinical advice.
Obioha C. Ukoumunne	Feasibility assessment for meta-analysis and calculation of diagnostic accuracy data.
Val Santo	Health economic modelling and writing in the assessment report.
Niamh Gale	Clinical expert advice and review of the assessment report.
Jonathan T Evans	Clinical expert advice and review of the assessment report.
Jonathan Evans	Clinical expert advice and review of the assessment report.
Jenny Lowe	Document management expert with responsibilities for supporting the conduct of evidence searches. Administrative support throughout the project.
G.J. Melendez-Torres	Literature screening and input into the writing of assessment documents.
Edward C.F. Wilson	Project director, economic evidence lead, and guarantor of the EAG assessment. Input into the writing of assessment documents.

This report should be referenced as follows: Farmer, Coelho, Muthukumar, Robinson, Meertens, Ukoumunne, Santo, Gale, Evans, Evans, Melendez-Torres, Wilson. Artificial intelligence software to help detect fractures on X-rays in urgent care: An Early Value Assessment. EAG assessment report. Peninsula Technology Assessment Group (PenTAG), 2024.

Artificial intelligence software to help detect fractures on X-rays in urgent care: Early Value Assessment

External Assessment Report: Changes after stakeholder consultation

Editorial corrections not tabulated.

Section	Description of change
4.2.1	Corrected ‘unassisted’ to ‘assisted’
4.2.1	Clarified description of Fu 2024 study design
Table 9	Corrected unassisted diagnostic accuracy for Nguyen 2022
Table 11	Corrected specificity for “Unassisted, mixed or unclear staff, hand/wrist”
Table 21	Standard deviations removed for Durations 1 and 2
8.3.4	Costs specified in text for False Negatives
Table 22	Added new table showing calculation of QALYs for false negative ankle/foot fractures
Tables 24, 26, 28, 30	Costs for hand/wrist corrected
Table 32	Costs for secondary prevention care following hip fracture corrected
Table 35	Costs for false negative hip fractures corrected
Tables 37, 39	Costs for RBFracture corrected [redacted]
Tables 42 – 48	Outputs of model updated following amends as described above [some redacted]
8.5	Updated text in response to consultee comment
Appendix D	New scenario added to investigate impact of set-up costs.

Table of Contents

List of tables

Abbreviations

Term	Definition
AI	Artificial Intelligence
CI	Confidence interval
DHSC	Department of Health and Social Care
DICOM	Digital Imaging and Communications in Medicine
EAG	External Assessment Group
ED	Emergency Department
ICER	Incremental Cost Effectiveness Ratio
INHB	Incremental Net Health Benefit
IQR	Interquartile range
MAUDE	Manufacturer and User Facility Device Experience
MHRA	Medicines & Healthcare products Regulatory Agency
MIU	Minor Injuries Unit
NA	Not applicable
NHB	Net Health Benefit
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NPV	Negative Predictive Value
NR	Not reported
PACS	Picture Archiving Communications System
PPV	Positive Predictive Value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QALY	Quality Adjusted Life Year
RCT	Randomised controlled trial
RFI	Request for Information
SD	Standard deviation
UTC	Urgent treatment centre

Executive Summary

Background and objectives

Plain film radiography or X-ray is the most common medical imaging approach used to detect fractures in urgent care settings, including the emergency department (ED), urgent treatment centre (UTC), and minor injuries units (MIU). X-rays are typically read in urgent care settings by healthcare professionals who are not radiology specialists or are inexperienced at interpreting X-rays, which may increase the likelihood of errors in decision-making, particularly in busy healthcare centres when staff are under significant pressure. Reduced staff numbers, such as outside normal working hours, may also influence the risk of errors in diagnosis. A definitive diagnosis of the injury will be produced by a consultant radiologist or reporting radiographer, although there may be a delay before this is available, meaning that this may arrive after people have been treated and/or discharged from urgent care. Delays vary across settings, and may be longer for children due to availability of specialist in paediatrics.

Artificial intelligence (AI) algorithms have been developed to support clinicians in diagnosing fractures, with the intention to improve the diagnostic accuracy of clinicians reviewing X-rays. Improving diagnostic accuracy means reducing the number of missed fractures (false negative diagnoses) and the number of people treated for a fracture who don't have one (false positive diagnoses).

The purpose of this rapid early value assessment (EVA) was to identify the existing evidence base for the technology and to assess whether there was a *prima facie* case for the technology to represent a value for money investment for people in the NHS. A rapid evidence review was conducted followed by 'light touch' early economic modelling to explore whether a plausible case could be made for cost-effectiveness at the prices charged by the companies. The approach was not suitable for a definitive assessment of the cost-effectiveness of one AI-algorithm against another, but rather to inform whether or not the NHS should consider adopting the technology whilst further evidence is collected.

Evidence review – clinical and service use outcomes

A broad evidence review was conducted to identify the existing evidence base for clinical, and service outcomes associated with the technology. The review identified 16 studies that evaluated the diagnostic accuracy of the technology as an aid to diagnosing fractures (i.e. when

used to assist reading clinicians, and not as a standalone diagnostic tool). Evidence was available for four of the eligible technologies: BoneView, Rayvolve, RBFracture, and TechCare Alert. None of the included studies were conducted in the UK and all were associated with limitations, including risks of bias and uncertain generalisability to the NHS. Few studies evaluated the technologies when used by clinicians who would typically provide the initial diagnosis in urgent care settings, with most evaluating readers who were clinicians specialising in radiology or amongst a varied of group of clinicians with varying levels of reading experience. Data were reported for a general sample of people with types of fractures that were eligible for consideration by the technologies. Subgroup data were also available for pre-specified fracture subgroups, for children and for 'less obvious' fractures. None of the included studies reported clinical outcomes associated with use of the technology and, aside from the reading time per scan, no service outcomes were reported. As compared to the list of outcomes specified in the NICE scope for this assessment, there was a major gap in the evidence base.

There was unexplained heterogeneity in the results reported across studies. To aid with interpretation of the results, where multiple results were reported by studies according to staff experience, the EAG grouped the data according to reader experience (as described by the included publications). This approach was considered imperfect and did not completely resolve the heterogeneity in the data. The EAG conducted a feasibility assessment to determine if meta-analysis of the data was possible, but where sufficient numbers of studies were available, these were considered too heterogeneous to pool. Notably, clinical advice to the EAG was that the diagnostic accuracy of unassisted readings in the included studies appeared lower than was expected by the EAG's clinical advisors, which adds uncertainty to the generalisability of the evidence base.

Overall, given the limitations in the evidence base and the heterogeneity in the study results, the EAG did not consider that the evidence base was suitable to determine reliable estimates of the diagnostic accuracy of the technologies for assisting in the diagnosis of fractures. However, based on evaluation of the evidence base as a whole and specifically in studies reporting outcomes for clinicians based in emergency care settings, the EAG identified a general trend for the technology to result in an improvement in sensitivity (i.e. a reduction in missed fractures) with no or minimal improvement in specificity (i.e. no change in false positive diagnoses). Use of the technology was still associated with varying levels of missed fractures, however, particularly in 'less obvious' fractures, where the technology was considered to be of most potential value. Further evidence is needed to determine robust evidence for any improvement in sensitivity,

and to establish whether the additional fractures identified would result in meaningful clinical benefits for patients.

Economic evidence and analysis

The evidence review identified no available economic evaluations for the technologies. The EAG constructed a simple decision model to establish whether there was a *prima facie* case for AI-assisted diagnosis to represent a value for money investment for NHS patients. As the long-term costs and outcomes for different fractures were substantially different, the EAG divided the analysis into three decision problems, concerning ankle and foot, wrist and hand, and hip fractures. These were chosen on the basis of availability of data and their different downstream costs and consequences. An overall estimate of the costs and consequences for a typical urgent care setting was estimated based on case mix for the three fracture types, with extension to all fractures considered in scenario analysis.

The decision model was a decision tree incorporating prevalence, sensitivity, specificity and cost per scan for each of the five diagnostic strategies (four AI algorithms and unassisted diagnosis). Estimates of the long-term costs and QALYs accrued from a true and false positive and negative for each fracture type were extracted from the literature. The tree was rolled back to estimate the expected cost and QALYs accrued from each diagnostic strategy. Scenario analyses explored key uncertainties.

Overall, most of the AI-assisted algorithms were associated with a positive incremental net health benefit at willingness to pay thresholds of £20,000 and £30,000 per QALY gained. Due to data limitations, the EAG did not consider the analysis appropriate to compare technologies against each other, although this would be required in a more thorough analysis in future to ensure that the diagnostic accuracy of each algorithm was matched to its price.

The results were mostly robust to the scenario analyses considered with the exception of diagnostic accuracy, where none of the algorithms were associated with a positive incremental net health benefit (compared with unassisted diagnosis) in the pessimistic scenario.

Key points for decision makers

- While a reasonable number of studies have evaluated the diagnostic accuracy of the technology as an aid to the identification of fractures, very few studies are specific to emergency care settings and all were associated with significant limitations due to risk of

bias or uncertain generalisability. The existing evidence base was not sufficient to determine an approximate estimate of the diagnostic accuracy of the technology for its intended use.

- Across the evidence base as a whole, there was a trend for the technology to reduce missed fractures without any change to false positive diagnoses. However, based on the existing evidence, the EAG was unclear whether the additional fractures identified would translate into meaningful benefits for patients. While there are some fractures that, if missed, can result in significant harm to patients, stakeholders to this assessment also considered it plausible that the technology would improve diagnosis of more subtle fractures that may not require a change in management.
- The evidence suggested that use of the technology would not eradicate the risk of missed fractures, meaning it was likely that health services would need to continue to take precautions to avoid the risk of a missed fracture in clinical practice (e.g. precautionary treatment of high risk suspected fractures). This means that use of the technology had an unclear impact on healthcare resource use.
- On average, based on a simple decision model for this EVA, most of the AI algorithms considered represent a positive incremental net health benefit compared with unassisted diagnosis at NICE's conventional threshold of £20,000 to £30,000 per QALY. The evidence base was not sufficient to compare different algorithms against one another.
- Results were mostly robust to scenario analyses with the exception of diagnostic accuracy.

Plain Language Summary

X-rays are the usual method for diagnosing broken bones (fractures) in urgent care settings, including Accident and Emergency (A&E), urgent treatment centres (UTC), and minor injuries units (MIU). Artificial Intelligence (AI) technologies have been developed to assist in identifying fractures on X-rays, and PenTAG was commissioned to conduct an Early Value Assessment (EVA) to provide an initial view about whether licensed AI technologies could be used for fracture detection in urgent care while further evidence is developed.

A search was conducted to identify all of the evidence that had evaluated AI to assist in fracture detection, including published evidence and confidential data from AI companies. The review identified 16 studies that evaluated how accurate AI was when used to assist diagnosis and 5 studies that reported how AI changed the time needed to interpret an X-ray. None of the studies were based within the UK and most were conducted with staff different to those who would normally read X-rays in the NHS. This meant that it was not possible to identify a good estimate for how accurate AI would be if it was used in urgent care. Overall, there appeared to be early signs that AI could help to reduce missed fractures, but further research would be needed to confirm this. No studies were identified that evaluated how using AI affected outcomes for people with a suspected fracture (such as their health and mobility) or how using the AI affected time and costs for the health service (such as the number of repeat appointments needed).

To assess whether AI would be good value for money for the NHS, we developed an economic model that combined information on how well AI diagnoses fractures, the cost of using AI, and information on what happens to a patient once their fracture is correctly – and incorrectly – diagnosed (their quality of life and costs to the NHS). Overall, our findings were that most of the AI technologies appeared to be fairly priced for the estimated benefits. We explored uncertainties related to the data and assumptions in our analysis and found that our conclusions did not change most of the time.

1. DECISION PROBLEM

The decision problem for this assessment is described in the NICE scope and EAG comments and planned assessment methods are included in the protocol.

During its assessment, the EAG made the following minor adjustments to the planned methods outlined in the protocol. These were:

Definition of frailty: None of the included clinical effectiveness evidence reported the proportion of participants who were assessed as being frail, had experienced a frailty fracture, or reported outcomes specific to this group. Though one of the included economic studies (Beaupre 2020) stated that frailty was considered and assessed bone mineral density (BMD) for elderly people when needed, it did not define “frailty” and indicated that only age was used to classify fragility fractures.

To help characterise the prevalence of people with frailty who were included in the evidence, the EAG reported other metrics that were imperfect but approximate indicators of frailty in the sample, where reported. This included the proportion of participants aged ≥80 years and the proportion of injuries due to falls. These indicators were considered to be imperfect, however.

Reference standard: The EAG included evidence from studies that used a reference standard that did not match that described in the NICE scope and review protocol (i.e. definitive report from a consultant radiologist or reporting radiographer). These decisions were taken due to a paucity of high-quality evidence directly relevant to the decision problem and due to uncertainty surrounding the correspondence between staff grades in other healthcare systems compared to consultant radiologists and reporting radiographers in the NHS. Where reference standards were considered to be indirect or flawed in some way, this is highlighted in the report.

Screening and prioritisation of evidence: The review protocol specified that the EAG would prioritise the clinical and economic evidence and outcomes that best addressed the decision problem for this assessment. As there were numerous quality considerations across the available evidence, it was not possible for the EAG to select a group of robust studies for priority inclusion in the review. Accordingly, the EAG included all of the evidence identified that reported diagnostic accuracy data for the included technologies, despite its limitations. As discussed later in the report, the limitations with the evidence means that the EAG did not aim to determine the diagnostic accuracy of the technologies, but rather to provide an overview of the existing

evidence base, initial interpretations from the results from patterns across the data, and recommendations for future research. One study[1] was de-prioritised following identification as it did not report any of the priority outcomes in the protocol. This study assessed healthcare professional user acceptability of AI for the detection of fractures before and following its implementation in a healthcare setting.

Selection of clinical evidence to inform the economic model: While all studies were associated with key limitations in quality, the EAG sought to select estimates of prevalence, sensitvity and specificity that could be used to inform the economic analysis. The studies with the most robust data – relative to the evidence base as a whole – were selected, though these were nevertheless considered to be unreliable. Studies with the following design features were prioritised for selection of sensitivity and specificity: studies with larger sample sizes; studies reported in peer-reviewed publications; studies that reported results for both AI assisted and unassisted clinicians; and studies that used a reference standard that did not include the results of the AI. Prevalence data selected for use in the economic analysis was derived from studies with: robust sample selection processes; larger samples; and relevant eligibility criteria. As this was an Early Value Assessment, these studies were not subjected to a formal quality appraisal, however key limitations of included studies are discussed within the report.

Information sought from companies. The EAG submitted clarification questions to four of the relevant companies (Gleamer, Radiobotics, AZmed, Milvue) on 29th July and companies were asked to return their responses by 7th August. An additional round of questions was submitted to one company (Radiobotics) on 5th August. These dates were later than scheduled in the protocol to coincide with the completion of evidence selection, at which point the questions could address uncertainties across the identified evidence base.

2. TECHNOLOGIES

A brief overview of the technologies included in the assessment can be found in Table 1. Please see the NICE scope for further details.

Table 1: Description of technologies included in the assessment

Technology (company)	Key Features
BoneView (Gleamer))	• BoneView detects fractures in X-rays of the appendicular skeleton, ribs and thoracic-lumbar spine. Also dislocations, effusions and bone lesions. • Exclusions: skull, cervical spine (IFU) • Suitable for people aged 2 years and over. • Compatible with all available X-ray imaging systems. • Uses X-ray radiographs in DICOM format. • Results are presented as either positive for fracture, negative, or doubt (IFU), with bounding boxes placed around identified anomalies. Where there is doubt about the presence of an anomaly, the bounding box is dashed. No response is given where an excluded body area is analysed (IFU).
Rayvolve (AZMed)	• Rayvolve detects fractures in the appendicular skeleton and ribs. Also dislocations and joint effusions. • Suitable for adults only. • Uses X-ray radiographs in DICOM format. • Integrated into hospitals' existing radiology workflows using Wellbeing's AI Connect gateway.
RBFracture (Radiobotics)	• RBFracture detects fractures in the appendicular skeleton, knee, elbow, rib and periprosthetic area • Exclusions: skull, face, spine (IFU), chronic and healed fractures (RFI) • Suitable for people aged 2 years and over (RFI) • Compatible with all available X-ray imaging systems • Uses X-rays in DICOM format • A red dot is placed on summary reports to indicate an identified anomaly. Bounding boxes are placed around identified anomalies; lower confidence results are indicated using a dashed line provided • Most recent version at time of assessment: 2.1. [redacted]
qMSK (Qure.ai)	• qMSK detects fractures in the appendicular skeleton and ribs. • Suitable for adults only.

Technology (company)	Key Features
TechCare Alert (Milvue)	• Detects fractures in X-rays of the appendicular skeleton and the ribs. Also dislocations. • No age limit for use. • Uses X-rays in DICOM format. • Boundary boxes are placed around areas of interest.

Abbreviations: DICOM, Digital Imaging and Communications in Medicine; IFU, instructions for use; RFI, request for information

Source: all information is reproduced from the NICE scope for this assessment unless stated otherwise.

3. CLINICAL CONTEXT

3.1. Care pathway

The care pathway for fractures within emergency settings in the NHS is described in the NICE scope for this assessment. As part of its assessment, the EAG noted the following additional considerations regarding the care pathway for fractures that was relevant to interpreting the evidence and understanding the potential role of AI as an aid to detecting fractures:

- The scope for this assessment was to consider the clinical and cost effectiveness of AI technologies as a diagnostic aid for detecting fractures in emergency settings, including the emergency department (ED), urgent treatment centre (UTC), and minor injuries units (MIU). These settings differ in the care pathways available to people with suspected fractures, for example in the grade of staff who are available to read radiographs and options for further imaging modalities. There may also be differences in the populations who are admitted to each of these settings, for example MIUs rarely assess people with suspected hip fractures, who in general would be referred to an ED, depending on local policy. In general, MIUs will generally receive a higher case mix of people with suspected fractures of the extremities. Variation in the care pathway and the populations treated within each setting will have implications for the potential value of using AI as a decision aid, as the value of AI is likely to vary according to the staff members using the technology, the fractures assessed, and the downstream impacts of AI on other parts of the care pathway. In order to interpret the evidence for AI as a decision aid presented in this report, it will be important to ensure that the evidence most relevant to the target setting is used and any variation between the evidence base and the target setting is considered carefully.
- Clinical experts noted that the care pathway for the assessment and treatment of fractures varies across the UK as each healthcare service develops and follows its own protocols. While there are established guidelines and standards that guide clinical practice, it is common for services to adapt the care pathway to adjust to the staff and resources they have. This means that there are variations in the staff who provide the initial assessment of x-rays, the use of additional imaging modalities, and the length of time until a definitive report is available. As with the importance of the target setting for the technology, the EAG also noted that the generalisability of the evidence for AI in this report to each healthcare service will vary according to local protocols for assessment and management. In addition to the issues discussed in the NICE scope, the EAG also highlight the following considerations:

- While NICE recommends[2] that a radiologist, radiographer or other trained reporter should provide a definitive report of an X-ray prior to discharge, the EAG was advised that this was very rarely possible and the typical time to a definitive report was between 24 hours and 2 weeks across different services. This is much longer than targets set by NHS England who have suggested, in consultation with The Royal College of Radiologists and The Society of Radiographers[3], a 12-hour target to a definitive report of X-rays for outpatients in emergency settings, with the aim that this should be reduced to a 4-hours. This will be very difficult to achieve without major changes to the service.
- The EAG received advice that services also vary in their approach to producing definitive reports, with some centres processing these on a first come, first served basis, while other centres may use alternative strategies, such as only providing a definitive report of X-rays where the initial assessment was negative (i.e. to confirm that a fracture had not been missed).
- A definitive report from a consultant radiologist or reporting radiographer may not always be necessary depending on the reader of the initial X-ray; for example, a consultant hip surgeon may make the initial diagnosis without the need for a further report.
- The EAG was aware that centres may take precautions to reduce the risk of missed fractures, particularly for those fractures that are known to be challenging to detect on X-ray. For example, suspected scaphoid fractures and intra-articular fractures may have long-term consequences for a person's health and mobility if missed, and so the injury may still be treated as a fracture and re-assessed in two weeks. A risk-averse approach may also be taken with more vulnerable patients for whom the potential consequences of a missed fracture would be considered to be greatest (e.g. children, people identified as frail). Such precautionary tactics are associated with a reduced risk of missed fractures but an increased risk of over-treatment, including the need for additional assessments after 2-weeks. This means that the potential value of AI as a decision aid will vary according to whether centres use any precautionary tactics.

- There are variations in the staff that would be involved in diagnosing fractures for people with suspected fractures that present to emergency settings out of hours (weekday evenings after 6pm, overnight, and weekends). During out of hours, some centres may outsource diagnosis of x-rays to centres overseas (i.e. where daytime staff are available due to the

time difference). Clinical experts suggested outsourcing diagnosis may not necessarily be as accurate, though the EAG did not have any data to confirm this. There may also be variations in the types of fractures seen in emergency settings during out of hours, due to variation in the cause of injuries; for example, sports injuries are more prevalent at weekends, while alcohol induced injuries (falls and injuries due to violence) may be more prevalent overnight. The EAG therefore considered it plausible that the potential value of AI as a diagnostic aid may vary according to whether people with suspected injuries were presenting to emergency settings during out of hours services or during weekdays.

- The EAG considered it plausible that the introduction of AI as a decision aid within the NHS may lead to broader changes to the care pathway, which may vary across centres and be challenging to predict. For example, centres that use AI may alter the staff that are required to provide the initial assessment, may change their practices for ordering and timing the definitive report, and may alter their use of precautionary tactics. Further consideration of issues related to the introduction of AI within a healthcare setting as discussed in Section 3.2.

3.2. Considerations for implementing AI as a diagnostic aid in clinical practice

- As discussed in the review protocol, the EAG received advice that AI would likely be valuable for use as a decision aid in a select group of fractures only, such as those that are challenging to diagnose on X-ray during the initial review and where the consequences for a missed fracture are greatest. However, the EAG was uncertain to what extent the technologies could be targeted towards specific fracture sites only or whether, once implemented, the technology would provide an analysis of all X-rays that are entered into the radiology picture archiving and communication systems (PACS). Information provided by the companies involved in this assessment was that services are charged on a fee for scan basis, with some offering volume-based discounts, meaning that the cost of the technology would vary considerably according to whether the technology is used for some or all suspected fractures. If a service chose to target the technology towards specific fracture types and locations, the EAG was uncertain how feasible this would be to implement in practice, for example whether the technology could be specified to not produce an analysis of certain body locations (as is the case for body locations not covered by the technology) or whether the treating clinician would need to choose to activate the technology for every exam. Expert advice to the EAG was that either approach could be feasible, with the

selective use of the technology ensuring that the technology was only used in the minority of circumstances where additional review was beneficial. Cost implications for the technology will vary according to the way in which the technology is implemented.

- The EAG considered that the successful integration of the technology with Radiology picture archiving and communication systems (PACS), and the perceived ease and acceptability of this, would be important for considering the potential value of the technology. This may include consideration of the effort required to produce a result, the notifications used, and the presentation of the findings. The EAG also considered that, were the technology used, in future it may be required to sit alongside other AI technologies analysing the same images for other anomalies. The EAG was uncertain to what extent the technologies already have this functionality but were aware that other technologies for analysing X-rays for other abnormalities are available.[4]
- The EAG considered it plausible that the successful integration of the technology and the extent to which staff have confidence in it may influence the way it's used in clinical practice and, therefore, its potential value for the service. For example, where there are discrepancies in the result given by the technology and clinical judgement, staff confidence in the technology may determine which result they choose to prioritise in the final decision. The EAG received advice from clinical experts that confidence in the technology may vary according to staff grade, in that more junior staff may be more likely to place greater weight on the result from the AI. The EAG was also advised that over-confidence in the AI could have negative consequences; for example, staff may rely on the AI diagnosis during busy periods.
- Clinical advice to the EAG was that the optimum order of use would be for the reader to form a judgement about the X-ray without use of the technology, and only then consult the results provided by the technology. This was partly to avoid over-reliance on the technology, with another benefit being to reduce the risk that readers become less skilled in reading X-rays over time, with potential knock-on consequences (for example, were the technology to not be available in the future). Were the technology to be used in this order, this would increase the reporting time required for X-rays, with potential knock-on consequences for the service.
- The algorithm within AI technologies used for detecting fractures specifies the threshold at which an identified anomaly is defined as a fracture or not, i.e. the level of confidence required for the technology to return a positive result. Depending on the threshold used, this may favour sensitivity or specificity of the technology, according to whether the threshold is

selected to prioritise to avoid missed fractures or false positives. The EAG noted the following considerations on this topic:

  - Generally speaking, the preferred threshold may vary according to the fracture being assessed or the needs of the patient. For example, a lower confidence threshold may be chosen for scaphoid fractures when there are significant consequences to the patient of a missed fracture. The EAG was uncertain whether it was possible for the operator to adjust the threshold used by the technology according to the fracture being assessed or whether this changed manually. Based on the information received, the EAG considered it more plausible that the same threshold was used for all fractures. As a consequence, the EAG considered it important that instructions and training for operators encourages the use of clinical judgement to interpret binary responses from technologies.
  - Some technologies, particularly more recent versions, include a note about the confidence of the result, such as by using a boundary box around the identified anomaly. The EAG considered that confidence metrics may be interpreted differently across users, although this was something that it was not possible to evaluate within this EVA.
  - There is no clear label or metric to identify the threshold used by the technologies during the published evaluations (for example, there is no apparent scale from which to report the threshold used and compare this across studies and different technologies). This makes it difficult to evaluate the performance of different thresholds, and it was also not clear to the EAG when studies of the same technology were using the same or different thresholds. As the sensitivity and specificity of the technology for detecting fractures will vary according to the threshold used, this creates significant uncertainty in the generalisability of the evidence base to clinical practice.

- AI technologies may be updated over time to refine and improve the technology. These iterative updates may lead to dynamic changes to the cost effectiveness of the technology and it was not clear to what extent NICE or local services could monitor this. Expert advice to the EAG was that frequent updates to the technology may be challenging for UK bodies to appraise, and it may be that an approved AI technology would not be expected to undergo update except to correct any defects or safety considerations.

- The EAG was unclear about any legal or ethical implications around using the technology within the NHS. For example, how the introduction of the technology would affect liability considerations for missed fractures discussed in the NICE scope.

3.3. Equality issues

Equality considerations for this assessment were noted in the NICE Scope. Further to these, the EAG identified a paucity of evidence for particularly populations who may be more vulnerable to missed fractures and over-treatment, including people with frailty and those with health conditions that affect their bone health and long-term recovery.

4. CLINICAL, SERVICE AND TECHNOLOGICAL EVIDENCE SELECTION

4.1. Evidence search strategies and study selection

The search strategies are presented in Appendix A and PRISMA diagram for the evidence selection is presented in Appendix B.

Searches were carried out in late June and early July 2024. The search strategies used relevant search terms for artificial intelligence, X-ray and different fracture types; each of these subjects comprised a combination of indexed keywords (e.g., Medical Subject Headings, MeSH) and free-text terms appearing in the titles and/or abstracts of database records. Searches were translated and adapted according to the configuration of each database. No date, language or publication status (published, unpublished, in-press, and in-progress) limits were applied. Searches for clinical and service outcomes and cost-effectiveness were combined and carried out in one search strategy.

Following deduplication, a total of 1,341 records of potentially relevant evidence on clinical and/or cost effectiveness were retrieved. Databases searched were Medline (including Medline in Process), Embase, Cochrane, Web of Science, CEA Registry and HERC. Additional trial registries searched were Clinicaltrials.gov (NLM) and ICTRP (WHO). The websites of the individual companies were searched; NICE and SIGN websites were searched for related guidelines; MAUDE and MHRA were searched for adverse events data. In addition, we scanned the reference lists for the Kuo 2022[5] and Pauling 2024[6] systematic reviews.

During study screening, the EAG identified several studies where the technology evaluated was unclear or where it was unclear whether the technology was evaluated as a standalone technology or with the interpretation of a clinician. In these cases, the EAG sought advice from three of the companies included in this assessment (Gleamer, Milvue, and AZmed) where they were stated to have sponsored the study and/or where staff from their companies were listed as authors. The EAG did not receive a response from Milvue or AZmed during its assessment, and therefore the studies queried were not included in the review. Following a response from Gleamer, one study was included in the assessment and two studies were excluded. Additional queries about included studies were also sent to companies (see Section 7) for details, which resulted in the merging of two studies: Radiobotics 2021[7] and Bonde [unpublished]. The former of these was a published document with little reported data (though following a request from the EAG the company, Radiobotics, provided additional data) and the latter was a full manuscript in

preparation that reported data for a subgroup of participants in one of the included countries. A full list of exclusions is provided in Appendix C.

4.2. Included and excluded studies

A total of 1,343 titles and abstracts were screened, 209 full-text publications were reviewed, and 16 studies (17 documents) met the review inclusion criteria. The included studies are summarised in Table 2.

One study included in the review (Bousson 2023[8]) conducted a head-to-head comparison of assisted reading using three technologies: BoneView, Rayvolve and TechCare Alert (please note that the study referred to TechCare Alert as 'SmartUrgences', which is an umbrella name for AI technologies developed by the same manufacturer. For pragmatic purposes, the EAG assumed that these were the same technology). The majority of other studies included in the review evaluated readings assisted with either BoneView or RBFracture.

A breakdown of the number of studies evaluating each technology is as follows:

- **BoneView**: nine studies, including two non-comparative studies where BoneView assisted readings were assessed against the reference standard (Cohen 2023[9], Meetschen 2024[10]), one head-to-head comparison (Bousson 2023[8]) and five studies that assessed both BoneView-assisted and unassisted readings (Canoni-Meynet 2022[11], Dell-Aria 2024[12], Duron 2021[13], Guermazi 2022[14], Nguyen 2022[15], Oppenheimer 2023[16])
- **RBFracture**: five comparative studies (Bachmann 2024[17], Jørgensen 2023[18], Radiobotics 2021[7] (also Bonde), Ruitenbeek 2024[19], Yogendra [unpublished][20])
- **Rayvolve**: one comparative study (Fu 2024[21]) and one head-to-head comparison (Bousson 2023[8])
- **TechCare Alert**: one comparative study (Suite 2020[22]) and one head-to-head comparison (Bousson 2023[23])
- **qMSK**: zero studies

Table 2: Overview of included studies with clinical and technological evidence

First author (date), location, publication type	Index test and comparators	Reference Standard	Participants, images, age range, included and excluded fractures	Type of outcome reported	EAG Notes
Head-to-head comparison					
Bousson (2023)[23], France, published article	BoneView/ Rayvolve/ TechCare Alert	Consensus between 4 radiologists (one senior)	1210 adults and adolescents (15 years or older), 1500 images, included clavicle, shoulder, humerus shaft, elbow, radius/ulna shaft, wrist/hand, finger, pelvis/hip, femur shaft, knee, tibia/fibula shaft, ankle, and foot	DTA	No unassisted results. The reference standard was based upon a combination of the AI output and clinician reports, therefore DTA results were considered to be at high risk of bias
BoneView					
Cohen (2023)[9], France, published article	BoneView assisted (no comparator)	Consensus between 3 radiologists with access to clinical information and additional imaging where available.	637 adults, 1917 images, included wrist only	DTA	Wrist fractures only.
Canoni-Meynet (2022)[11], France, published article	BoneView assisted vs unassisted	Consensus between 3 radiologists and AI, or 2 senior radiologists and AI	500 adults and children, 500 images, included skeletal, excluded skull and facial	DTA, service	The reference standard was based upon a combination of the unassisted readings and the AI output, therefore DTA results were considered to be at risk of bias
Dell-Aria (2024)[24], Belgium, published article	BoneView assisted vs unassisted	Decision by one radiologist (in consultation with an orthopaedic surgeon) who had access to all medical information and could examine the patient if necessary. CT sought where required.	101 adults (>/=18 years), included upper and lower limbs including shoulder and hip, excluded other locations	DTA	Low-velocity trauma injuries to upper and lower limbs only (included shoulder and hip)

First author (date), location, publication type	Index test and comparators	Reference Standard	Participants, images, age range, included and excluded fractures	Type of outcome reported	EAG Notes
Duron (2021)[13], France, published article	BoneView assisted vs unassisted	Consensus between 2 radiologists (disagreements resolved by another radiologist)	600 adults, 600 images, included shoulder, arm, hand, pelvis, leg, foot	DTA, service	Excluded obvious fractures from the sample.
Guermazi (2022)[14], USA, published article	BoneView assisted vs unassisted	Consensus between 2 radiologists without clinical information	480 participants, age NR but appear to be only or mainly adults. Appendicular skeleton with equal number of obvious and non-obvious fractures.	DTA, service	
Meetschen (2024)[10], Germany, published article	BoneView assisted (no comparator)	Consensus between 2 consultant radiologists with access to additional imaging where required.	200 adults and children, 200 images, included hand, wrist, arm, elbow, shoulder, scapula, clavicle, ribs, spine, pelvis, hip joints, legs, knees, ankles, and feet	DTA, service	
Nguyen (2022)[25], USA/France, published article	BoneView assisted vs unassisted	Consensus between 2 radiologists (disagreements resolved by another radiologist)	300 children and young adults (2 to 21 years old), 300 images, included appendicuar skeleton, excluded skull, pelvis, rib cage, spine	DTA	Children and young people only.
Oppenheimer (2023)[16], Germany, published article	BoneView assisted vs unassisted	Decision by one radiologist with access to the preliminary reports. CT, MRI, or PET were used as the reference standard when eprfrmend within 1 week of intiial X-ray and no new trauma or symptoms.	735 children and adults (range 2 - 100 years), 1163 images, included skeletal, excluded cervical spine, skull, face	DTA	It was unclear whether the reference standard included consideration of the AI results, therefore DTA results were considered to be at risk of bias
Rayvolve					
Fu (2024)[21], USA, published article	Rayvolve assisted vs unassisted	Consensus between at least 2 of 3 radiologists	Adults (>=22 years), sample size NR but 186 exams. Fractures included ankle, clavicle, elbow, forearm, humerus, hip, knee, pelvis, shoulder, tibia/fibula, wrist, hand, foot	DTA, service	
RBFracture					

First author (date), location, publication type	Index test and comparators	Reference Standard	Participants, images, age range, included and excluded fractures	Type of outcome reported	EAG Notes
Bachmann (2024)[17], US/Denmark, published article	RBFracture assisted vs unassisted	Consensus between 2 consultant radiologists (disagreements resolved by another consultant radiologist) with access o clinical information and original radiology reports.	340 adults (>=21 years) and children (>=2years), 340 images, included appendicular skeleton, excluded other locations (e.g. ribs, spine)	DTA, service	
Radiobotics (2021)[7], USA/Denmark, document. Bonde, unpublished manuscript[26], [redacted]	RBFracture assisted vs unassisted	Consensus between 2 radiologists (disagreements resolved by a reporting radiographer)	312 adults (>/=21 years), 312 images, included unobvious hip, excluded all other locations and obvious hip	DTA	Non-obvious suspected hip fractures only.
Jørgensen (2023)[18], Denmark, abstract	RBFracture assisted vs unassisted	Radiology report was used, including information about subsequent imaging where ordered.	214 adults, 214 images, included hip	DTA	Hip fractures only
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]DT A results were therefore considered to be at risk of bias
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	
TechCare Alert					

First author (date), location, publication type	Index test and comparators	Reference Standard	Participants, images, age range, included and excluded fractures	Type of outcome reported	EAG Notes
Suite (2020)[22], France, document	TechCare assisted vs unassisted	The original radiology report produced by a radiologist was used.	650 adults (>/=18 years), included lower limbs (pelvis, ankle, knee, hip, leg, foot), upper limbs (arm, elbow, shoulder, hand, wrist), thorax (ribs), excluded other locations	DTA	

Abbreviations: CT, computed tomography; DTA, diagnostic test accuracy; MRI, magnetic resonance imaging; NR, not reported; PET, positron emission tomography

4.2.1. Study design and diagnostic tests

Details about the methodological approach used by the included studies are provided in Table 3.

Six studies[9 11 16 23 24 27] included consecutive cases presenting to participating centres during the study period and one study[21] included a random selection of cases from a database of patients who presented with a suspected fracture. The remaining nine studies[7 10 13 14 17 18 20 22 25] used a case-control design to stratify inclusion towards a set fracture prevalence rate (typically 50%, with additional requirements, such as spread of specific fracture types and age groups). Consecutive and random sampling study designs are generally more robust for diagnostic evaluation, as the study samples more closely represent the prevalence of the target condition that would be seen in clinical practice, meaning that prevalence data and PPV and NPV outcomes (each affected by the prevalence rate) will be more reliable. However, as the diagnostic accuracy of X-ray varies according to the fracture location, even consecutive and random sampling designs may be limited if they do not include a mix of fracture types that is representative of clinical practice (meaning that sensitivity and specificity data may not be generalisable). This issue is discussed further in Section 4.2.1.

The vast majority of studies used a retrospective study design, with only two studies (Oppenheimer 2023[16] and Dell-Aria 2024[24]) using a prospective design. Retrospective designs are unlikely to be representative of clinical practice for several reasons, including that readers have knowledge that their judgement will not impact upon patient care, readers are less likely to have the same access to support from colleagues, and readers may not dedicate the same length of time and consideration to reading X-rays within these circumstances compared to clinical practice.

None of the included studies were set in the UK, though the majority (12 studies) included sites in Europe, five studies included sites in USA, and one study was set in Asia (Singapore). Based on the information reported in the included studies, it was not clear to the EAG how applicable the study findings were to the decision problem for the assessment; i.e. the target emergency settings of ED, MIU and UTC, and the extent to which readers of index tests and the reference standard were comparable to staff in the target settings. This was a major limitation of the evidence base, since – as noted in the review protocol – the diagnostic accuracy of the index test and technologies would be expected to vary according to both case mix and reader experience. Eight of the included studies[9-11 16 18 20 22-24 27] were described to be set within a

hospital and/or trauma centre.

[redacted]. However, the EAG was uncertain to what extent the care pathway and treating clinicians would vary between hospital settings in included studies as compared to the UK. Of the other studies, two studies[13 21] were described as being set in medical settings and one study[17] was described as being based within a virtual centre. Two studies[14 25] were described as using data from unspecified settings with the USA.

Many of the included studies reported multiple analyses for the diagnostic accuracy of assisted and unassisted readers according to different staff grades (in addition to other relevant subgroups, such as age group and fracture location). As it was not possible to determine which analyses were most relevant to NHS staff who would be expected to use the technology in clinical practice, the EAG instead sought to categorise available data according to the level of experience as described by the publications: less experienced staff; highly experienced staff, and mixed or unclear levels of experience. The EAG noted that these categorisations were based on highly limited information and there is a high risk of error. Further information about this is provided in Section 5.1. Very few studies included staff that appeared comparable to those who would typically work within urgent care settings in the NHS, such as emergency care doctors or specialist trauma nurses.

Only three of the included studies specified the version of the technology under evaluation: in these studies, it was stated that Bousson 2023 evaluated BoneView version 1.0.2. and TechCareAlert version 1.7 (no version reported for Rayvolve), Radiobotics (2021) evaluated RBFracture version [redacted] and Yogendra [unpublished] evaluated RBFracture version [redacted]. Descriptions of the output of the evaluated technologies was typically poor across studies, and therefore the EAG did not feel confident in differentiating between studies on the basis of features mentioned or not mentioned in the publications. For example, two studies[13 16] evaluating BoneView reported that the technology provided readers with an indication of confidence in the result, either as a note on the result and/or as an altered boundary box around the area of interest.

[redacted]. However, the EAG considered it plausible that other studies included in the assessment also evaluated versions of the technology that provided a rating of confidence with results, and therefore did not use reporting of this feature to draw a comparison across studies. Similarly, three studies mentioned that staff received training in the AI prior to the study (Canoni-Meynet

2022, Bachmann 2024, Radiobotics 2021), with varying levels of training ranging from months (Canoni-Meynet 2022) to written instructions and five training cases only (Bachmann 2024). The EAG considered it plausible that all studies would have trained readers in using the technology prior to the study, and that training was simply not described for most studies. The EAG therefore did also not use training requirements as a factor for comparing findings.

Where the same readers read the same exams assisted and unassisted, the washout period between assisted and unassisted readings ranged between no washout to three months (two studies [16 25]had no washout, four studies[11 14 17 21] had 1 month washout, [redacted], one study[23] had 2 months washout, and one study[24] had 3-months washout). The EAG considered that a washout of 1-month or longer was sufficient to ensure that the reading clinicians did not recall their previous responses to an X-ray, or at least would do so in very few cases. The EAG considered that there was a risk of bias associated with no washout period, as readers may be guided by their previous responses using the alternative method. Where washout was not reported, the EAG considered it more likely that no washout period was used, though of course this was not clear.

One of the studies used an index test that was considered not to be representative of the likely use of the technology: in this study to detect wrist fractures, rather than readers using the technology to reach a decision on diagnosis, the standalone results from the AI results and results of the original radiology report were artificially combined: an observation was considered positive when it was detected by either the AI or reported on IRR, regardless of the other's group result. The EAG considered this study to be at a risk of bias.

The reference standard used in the included studies generally included a decision from a senior radiologist, though there was some variation in the information that readers had to make their diagnosis (e.g. clinical information and medical notes, and access to further imaging). Where the reference standard was based on limited access to information about the patient and injury (as would be available within routine clinical practice), the EAG considered there to be an increased risk of incorrect judgements. Where a reference standard is determined to be imperfect, this affects the reliability of all the study results. The reference standard used in three studies (Canoni-Meynet 2022, Bousson 2023 and [redacted]) included the results of the AI technology, and in a further study it was unclear whether this was the case (Oppenheimer 2023). The inclusion of these studies was a deviation from the review protocol (as described in Section 1) and was also associated with the potential for bias in the results, as it created closer

alignment between the results of the AI-assisted assessment and the reference standard. The EAG considered that the evidence from these studies should be interpreted with caution.

On the basis of information about the study design used by the included studies, and as described in Section 1), the EAG identified studies that provided the best quality evidence available within the evidence base and therefore could be used to inform the economic analysis. These decisions were based upon the following *post hoc* criteria: having a reference standard that did not include the AI reports, inclusion of results for both AI assisted and unassisted readers, relatively large sample sizes and peer-reviewed publication. Based upon these key criteria, the EAG considered the pivotal studies for each technology, to be as follows:

- **BoneView**: Duron 2021 (for adults) and Nguyen 2022 (for children and young adults).
- **RBFracture**: Bachmann 2024 (adults and children).
- **Rayvolve**: Two studies evaluated Rayvolve (Fu 2024 and Bousson 2023). Although limited by a relatively small sample size, Fu 2024 avoided the limitations noted above and was therefore considered by the EAG to be the pivotal study for this technology.
- **TechCare Alert**: Two studies evaluated TechCare Alert (Suite 2020 and Bousson 2023), both of which were associated with limitations as described above. Suite 2020 was also not presented in a peer-reviewed publication. Evidence for this technology was therefore of poorer quality than the studies listed above.

The EAG emphasises that all of these studies are associated with their own quality limitations that may undermine the reliability of the findings. In particular, none of the studies named above used a prospective design and may not be representative of clinical practice. Unfortunately, the two studies that included a prospective design were limited in other ways (see section 4.31.1.1) and were considered to be less reliable than those selected. Specifically, in Oppenheimer (2023), it was unclear whether the reference standard included the results of the technology, and Dell-Aria (2024) was a relatively small study.

Participant selection was also a key consideration in the selection: studies where participants were selected to ensure sufficient numbers of fractures (i.e. case-control designs, where 50% of the sample were selected for the presence of a fracture) would not have a fracture prevalence rate representative of clinical practice (and may also be unlikely to have a representative fracture location mix). These studies were not used to determine fracture prevalence in

economic analysis, and PPV and NPV estimates were considered to be unreliable. The pivotal studies by Duron 2021, Nguyen 2022 and Bachmann 2024 used a case-control design, but due to methodological issues with the other studies for these technologies, these studies remained the most robust for estimating sensitivity and specificity (prevalence, PPV and NPV were not extracted or considered from these studies). One of the studies that evaluated TechCare Alert (Suite 2020) also used a case-control design, with the other being limited in other ways (see 4.3). The relevant study sample from the pivotal Rayvolve study (Fu 2024) used random sampling. However, it is unclear, but appears likely, that the original dataset from which the sample were selected was based on a case-control design.

Important subgroups in the review were different staff grade/type and specific fracture locations (hand/wrist, foot/ankle, hip, elbow fractures in children and Salter-Harris fractures). The majority of studies provided some relevant subgroup data, with four BoneView studies (Canoni-Meynet 2022, Meetschen 2024, Dell-Aria 2024, Guermazi 2022), three RBFracture studies (Radiobotics 2021 (also Bonde), Bachmann 2023, Yogendra) and the Rayvolve and TechCare studies (Fu 2024 and Suite 2020 respectively) all reporting some DTA data for different staff types or grades and five BoneView studies (Canoni-Meynet 2022, Duron 2021, Ngyuen 2022, Oppenheimer 2023, Guermazi 2022) and the head-to-head study (Bousson 2023) reporting some DTA data for different fracture locations. Oppenheimer (2023) and Bachmann (2024) each reported subgroup data in children and young people.

Table 3: Study design of included studies

First author (date), location, publication type	Index test and comparators	Design	Order of tests and washout	Index tests details	Reference standard details	Subgroup data
Head-to-head comparison						
Bousson (2023), France, published article	BoneView/ Rayvolve/ TechCare Alert	Retrospective, consecutive sampling	Unassisted/assisted, ground truth 2 months later	BoneView (version: v1.0.2; output: doubtful fractures were categorised as a fracture). Readers: six radiology residents (4 years of residency)	Four musculoskeletal radiologists, three fellows and one senior radiologist. Combination of radiology reports and AI results. Timing: 2 months later. N = 1500	Fracture location
BoneView						
Canoni-Meynet (2022), France, published article	BoneView assisted vs unassisted	Retrospective, consecutive sampling	Unassisted, 1 month washout, assisted, ground truth	BoneView (version: NR; output: Bounding boxes overprinted on X-rays). Reading clinicians: 3 radiologists with >= four months daily practice with the AI system (Clinical readers were radiologists of varying seniority: 15 and 2 years of experience in MSK imaging, and a third-year resident). Process: NR	Agreement between AI and radiologists' decisions, Timing: NR. N = 500. Note: some decisions were based only on the two senior radiologists.	Staff grade/type Fracture location
Cohen (2023), France, published article	BoneView standalone + radiology report vs. unassisted	Retrospective, consecutive sampling	Unclear	BoneView (version: NR; output: Binary as doubtful fractures were categorised as a fracture). Reading clinicians: Initial radiology reports (IRRs) were used to provided radiologist support to the AI. These IRRs were made by a total of 41 radiologists with various levels of experience, including 29 residents (4th or 5th year of residency), eight fellows in radiology, and four attending radiologists.). Process: NR	Consensus between three expert radiologists (5+ years of experience in MSK radiology) with access to additional imaging where available. Timing: NR. N = 637.	None
Dell-Aria (2024), Belgium,	BoneView assisted vs unassisted	Prospective, consecutive sampling	Ground truth, unassisted, 3	BoneView (version: NR; output: Binary). Multiple suspected fractures in one location had one result.	A radiologist with >15 years' experience read the X-rays and had access to all the medical	Staff grade/type

First author (date), location, publication type	Index test and comparators	Design	Order of tests and washout	Index tests details	Reference standard details	Subgroup data
published article			month wash-out, assisted,	Reading clinicians: Radiologist with <5 years' experience reading X-rays that were not ultimately sent for CT imaging (No consultant support). Process: Reading of the X-rays with the AI results was performed without access to any other clinical information, including the clinical history of the patient. It took place 3 months after the comparison test (unassisted reading of the X-rays)	information and was able to examine the patient if necessary. They consulted with an orthopaedic surgeon to determine if CT imaging was required to confirm diagnosis. Patients where there was/wasn't a decision reached between radiologist and surgeon were analysed separately. Timing: NR. N = 101.	
Duron (2021), France, published article	BoneView assisted vs unassisted	Retrospective, case-control with random selection	Ground truth, Assisted and Unassisted (reader cross over no wash out): Note: The readers did not look at all images assisted and unassisted, instead they each looked at 150 assisted and unassisted	BoneView (version: NR; output: highlight of each region of interest with a box and provision of a confidence score regarding the existence of a fracture in the region of interest). Reading clinicians: Twelve independent readers (six radiologists and six emergency physicians) of various levels of experience (including residents and experts) (No consultant support though a number of the readers were considered "experts"). Process: Readers were blinded to clinical data, and no time constraints for reading. Readers were blinded to one another and to expert's judgments.	Two skeletal imaging radiologists with >9 years of experience identified the presence and location of a fracture. The presence or absence of fracture was determined by the majority opinion of at least two of the three MSK radiologists. Timing: NR. N = 600.	Fracture location
Guermazi (2022), USA, published article	BoneView assisted vs unassisted	Retrospective, case-control with stratification	Ground truth, then all X-rays were read either assisted or unassisted (randomised 50% of each, switched every 120 exams), 1	BoneView (version NR). AI-assisted mixed group of clinicians: 4 radiologists, 4 orthopaedic surgeons, 4 emergency medicine physicians, 4 emergency medicine physician assistants, 4 internal medicine physicians, and 4 family practice physicians. They had 2 - 18 years of experience in radiographic	Two experienced musculoskeletal radiologists with 12 and 8 years of experience independently interpreted all examinations without clinical information	Staff grade

First author (date), location, publication type	Index test and comparators	Design	Order of tests and washout	Index tests details	Reference standard details	Subgroup data
			month washout, then the opposite x-rays were read assisted and unassisted (switched every 120 exams)	interpretation. All readers had 1 hour of training to get accustomed to the AI.		
Meetschen (2024), Germany, published article	BoneView assisted (no comparator)	Retrospective, case-control with random selection	Ground truth timing unclear. Note: The readers did not look at all 200 images assisted and unassisted, instead they each looked at 100 assisted and 100 unassisted (crossover across readers))	BoneView (version: NR, stated to be built on the "Detectron 2" framework; output: Binary. Reading clinicians: 4 radiology residents with different levels of training in the radiology residency program, ranging from 4.5 to 24.5 months of experience.	Consensus between 2 consultant radiologists with 7 and 10 years of experience in musculoskeletal imaging with access to additional imaging where indicated. Timing: NR. N = 200.	Staff grade/type
Nguyen (2022), USA/France, published article	BoneView assisted vs unassisted	Retrospective, case-control with random selection	Ground truth, Each radiograph examined in turn unassisted then assisted (no wash out)	BoneView (version: NR; output: Binary). Reading clinicians: Eight readers: 5 radiology residents (between the 2nd and 4th year of residency) and 3 expert paediatric radiologists (at least 7 years of experience, including >3 years specialising in paediatric radiology (No consultant support). Process: After being trained with the AI on 10 independent cases, they were presented with the 300 radiographic examinations in random order. The readers analysed each radiograph without AI, then with AI before moving to the next radiograph. They	Consensus between 2 board-certified musculoskeletal radiologists with >8 of experience in musculoskeletal imaging . If they disagreed, a third board-certified expert MSK radiologist (more than 25 years of experience in musculoskeletal imaging) was involved to settle the matter by consensus. Fractures were classified as obvious or nonobvious depending on their difficulty by the three radiologists who established the ground truth. A fourth board-certified paediatric radiologist	Fracture location

First author (date), location, publication type	Index test and comparators	Design	Order of tests and washout	Index tests details	Reference standard details	Subgroup data
				had no time constraints to analyse the radiographs. They were asked first to mark the fractures by level of confidence, doubtful or certain. After consulting the AI-aided results, they could modify their diagnosis or change the confidence in their diagnosis.	(T.N. with 5 years of specialisation in paediatric imaging) also reviewed all radiographs after the ground truth was determined to classify fractures by type. Timing: NR. N = 300.	
Oppenheimer (2023), Germany, published article	BoneView assisted vs unassisted	Prospective, consecutive sampling	Unassisted, assisted (no wash-out), ground truth	BoneView (version: NR; stated to be based on Detectron 2 framework output: this software returned a diagnosis of “Positive,” “Doubt,” or “Negative” with a through line where a fracture was diagnosed and a bounding box with a dashed line where a fracture was possible (“Doubt”). The AI set the threshold of “Doubt” at 50–89% confidence and “fracture” at greater than or equal to 90%. Additionally, the AI marks regions of interest where it diagnoses a joint effusion or dislocation. Reading clinicians: Resident radiologist	An experienced board-certified radiologist reviewed the index results and either signed off unchanged or corrected accordingly. Unclear if this process included consideration of the AI results. Where CT, MRI, PET was performed within one week after the initial radiographic exam, and no new trauma or symptoms were indicated, this diagnosis of this imaging was used as the gold-standard reference, noting where the final report diagnosis was overruled (12.8%). Timing: NR but possibly 24 hours based on typical practice. N = 1163.	Fracture location
Rayvolve						
Fu (2024), USA, published article	Rayvolve assisted vs unassisted	Retrospective, case-control with random selection	Ground truth, Assisted, 1 month wash out, Unassisted	Rayvolve (version: NR; output: binary). Reading clinicians: The readers included eight each of emergency physicians, non-MSK radiologists, and MSK radiologists. This portion of the study consisted of two independent reading sessions separated by a washout period of at least one month to avoid memory bias.	Consensus between 3 US board-certified MSK radiologists, each with 7–16 years of experience, who independently interpreted images using the standard clinical definition of a fracture. Timing: NR. N = 2626.	Staff grade/type

First author (date), location, publication type	Index test and comparators	Design	Order of tests and washout	Index tests details	Reference standard details	Subgroup data
RBFracture						
Bachmann (2024), US/Denmark, published article	RBFracture assisted vs unassisted	Retrospective, case-control with random selection	Ground truth, Assisted and Unassisted (reader cross over with 4-week washout)	RBFracture (version: NR; output: binary using bounding boxes). Reading clinicians: 15 readers: 2 advanced trauma care nurses, 3 diagnostic radiographers, 4 A&E trainees, 3 orthopaedic specialty registrars, 3 radiology specialty registrars. All provided with written instructions and 5 training cases. Process: Two sessions with a 4 week wash out period. In each session, half of the images were read aided and the other half unaided. This was reversed in the second reading session.	Consensus between 2 consultant radiologists with 1- and 10-years post-specialty experience. Disagreements arbitrated by a third consultant radiologist with 14 years' experience. Each radiologist had access to clinical referral notes and the original radiology reports. Timing: NR. N = 340 (6 didn't receive reference standard).	Staff grade/type
Jørgensen (2023), Denmark, abstract	RBFracture assisted vs unassisted	Retrospective, case-control with stratification	Unclear	RBFracture (version: NR; output: NR). Reading clinicians: Two radiographers, two medical interns and two consultants (not clear if they were radiologists). Process: Evaluated fracture status on all exams with support from the AI tool	The radiological reports, taking additional information from CT into consideration when relevant. Timing: NR. N = 214.	None
Radiobotics (2021), USA/Denmark, company document. Bonde [unpubished] contained methods info and results for Denmark only.	RBFracture assisted vs unassisted	Retrospective, case-control	Order unclear, different readers read different sets of images.	RBFracture (version: [redacted]; output: binary. Bounding boxes were placed around the location of a positive result). Reading clinicians [redacted]	Two radiologists with more than 4 years of musculoskeletal experience. Any disagreements were resolved by a reporting radiographer with 11 years of experience, specialising in MSK-reporting. [redacted] Timing: NR. N = 312.	Staff grade/type

First author (date), location, publication type	Index test and comparators	Design	Order of tests and washout	Index tests details	Reference standard details	Subgroup data
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
TechCare Alert						
Suite (2020), France, document	TechCare assisted vs unassisted	Retrospective, case-control with random selection	Ground truth, Unassisted, Assisted (No wash out: the readers looked at different images assisted and unassisted)	TechCare Alert (version: described as SmartUrgences 'Arterys' MSK AI module'; output: not described although assume binary by the results). Reading clinicians: Four junior radiologists with AI. Process: a random selection of previous cases with known diagnoses were identified. The 650 cases were split between the eight radiologists (four junior and four senior) for review without and with AI. The readings were conducted without access to other information, such as medical history.	The original radiology report (i.e. when the patient was first treated) was used. Stated that a radiologist reads all X-rays after a 24-hour delay. Timing: NR; prior to study. N = 650 (30 didn't receive reference standard).	Staff grade/type

Abbreviations: CT, computed tomography; MRI, magnetic resonance imaging; MSK, musculoskeletal; NR, not reported; PET, positron emission tomography

4.2.2. Populations and reported prevalence

Demographics about the study sample used in the included studies, including prevalence rates of any and subgroup fracture types, are shown in As noted in Section 4.2.1, the EAG considered that studies using a case-control design would not provide prevalence data that was representative of clinical practice. The data from these studies should therefore be considered to only represent the case mix in the included studies and was not used as an estimate of fracture prevalence in the economic analysis. The EAG noted that the most methodologically robust study for the estimation of prevalence was likely Cohen 2023, but this study focused only on wrist fractures. Other studies that were reasonably robust for estimating prevalence were Dell-Aria 2024 and Fu 2024 (albeit both limited due to relatively small sample sizes), and [redacted]). Amongst the studies that used consecutive and random sampling, rates of subgroup fracture types were rarely reported. Where these were available, these varied between studies, suggesting that the study samples included different underlying populations. As diagnostic accuracy for detecting fractures on X-rays varies according to fracture location, variation in case mix across studies will likely influence the findings.

Table 4. Studies with more reliable prevalence data (i.e. studies with consecutive and random sampling) are highlighted in bold.

Studies typically only included adults, or the vast majority of the sample were adults (see

). There were five studies[10 11 16 17 23] that included a mix of adults, children and young people, of which two studies[16 17] reported subgroup data specifically in children and young people. Two studies were conducted only in children and young people (Nguyen 2022 and Yogendra [unpublished]).

No studies reported frailty measures for its participants. Two studies[7 18] reported that the mean age of the sample was > 70 years. One study[24] conducted with participants with a median age of 39 years (no range reported) who had all experienced a low velocity trauma, and one study reported that 83.0% of injuries were due to falls. The EAG considered that all of these indicators may be correlated with frailty, but nevertheless wouldn't provide a meaningful representation of outcomes in people with frailty.

The EAG sought additional information from studies about the health of participants, such as the number of participants with diseases that affect bone health (e.g. osteoporotic disease). No studies reported this information.

Eight studies[9-11 13 14 17 24 25] reported that a minority of participants had multiple fractures; where rates were reported, these ranged from 3.3% to 26% of the study sample. None of the studies excluded suspected dislocations and effusions, although only one study[22] stated that the study sample included these injuries. Three studies[7 13 24] explicitly excluded 'obvious' fractures, such as open fractures, displaced, and multi-fragmented and those caused from polytrauma, while other studies neither included nor excluded these. Within the EVA, the EAG did not extract information of the full range of fracture types included in the study samples from which to consider the case mix in which the technologies were evaluated. As noted in Section 4.2.1, the EAG considered that studies using a case-control design would not provide prevalence data that was representative of clinical practice. The data from these studies should therefore be considered to only represent the case mix in the included studies and was not used as an estimate of fracture prevalence in the economic analysis. The EAG noted that the most methodologically robust study for the estimation of prevalence was likely Cohen 2023, but this study focused only on wrist fractures. Other studies that were reasonably robust for estimating prevalence were Dell-Aria 2024 and Fu 2024 (albeit both limited due to relatively small sample sizes), and [redacted]). Amongst the studies that used consecutive and random sampling, rates of subgroup fracture types were rarely reported. Where these were available, these varied between studies, suggesting that the study samples included different underlying populations. As diagnostic

accuracy for detecting fractures on X-rays varies according to fracture location, variation in case mix across studies will likely influence the findings.

Table 4: Study participants and fracture types

First author (Date)	Tech	Mean age (SD) and sex	N participants/N exams	N with fractures (%)	N (%) multiple fractures	Prevalence hip	Prevalence foot/ ankle	Prevalence hand/ wrist	Prevalence elbow in children	Prevalence Salter Harris in children
Head-to-head comparison										
Bousson (2023)	**BoneView /Rayvolve /SmartUrgence**	**Age: 41.3 (18.5) years (range=NR). Sex: 468F:742M**	**1210 people/ 1500 exams**	**326* (21.7%)**	**NR**	**50 (3.3%; pelvis/hip)**	**Ankle 232 (15.5%); Foot 186 (12.4%)**	**314 (20.9%)**	**NR**	**NR**
BoneView										
Cohen (2023)	**BoneView**	**Age: NR (range=NR). Sex: NR**	**637 people/ 1917 X-rays**	**247 (38.8%)**	**166 (26%)**	**0%**	**0%**	**247 (38.8%)**	**0%**	**NR**
Canoni-Meynet (2022)	**BoneView**	**Age: 37 yrs (28); 80.2% were adults (range=0.25−99 yrs). Sex: 232F:268M**	**500**	**188 (37.6%)**	**35 (7%)**	**NR**	**39 (7.8%)**	**38 (7.6%)**	**NR**	**NR**
Dell-Aria (2024)	**BoneView**	**Age: median 39 (range=NR). Sex: 51F:50M**	**101**	**54 (53.9%)**	**15 (14.8%)**	**NR**	**NR**	**NR**	**NR**	**NR**
Duron (2021)	BoneView	Age: 57 (22) (range=18–100 years). Sex: 358F:242M	600 people/ 600 exams	300 (50%)	80 (13%)	NR	Foot: 44 (7.4%)	Hand: 44 (7.3%)	NR	NR
Guermazi 2022	BoneView	Age: 59 (16) (range NR). Sex: 327F:153	480	240 (50%)	16 (3.3%)	44 (9.2%)	38 (7.9%)	52 (10.8%)	NR	NR
Meetschen (2024)	BoneView	Age: 40.7 (24.5) (range=1-95 years). Sex: 95F:105M	200	100 (50%)	NR; stated that 135 fractures were in 100 X-rays	8 (4%; pelvis or hip)	46 (23%)	49 (24.5%)	NR	NR

First author (Date)	Tech	Mean age (SD) and sex	N participants/N exams	N with fractures (%)	N (%) multiple fractures	Prevalence hip	Prevalence foot/ ankle	Prevalence hand/ wrist	Prevalence elbow in children	Prevalence Salter Harris in children
Nguyen (2022)	BoneView	Age: 10.8 (4.9) (range=2 to 21). Sex: 133F:167M	300	150 (50%)	NR; stated that 173 fractures were found in 150 people	0	30 (10%)	30 (10%)	60 (20%; elbow or arm)	Salter II: 21 (7%), Salter IV: 3 (1%)
Oppenheimer (2023)	**BoneView**	**Age: 61.39 (21.9) (range=2 - 100). Sex: 426F:309M**	**735 people/1163 exams**	**367* (31.56%)**	**NR**	**NR**	**NR**	**NR**	**NR**	**NR**
Rayvolve										
Fu (2024)	**Rayvolve**	**Age: NR (range=NR but minimum 21 years). Sex: 1223F:1403M**	**2626 X-rays**	**587* (22.4%)**	**NR**	**53 (2%)**	**Foot: 56 (2.1%) Ankle: 55 (2.1%)**	**Hand: 48 (1.8%) Wrist: 72 (2.7%)**	**35 (1.3%)**	**NR**
RBFracture										
Bachmann (2024)	RBFracture	Age: NR; 236 Adults, 98 children (range=NR). Sex: 136F:106M missing n=92	334 people/ 340 exams	164 (49.1%)	NR although stated that participant with multiple fractures were included	19 (5.7%; pelvis or hip)	30 (9.0%)	30 (9.0%)	9 (2.7%)	NR
Jørgensen (S9 Abstract) (2023)	RBFracture	Age: 78.0 (11.9) (range=NR). Sex: 149F:65M	214	107 (50%)	NR	107 (50%)	0	0	0	0

First author (Date)	Tech	Mean age (SD) and sex	N participants/N exams	N with fractures (%)	N (%) multiple fractures	Prevalence hip	Prevalence foot/ ankle	Prevalence hand/ wrist	Prevalence elbow in children	Prevalence Salter Harris in children
Radiobotics (2021); Bonde (manuscript in prep); confidential study report	RBFracture	Age: [redacted] (range=[redacted]). Sex: [redacted]	312	156 (50%)	NR	156 (50%)	0	0	0	0
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
TechCare Alert										
Suite (2020)	TechCare	Age: 53.6 (23.9) (range=18, 98). Sex: NR	620	253 (40.8%); dislocation 28 (4.5%); effusion 25 (36.2%)	NR	67 (10.8%)	144 (23.2%)	134 (21.6%)	30 (9.4%)	NA

Abbreviations: F, female; M, male; NA, not applicable; NR, not reported; SD, standard deviation

Note: Studies that used consecutive or random sampling, and therefore offer more reliable prevalence rate data, are highlighted in bold. * this is the number of fractures identified across the exams, meaning that a participant with multiple fractures would be counted more than once.

4.2.3. Outcomes

The outcomes reported by, or calculable from, the included studies are shown in Table 5. Sensitivity, specificity, and contingency tables were reported or calculable for all studies, although some data were missing for specificity and contingency tables for subgroup analyses (staff grade and fracture types) in two studies[11 16]. PPV and NPV were either not reported or were not extracted for case-control studies. Additional outcomes reported by a small number of studies were the prevalence of fractures, as assessed using the reference standard (see Section 4.2.21.1.1), and the reading time for radiographs. The vast majority of outcomes listed on the NICE scope were not reported in any of the included studies (see Section 9).

Diagnostic accuracy that was reported per patient (i.e. a binary decision about whether a patient had a fracture) was typically reported across studies and prioritised for inclusion in the review. If these data were not provided, available per-exam or per-fracture data were extracted and these instances are highlighted in the results. The decision to prioritise per-patient data was made primarily because the majority of studies report data in this way and because these data are most relevant to the economic analysis. The EAG noted that the per patient approach allows for the assessment of one injury only, while the per-fracture approach accounts for accuracy across multiple fractures in the same person, and thus means that there will be multiple reports for the same participant. The per-exam approach is similar to the per-patient approach and usually reports on one injury per patient but may plausibly detect more than one injury in the same location. As noted in Section 4.2.2, the samples of eight of the studies included people with multiple fractures, all of which reported per-patient data, which were extracted for the assessment. These data therefore do not fully capture the rates of fractures in the study samples and may overestimate the sensitivity of the technologies (since a true positive result may be acquired even if all fractures in the same person are not identified).

Table 5: Outcomes available from the included studies

First author (Date)	True positive	True negative	False positive	False negative	Sensitivity	Specificity	PPV	NPV	Reading time
BoneView									
Bousson (2023)	Y	Y	Y	Y	Y	Y	Y	Y	N
Cohen (2023)	Y	Y	Y	Y	Y	Y	Y	Y	N
Canoni-Meynet (2022)	Y	Y	Y	Y	Y	Y	Y	Y	Y
Dell-Aria (2024)	Y	Y	Y	Y	Y	Y	Y	Y	N
Duron (2021)	Y	Y	Y	Y	Y	Y	NA	NA	Y
Guermazi (2022)	Y	Y	Y	Y	Y	Y	NA	NA	Y
Meetschen (2024)	Y	Y	Y	Y	Y	Y	NA	NA	Y
Nguyen (2022)*	Y	Y	Y	Y	Y	Y	NA	NA	N
Oppenheimer (2023)	Y	Y	Y	Y	Y	Y	Y	Y	N
Rayvolve									
Bousson (2023)	Y	Y	Y	Y	Y	Y	Y	Y	N
Fu 2024	Y	Y	Y	Y	Y	Y	Y	Y	Y
RBFracture									
Bachmann (2024)	Y	Y	Y	Y	Y	Y	NA	NA	N
Jørgensen (S9 Abstract) (2023)	Y	Y	Y	Y	Y	Y	NA	NA	N
Radiobotics 2021 / Bonde	Y	Y	Y	Y	Y	Y	NA	NA	N
Ruitenbeek (2024)	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Yogendra (NA)*	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
TechCare Alert									
Bousson (2023)	Y	Y	Y	Y	Y	Y	Y	Y	N
Suite (2020)	Y	Y	Y	Y	Y	Y	NA	NA	N

Abbreviations: N, no; NA, not applicable; NPV, negative predictive value; positive predictive value; Y, yes

Notes: *studies conducted in children only.

4.3. Quality appraisal of studies

As this was an EVA, no formal quality assessment of the included studies was conducted. Quality considerations that were considered to represent a potential source of bias have been discussed throughout the previous sections.

An overview of the way quality considerations influenced the interpretation of diagnostic evidence and the selection of evidence to inform the economic analysis are shown in Table 6. Coloured ratings are provided for ease of reference, though the EAG emphasise that these ratings are relative to each other and none of the included evidence was considered to be robust for diagnostic outcomes.

Table 6: Quality considerations for the interpretation of diagnostic accuracy of the technologies

First author (date)	Index test and comparators	Good for sensitivity and specificity?	Good for prevalence, NPV and PPV?
Head-to-head study			
Bousson (2023)	BoneView/Rayvolve/TechCare Alert (no unassisted comparator)	**AMBER**. Limited by the reference standard including AI results	**AMBER**. Good due to consecutive sampling, but limited by the reference standard including AI results
BoneView			
Canoni-Meynet (2022)	BoneView assisted vs unassisted	**AMBER**. Limited by the reference standard including AI results	**AMBER**. Good due to consecutive sampling, but limited by the reference standard including AI results
Cohen (2023)	BoneView standalone + reader (no comparator) vs unassisted	**AMBER**. Limited by having no comparator (assisted only)	**GREEN**. Good due to consecutive sampling, NPV and PPV limited to assisted only, limited to wrist only
Dell-Aria (2024)	BoneView assisted vs unassisted	**AMBER**. Limited by relatively small sample size and reference standard decision by a single radiologist	**AMBER**. Good due to consecutive sampling, but limited by relatively small sample size and reference standard decision by a single radiologist

Duron (2021)	BoneView assisted vs unassisted	**GREEN**. Good for sensitivity and specificity estimates	**RED**. No, case-control design
Guermazi (2022)	BoneView assisted vs unassisted	**AMBER**. Limited by high number of obvious fractures	**RED**. No, case-control design
Meetschen (2024)	BoneView assisted (no comparator)	**AMBER**. Limited by relatively small sample size, also by having no comparator (assisted only)	**RED**. No, case-control design
Nguyen (2022)	BoneView assisted vs unassisted	**GREEN**. Good for sensitivity and specificity estimates	**RED**. No, case-control design
Oppenheimer (2023)	BoneView assisted vs unassisted	**AMBER**. Limited by the reference standard including AI results	**AMBER**. Good due to consecutive sampling, but limited by the reference standard including AI results
RBFracture			
Bachmann (2024)	RBFracture assisted vs unassisted	**GREEN**. Good for sensitivity and specificity estimates	**RED**. No, case-control design
Jørgensen (2023)	RBFracture assisted vs unassisted	**AMBER**. Limited by unclear reference standard and publication type (abstract)	**RED**. No, case-control design
Radiobotics (2021) and Bonde	RBFracture assisted vs unassisted	**AMBER**. Limited by publication type (not peer-reviewed publication)	**RED**. No, case-control design
Ruitenbeek (2024)	[redacted]	[redacted]	[redacted]
Yogendra (unpublished)	[redacted]	[redacted]	[redacted]
Rayvolve			
Fu (2024)	Rayvolve assisted vs unassisted	**AMBER**. Limited by relatively small sample size	**AMBER**. Good due to random sampling, but limited by relatively small sample size
TechCare			

Suite (2020)	TechCare assisted vs unassisted	**AMBER**. Limited by publication type (not peer-reviewed) and reference standard decision by a single radiologist	**RED**. No, case-control design

Abbreviations: NPV, negative predictive value; PPV, positive predictive value

5. CLINICAL, SERVICE AND TECHNOLOGICAL EVIDENCE RESULTS

5.1. Results from the evidence base and evidence synthesis

5.1.1. Diagnostic accuracy

The diagnostic accuracy of the technologies as reported by the included studies and/or calculated by the EAG are shown in Table 7. Data were calculated where these were not reported in publications. In some cases, calculated data varied slightly from that reported in the publication, which the EAG assumed was due to rounding errors, alternative methods for handling missing data, or reporting errors. The EAG selected the most reliable data for inclusion in the report tables and the evidence synthesis, meaning that some data (in all cases, rates of TP, FP, TN, FN) differed from those reported in the publications. As noted in Section 4.2.3, diagnostic data were calculated per patient (i.e. one result per patient, meaning that the analysis did not account for multiple fractures in the same person), unless otherwise indicated.

Only three studies[13 17 21] reported outcomes for readers based within emergency care settings, each evaluating a different technology: one study[13] compared assisted and unassisted readings with BoneView as read by emergency physicians; one study[21] reported assisted and unassisted readings with Rayvolve as read by emergency physicians; and one study[17] reported assisted and unassisted readings with RBFracture as read by A&E trainees or trauma-care nurses. Most of the included studies reported multiple analyses according to the level of experience of clinicians reading the X-rays. As the protocol for this assessment included a comparison of diagnostic accuracy across staff reading experience, evidence reported for readers of varying levels of experience was extracted. The EAG noted that those with more seniority and expertise in reading X-rays, would not be expected to use the technology in clinical practice and that these data are presented for comparison purposes only. To aid interpretation of the findings across reader experience, the EAG grouped results according to the description of reader experience and seniority described in the publications. As all of the included studies were based outside of the UK, staff grades used in the publications had unclear relevance to target clinicians within NHS settings, which meant that some groupings were uncertain. Three staff groupings were created: readers described by studies as having less experience with reading X-rays, including staff specified as being based within emergency settings as well as those stated to have less experience in reading X-rays; senior and expert staff, intended to be consistent with a consultant radiologist and reporting radiographer grade; and a mixed and unclear grouping, where results were reported for groups of readers who varied widely in experience level and

studies where reader descriptions did not fit easily within the less experienced and senior groups. The EAG noted that, in some studies, staff described as radiologists were nevertheless described as having less experience in reading X-rays, leading the EAG to assume that the term ‘radiologist’ may have a broader range of experience than would be expected within the NHS. In these cases, the EAG classified these as less experienced staff, though considered these groupings to be particularly uncertain, since descriptions in the publications may have been misleading.

Inspection of the results suggested that the groupings of studies according to reader experience was not particularly successful: in general, readers with greater seniority and expertise at reading X-rays were associated with more accurate diagnosis unassisted, though this was not universal across groupings. As this lacked face validity, the EAG assumed that the groupings according to reader experience were incorrect in some instances and so should be interpreted with caution when pooled (see Section 5.2).

Evidence from across the comparative studies generally suggested a trend for the technologies to improve sensitivity for diagnosing fractures in a mix of fracture types and across all reader groupings, but to have little or no impact on specificity. This trend was also present in those studies[13 17 21] specific to readers based within emergency care settings. The comparability of outcomes between studies in emergency care settings was uncertain due to variation in the sensitivity of unassisted emergency physicians considered in both the Duron (2021) and Fu (2024) studies (sensitivity was reported as 61.3% in Duron and 79.2% in Fu), suggesting that staff experience or approach to decision-making varied between the studies. Nevertheless, the technologies improved sensitivity by a similar amount, with little or no change in specificity. Assistance with RBFracture in Bachmann (2024) improved sensitivity for trauma care nurses also by a similar amount, also with some benefit to specificity. The benefit to sensivitity for A&E trainees with RBFracture in the same study was smaller, with minimal benefit to specificity. However, in this study, unassisted A&E trainees had higher accuracy for detecting fractures unassisted than trauma care nurses and were no worse than either of the emergency physicians in Duron (2021) and Fu (2024).

5.1.1.1. Evidence from selected pivotal studies

The selection of pivotal studies for informing the economic analysis is described previously in this report. Nevertheless, the EAG cautions again that these studies were selected for being the

most robust estimates within their design compared to other studies identified in the evidence base, but nevertheless still have limitations and should be interpreted with caution.

For BoneView, in a mixed group of readers including radiologists and ED doctors, Duron 2021 reported sensitivity as 79.4% (SD 7.4) when assisted by the technology and 70.8% (SD 12.5) when unassisted. Specificity was 93.6% (SD 4.6) assisted and 89.5% (SD 6.5) unassisted. All participants in this study were adults. Similarly, using readers who were radiology residents with variable levels of experience, Nguyen 2022 reported a higher sensitivity with AI assistance (82.67%, 95% CI 75.65, 88.36) than without assistance (73.17%, 95% CI 65.33, 80.07) and a higher AI assisted specificity (90.33%, 95% CI 84.43, 94.55) than without AI assistance (89.58%, 95% CI 83.55, 93.97). Nguyen 2022 reported subgroup data for children and young adults and these data are presented in section 5.1.2 alongside data for children from other studies.

For RBFracture the pivotal study of sensitivity and specificity estimates was Bachmann 2024 (albeit with the caveats described above). In this study, when readers were a mixed group of emergency care staff with a moderate level of experience, sensitivity was reported as higher with AI assistance (0.80, 95% CI 0.78-0.82) than without (0.72, 95% CI 0.70-0.73). Likewise, AI assisted specificity was higher (0.85, 95% CI 0.84-0.87) than unassisted (0.81, 95% CI 0.80-0.83). Subgroup data were also available for children (see section 5.1.2), and for more junior staff (Table 7).

Similarly, the pivotal Rayvolve study (Fu 2024), provided sensitivity and specificity data for a mixed group of readers (emergency physicians, non-MSK radiologists, and MSK radiologists) and reported a higher sensitivity with AI assistance (0.955, 95% CI 0.944, 0.964) than without assistance (0.865, 95% CI 0.848, 0.881). In this study, specificity was similar with and without AI assistance (0.831, 95% CI 0.817, 0.845 and 0.826, 95% CI 0.812, 0.840 respectively). Subgroup data were available for emergency physicians (Table 7). The TechCare Alert study (Suite 2020) provided data only for junior radiologists and data were sparse and without confidence intervals (Table 7).

The EAG noted that although these sensitivity and specificity estimates may appear to differ across technologies, due to clinical and methodological heterogeneity between studies, and the overall paucity of robust data, it remains difficult to comment on whether any AI technology performed better than another. Although the head-to-head study (Bousson 2023) provided sensitivity and specificity data across three of the included technologies, the EAG highlighted

that these data were likely limited to a greater extent than the pivotal studies by the inclusion of the AI results in the reference standard. This study also did not include readers likely to be involved in emergency assessments of x-rays in UK clinical practice.

The pivotal studies for sensitivity and specificity estimates did not report other DTA data, due to study design features previously described. The EAG emphasised the need for caution when interpreting all DTA results (all studies are likely to be very limited), and in particular any prevalence, PPV and NPV estimates provided in the studies. When looking at the PPV and NPV data, the most robust data were from Cohen 2023, but these estimates were only relevant to wrist fractures (Table 7).

5.1.1.2. PPV and NPV from other included studies

Other studies with potentially reasonable designs for estimating PPV and NPV were Dell-Aria (2024) for BoneView, Ruitenbeek (2024) for RB Fracture, and Fu (2024) for Rayvolve (noting that Ruitenbeek and Fu are limited by the use of retrospective study designs, and Dell-Aria 2024 by a relatively small sample size). Indeed, of these three studies, only Dell-Aria reported PPV and NPV data (Table 7). The EAG calculated PPV and NPV for the other two studies but highlighted the potential lack of generalisability to clinical practice. For Fu 2024, for a mixed group of readers, the EAG calculated PPV was slightly higher with Rayvolve than without (61.92 and 58.86 respectively), as was NPV (98.49 and 95.52 respectively). Similar results were found for the junior readers, i.e. the EAG calculated PPV was slightly higher with Rayvolve than without (64.75 and 60.63 respectively), as was NPV (97.97 and 93.44 respectively). For Ruitenbeek, for a mixed group of readers, the EAG calculated PPV was slightly higher with RBFracture than without ([redacted] and [redacted] respectively), as was NPV ([redacted] and [redacted] respectively). PPV and NPV were overall lower for the junior readers but were similarly higher with RBFracture than without (PPV [redacted] and [redacted] respectively; NPV [redacted] and [redacted] respectively).

Table 7: Diagnostic accuracy of included technologies (mixed fracture and age groups)

Author (date). Target fractures.	Prev (reference standard)	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear staff level assisted	Mixed or unclear staff level unassisted	Senior and highly experienced staff assisted	Senior and highly experienced staff unassisted
BoneView							
Bousson (2023). Age range NR; likely all or mostly adults. Hand, wrist, arm, elbow, shoulder, pelvis, hip, leg, knee, ankle, foot.	326/1500 (26.9%)	.	.	Radiology residents (4 years of residency). Sensitivity 91.3 (88.2-93.6); specificity 90.5 (89.1-92.3); PPV NR; NPV NR. TP 298 FP 112 TN 1062 FN 28	.	.	.
Canoni-Meynet (2022) Age 0.25 – 99 years; 80.2% were adults. All fractures excluding skull and face.	188/500 (37.6%)	Third year radiology resident. Sensitivity 89 (83−93); specificity 93 (89−95); PPV 88 (83−92); NPV 93 (90−96) TP 167 FP 23 TN 189 FN 21	Third year radiology resident. Sensitivity 68 (60−74); specificity 96 (93−98) ; PPV 91 (87−96); NPV 83 (79−87) TP 127 FP 12 TN 300 FN 61	.	.	Senior radiologist. Sensitivity 88 (83−93); specificity 99 (98−100) ; PPV : 98 (97−100); NPV 93 (91−96) TP 166 FP 2 TN 310 FN 22	Senior radiologist. Sensitivity 70 (63−77); specificity 96 (94−98); PPV 92 (88−97); NPV 84 (80−88) TP 132 FP 11 TN 301 FN 56
Dell-Aria (2024). Age range NR; mean age 39 years.	54/101 (53.9%) Injuries that required additional	AI assisted radiologist with <5 years' experience. Injuries that did not require other imaging. Sensitivity	Unassisted radiologist with <5 years' experience. Injuries that did not require other imaging. Sensitivity	.	.	AI assisted radiologist with >15 years' experience Injuries that did not require other imaging. Sensitivity	Unassisted radiologist with >15 years' experience Injuries that did not require other imaging. Sensitivity

Author (date). Target fractures.	Prev (reference standard)	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear staff level assisted	Mixed or unclear staff level unassisted	Senior and highly experienced staff assisted	Senior and highly experienced staff unassisted
Low velocity trauma to upper or lower limbs (including shoulder and hip).	imaging: 32/51 (62.7%)	77.27; specificity 88.29; PPV 85; NPV 83.33 Injuries that did require additional imaging. Sensitivity 56.25; specificity 78.95; PPV 81.82; NPV 51.72 TP 18 FP 4 TN 15 FN 14	50.0; specificity 82.14; PPV 68.75; NPV 67.65 Injuries that did require additional imaging. Sensitivity 31.25; specificity 89.47; PPV 83.33; NPV 43.59 TP 10 FP 2 TN 17 FN 22			95.45; specificity 89.29; PPV 87.5; NPV 96.15 Injuries that did require additional imaging. Sensitivity 81.25; specificity 89.47; PPV 98.86; NPV 73.91 TP 26 FP 2 TN 17 FN 6	86.36; specificity 89.29; PPV 86.36; NPV 89.29 Injuries that did require additional imaging. Sensitivity 56.25; specificity; 89.47; PPV 90; NPV 54.84 TP 18 FP 2 TN 17 FN 14
Duron (2021). Age range 18 – 100 years; mean 57. Shoulder, arm, hand, pelvis, leg, foot	300/600 (50%)	Assisted emergency physicians Sensitivity 74.3 (SD 6.6); specificity 96.6 (1.9) TP 223 FP 10 TN 290 FN 77	Unassisted emergency physicians Sensitivity 61.3 (SD 9.3); specificity 90.6 (5.8) TP 184 FP 28 TN 272 FN 116	AI assisted six radiologists and six emergency physicians, including residents and experts. Sensitivity 79.4% (SD: 7.4); specificity 93.6% (4.6); PPV NR; NPV NR TP 238 FP 19 TN 281 FN 62	Unassisted six radiologists and six emergency physicians, including residents and experts. Sensitivity 70.8% (SD: 12.5); specificity 89.5% (6.5); PPV NR; NPV NR TP 212 FP 31 TN 269 FN 88	.	.
Meetschen (2024).	100/200 (50%)	.	.	Radiology residents 4.5 to	A radiology resident without AI.		

Author (date). Target fractures.	Prev (reference standard)	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear staff level assisted	Mixed or unclear staff level unassisted	Senior and highly experienced staff assisted	Senior and highly experienced staff unassisted
Age range 1 – 95 years; mean 40.7. Hand, wrist, arm, elbow, shoulder, scapula, clavicle, ribs, spine, pelvis, hip joints, legs, knees, ankles, and feet				24.5 months of experience). Sensitivity 77 (72 - 82); specificity 79 (73 -84); PPV 81% (76 - 86); NPV 75% (69 - 80) TP 77 FP 21 TN 79 FN 23	Sensitivity 58 (52-64); specificity 77 (71-81); PPV 74 (67-79); NPV 62 (56-67) TP 58 FP 23 TN 77 FN 42		
Nguyen (2022). Children and young people, age range 2 – 21 years; mean 10.8. Appendicular skeleton.	150/300 (50%)	.	.	AI assisted, eight readers: 5 radiology residents (between the 2nd and 4th year of residency) and 3 expert paediatric radiologists (at least 7 years of experience, including >3 years specialising in paediatric radiology. Sensitivity 82.67 [75.65, 88.36]; specificity 90.33 [84.43, 94.55]; PPV NR; NPV NR TP 124 FP 15	Unassisted eight readers: 5 radiology residents (between the 2nd and 4th year of residency) and 3 expert paediatric radiologists (at least 7 years of experience, including >3 years specialising in paediatric radiology . Sensitivity 73.17 [65.33, 80.07]; specificity 89.58 [83.55, 93.97]; PPV NR; NPV NR TP 110 FP 16 TN 134	.	.

Author (date). Target fractures.	Prev (reference standard)	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear staff level assisted	Mixed or unclear staff level unassisted	Senior and highly experienced staff assisted	Senior and highly experienced staff unassisted
				TN 135 FN 26	FN 40		
Oppenheimer (2023).^ Age range 2 – 100 years; mean 61.4. All excluding cervical spine, skull and face.	367/1163 (31.56%)	.	.	Resident radiologist with AI Sensitivity 91.28 (91.25, 91.31); specificity 97.36 (97.35, 97.37); PPV 94.10 (94.08, 94.12); NPV 96.03 (96.02, 96.04) TP 335 FP 21 TN 775 FN 32	Resident radiologists without AI Sensitivity 84.74 (84.70, 84.78); specificity 97.11 (97.10, 97.12); PPV 93.11 (98.08, 93.14); NPV 93.24 (93.22, 93.26) TP 311 FP 23 TN 773 FN 56	.	.
Rayvolve							
Bousson (2023). Age range NR; likely all or mostly adults. Hand, wrist, arm, elbow, shoulder, pelvis, hip, leg, knee, ankle, foot.	326/1500 (26.9%)			AI assisted radiology residents (4 years of residency) Sensitivity 92.6 (90.1-94.6); specificity 70.4 (68.1-73); PPV NR; NPV NR TP 302 FP 348 TN 826 FN 24			
Fu (2024).# Adults aged ≥21	587/2626 (22.4%)	Emergency physician (assisted).	Emergency physician (unassisted).	Eight each of emergency physicians, non-	Eight each of emergency physicians, non-	.	.

Author (date). Target fractures.	Prev (reference standard)	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear staff level assisted	Mixed or unclear staff level unassisted	Senior and highly experienced staff assisted	Senior and highly experienced staff unassisted
years. Ankle, clavicle, elbow, forearm, humerus, hip, knee, pelvis, shoulder, tibia/fibula, wrist, hand, foot		Sensitivity 0.938 (0.915, 0.955); specificity 0.853 (0.828, 0.875); PPV NR; NPV NR TP 551 FP 300 TN 1739 FN 36	Sensitivity 0.792 (0.757, 0.824); specificity 0.852 (0.828, 0.874); PPV NR; NPV NR TP 465 FP 302 TN 1737 FN 122	MSK radiologists, and MSK radiologists. Sensitivity 0.955 (0.944, 0.964); specificity 0.831 (0.817, 0.845); PPV NR; NPV NR TP 508 FP 355 TN 1684 FN 79	MSK radiologists, and MSK radiologists Sensitivity 0.865 (0.848, 0.881); specificity 0.826 (0.812, 0.840); PPV NR; NPV NR TP 561 FP 345 TN 1694 FN 26		
RBFracture							
Bachmann (2024). Age range NR; 70.7% adults. Appendicular skeleton	164/334 (49.1%)	Assisted A&E Trainees. Sensitivity 0.83 (0.81;0.86); specificity 0.90 (0.87;0.92); PPV NR; NPV NR TP 136 FP 17 TN 153 FN 28 Trauma-care nurses Sensitivity 0.70 (0.65;0.75); specificity 0.67 (0.62;0.72) TP 115	Unassisted A&E Trainees. Sensitivity 0.74 (0.71;0.78); specificity 0.87 (0.84;0.89); PPV NR; NPV NR TP 121 FP 22 TN 148 FN 43 Trauma-care nurses Sensitivity 0.58 (0.53;0.64) specificity 0.60 (0.55;0.65) TP 95	2 advanced trauma care nurses, 3 diagnostic radiographers, 4 A&E trainees, 3 orthopaedic specialty registrars, 3 radiology specialty registrars. Sensitivity 0.80 (0.78-0.82); specificity 0.85 (0.84-0.87); PPV NR; NPV NR TP 131 FP 25 TN 145 FN 33	Unassisted 2 advanced trauma care nurses, 3 diagnostic radiographers, 4 A&E trainees, 3 orthopaedic specialty registrars, 3 radiology specialty registrars. Sensitivity 0.72 (0.70-0.73); specificity 0.81 (0.80-0.83); PPV NR; NPV NR TP 118 FP 32 TN 138 FN 46	.	.

Author (date). Target fractures.	Prev (reference standard)	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear staff level assisted	Mixed or unclear staff level unassisted	Senior and highly experienced staff assisted	Senior and highly experienced staff unassisted
		FP 56 TN 114 FN 49	FP 68 TN 102 FN 69				
Ruitenbeek (S11 Abstract) (2024). [redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.	.
Yogendra (S7 Manuscript) (NA) [redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
TechCare Alert							
Bousson (2023). Age range NR; likely all or	326/1500 (26.9%)	.	.	Assisted radiology residents (4 years of residency) Sensitivity 90.2 (87.2-92.8); specificity 92.5	.	.	.

Author (date). Target fractures.	Prev (reference standard)	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear staff level assisted	Mixed or unclear staff level unassisted	Senior and highly experienced staff assisted	Senior and highly experienced staff unassisted
mostly adults. Hand, wrist, arm, elbow, shoulder, pelvis, hip, leg, knee, ankle, foot.				(91.1-94); PPV NR; NPV NR TP 294 FP 88 TN 1086 FN 32			
Suite (2020). Adults aged 18 – 98 years; mean age 53.6. Lower limbs, upper limbs, ribs	253/620 (40.8%)	Four junior radiologists with AI. Sensitivity 95; specificity 98; PPV 97; NPV 96 TP 240 FP 7 TN 360 FN 13	Junior radiologist without AI. Sensitivity 92; specificity 97; PPV 96; NPV 94 TP 233 FP 11 TN 356 FN 20	.	.	Senior radiologist with AI. Sensitivity 93; specificity 98; PPV 97; NPV 96 TP 235 FP 7 TN 360 FN 18	Senior radiologist without AI. Sensitivity 93; specificity 98; PPV 97; NPV 96 TP 235 FP 7 TN 360 FN 18

Abbreviations: CI, confidence interval; FN, false negative; FP, false positive; MSK, musculoskeletal; NA, not applicable; NPV, negative predictive value; PPV, positive predictive value; prev, prevalence; TN, true negative; TP, true positive

Notes: * calculated based on a crude midpoint between the two readers. ^ data is exam-wise. # data is fracture-wise.

5.1.2. Subgroup results (paediatric participants)

Four studies reported diagnostic accuracy data in a sample of children and young people: two studies[20 25] were conducted only with children and young people, and are also included in the previous section, and two studies[16 17] reported subgroup data for children and young people. These data were in a mixed fracture population only and were only available in mixed/unclear and senior/expert reader groupings. The data are summarised in Table 8.

One of the studies[16] evaluating BoneView reported 100% sensitivity for both assisted and unassisted readers. In other studies,[17 20 25] the use of assisted readings with BoneView and RBFracture improved sensitivity for detecting fractures in children and young people but made no difference to specificity. Sensitivity for assisted diagnosis in these studies was high but would nevertheless result in more than 10% of positive fractures being missed.

Table 8: Diagnostic accuracy data for children and young people

Author (date).	Prevalence (reference standard)	Test	Mixed or unclear reader assisted	Mixed or unclear reader unassisted	Senior and expert staff assisted	Senior and expert staff unassisted
BoneView						
Nguyen (2022). Appendicular skeleton.	150/300 (50%)	BoneView	AI assisted, eight readers: 5 radiology residents (between the 2nd and 4th year of residency) and 3 expert paediatric radiologists (at least 7 years of experience, including >3 years specialising in paediatric radiology. Sensitivity 82.67 [75.65, 88.36]; specificity 90.33 [84.43, 94.55]; PPV NR; NPV NR TP 124 FP 15 TN 135 FN 26	Unassisted eight readers: 5 radiology residents (between the 2nd and 4th year of residency) and 3 expert paediatric radiologists (at least 7 years of experience, including >3 years specialising in paediatric radiology. Sensitivity 73.17 [65.33, 80.07]; specificity 89.58 [83.55, 93.97]; PPV NR; NPV NR TP 110 FP 16 TN 134 FN 40	.	.
Oppenheimer (2023) All fractures excluding cervical spine, skull and face.	6/31 (19.4%)	BoneView	Resident radiologist with AI Sensitivity 100%; specificity 92% TP 6	Resident radiologist without AI Sensitivity 100%; specificity 92% TP 6	.	.

Author (date).	Prevalence (reference standard)	Test	Mixed or unclear reader assisted	Mixed or unclear reader unassisted	Senior and expert staff assisted	Senior and expert staff unassisted
			FP 2 TN 23 FN 0	FP 2 TN 23 FN 0		
RBFracture						
Bachmann (2024). Appendicular skeleton.	49/98 (50%)*	RBFracture	2 advanced trauma care nurses, 3 diagnostic radiographers, 4 A&E trainees, 3 orthopaedic specialty registrars, 3 radiology specialty registrars. Sensitivity 0.89 (0.87, 0.92); specificity 0.80 (0.77, 0.83) TP 44 FP 10 TN 39 FN 5	2 advanced trauma care nurses, 3 diagnostic radiographers, 4 A&E trainees, 3 orthopaedic specialty registrars, 3 radiology specialty registrars. Sensitivity 0.78 (0.74, 0.81); specificity 0.77 (0.74, 0.80) TP 38 FP 11 TN 38 FN 11	.	.
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Abbreviations: NPV, negative predictive value; NR, not reported; PPV, positive predictive value

Notes. *estimated. Prevalence in paediatric population wasn't reported, but case-control specified 50% fracture rate in overall sample and stratified by age

5.1.3. Subgroup results (fracture type)

The EAG identified diagnostic evidence separately for hand/wrist fractures, foot/ankle fractures, hip/pelvis fractures and Salter-Harris II fractures. These data are shown in Table 9.

Data for hand/wrist fractures were available for BoneView, Rayvolve and Smarturgences (Table 9). For BoneView, the studies with the most robust sensitivity and specificity data were likely to be Duron 2021 and Nguyen 2022 (albeit limited by their retrospective designs). For hand fractures only, and a mixed group of readers, Duron 2021 reported sensitivity to be higher with BoneView assistance than without (80.2%, SD 11.4 and 59.6%, SD 20.5 respectively). Specificity was also higher with BoneView assistance than without (91.0%, SD 6.4 and 84.7%, SD 11.0 respectively). Similarly, in Nguyen 2022, for a mixed group of readers interpreting hand/wrist X-rays, sensitivity was higher with BoneView assistance than without (87.08 95% CI 69.79, 96.46 and 68.75 95% CI 49.31, 84.32 respectively) although specificity was similar with and without BoneView assistance (88.33 95% CI 71.34, 97.1 and 87.92 95% CI 85.44, 89.48 respectively). PPV and NPV data were not available for these studies but are provided by Cohen 2023 in Table 9. For Rayvolve and Smarturgences one the head-to-head study provided subgroup data for hand/wrist fractures. The EAG again highlight the need to interpret these results with caution; the study was a retrospective which included the results of the AI in the reference standard.

For foot/ankle fractures, data were also available for BoneView, Rayvolve and Smarturgences (Table 9). Again, for Rayvolve and Smarturgences only the limited head-to-head study provided these subgroup data. For BoneView, the studies with the most robust sensitivity and specificity data were again likely to be Duron 2021 and Nguyen 2022 (albeit limited by their retrospective designs). For foot fractures only, and a mixed group of readers, Duron 2021 reported sensitivity to be higher with BoneView assistance than without (86.9%, SD 8.3 and 71.8% SD 13.6 respectively). Specificity was also higher with BoneView assistance than without (92.9% SD 5.8 and 88.0% SD 9.9 respectively). Similarly, in Nguyen 2022, for a mixed group of readers interpreting foot/ankle X-rays, sensitivity was higher with BoneView assistance than without (70.83 95% CI 51.47, 85.89 and 70.83 95% CI 51.47, 85.89 respectively). Specificity with BoneView assistance was 86.25 95% CI 68.77, 96.02 and without assistance was 85.83 95% CI 68.26, 95.79.

Although hip fracture data were available for BoneView and RBFracture (Table 9), the EAG noted that none of the pivotal studies reported these data. For BoneView, only Oppenheimer

2023 provided subgroup data for hip fractures, and despite being a prospective study, was limited by the inclusion of the AI results in the reference standard. Hip fracture data for RBFracture was either methodologically limited and unpublished (Bonde and Radiobotics 2021) or limited data from a published abstract (Jørgensen 2024, Table 9).

Only one study (Nguyen 2022; BoneView) provided data for Salter-Harris II fractures in children. As previously noted, although this study may provide reasonable sensitivity and specificity estimates, there is uncertainty surrounding this due to the retrospective study design. BoneView assisted sensitivity, for a mixed group of readers, was given as 92.26, 95% CI 71.29, 99.4 and unassisted as 80.95, 95% CI 57.41, 94.78. Specificity was not reported (Table 9).

Table 9: Diagnostic accuracy data for different fracture locations

Author (date).	Prevalence (reference standard)	Test	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear reader assisted	Mixed or unclear reader unassisted	Senior and expert readers assisted	Senior and experts unassisted
Hand/wrist								
Bousson (2023)	97/314 (30.9%)	BoneView			Radiology residents (4 years of residency). Sensitivity 91.5 (84.9-95.6); specificity 92.1 (89.2-95.8) TP 89 FP 17 TN 200 FN 8			
Canoni-Meynet (2022)^	38/NR	BoneView	.	.	Radiologists assisted sensitivity: Hand 89.5% (33.9/38)	Radiologists unassisted sensitivity: Hand 68.4% (25.9/38)	.	.
Cohen (2023).	247/637 (38.8%);	BoneView	.	.	.	.	Artificial combination of AI + initial radiology report. Sensitivity 88%(84–92); specificity 92%(89–95); PPV 88% (84–92); NPV 93% (90–95) TP 218 FP 29	Initial radiology report. Sensitivity 76 (70, 81); specificity 96 (94, 98); PPV 93 (90, 97); NPV87 (83, 90) TP 189 FP 14 TN 376 FN 58

Author (date).	Prevalence (reference standard)	Test	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear reader assisted	Mixed or unclear reader unassisted	Senior and expert readers assisted	Senior and experts unassisted
							TN 361 FN 29	
Duron (2021)	Hand only: 44/108 (40.7%)	BoneView			Six radiologists and six emergency physicians Sensitivity 80.2% (SD: 11.4); specificity 91.0% (SD 6.4) TP 35 FP 6 TN 58 FN 9	Six radiologists and six emergency physicians Sensitivity 59.6% (SD: 20.5); specificity 84.7% (SD 11.0) TP 26 FP 10 TN 54 FN 18		
Nguyen (2022)	30/60 (50%)	BoneView			Average across 5 radiology residents and 3 expert paediatric radiologists. Sensitivity 87.08 [69.79, 96.46]; specificity 88.33 [71.34, 97.1] TP 26 FP 4 TN 26 FN 4	Average across 5 radiology residents and 3 expert paediatric radiologists. Sensitivity 68.75 [49.31, 84.32]; specificity 87.92 [85.44, 89.48] TP 21 FP 4 TN 26 FN 9		
Oppenheimer (2023)	NR	BoneView			Resident radiologist.	Resident radiologist		

Author (date).	Prevalence (reference standard)	Test	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear reader assisted	Mixed or unclear reader unassisted	Senior and expert readers assisted	Senior and experts unassisted
					Sensitivity 95.65%	Sensitivity 78.26%		
Bousson (2023)	97/314 (30.9%)	Rayvolve			Radiology residents (4 years of residency). Sensitivity 97.8 (94.5-99); specificity 74.6 (70.1-80.9) TP 95 FP 55 TN 162 FN 2	.		
Bousson (2023)	97/314 (30.9%)	TechCare Alert			Radiology residents (4 years of residency). Sensitivity 93.6 (88-96.3); specificity 91.7 (89-95.3) TP 91 FP 18 TN 199 FN 6			
Foot/ankle								
Bousson (2023)	Foot: 56/186 (30.1%), ankle 42/232 (18.1%)	BoneView			Radiology residents (4 years of residency). Sensitivity Foot 98.1% (94.4, 98.3), Ankle 89.9% (79.4, 95.4%)			

Author (date).	Prevalence (reference standard)	Test	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear reader assisted	Mixed or unclear reader unassisted	Senior and expert readers assisted	Senior and experts unassisted
					Foot TP 55 FP 27 TN 103 FN 1 Ankle TP 38 FP 13 TN 177 FN 4			
Canoni-Meynet (2022)^	39/NR	BoneView	.	.	Radiologists assisted sensitivity: Foot 82.9% (32.3/39)	Radiologists unassisted sensitivity: Foot 57.3% (22.3/39)	.	.
Duron (2021)	Foot only 44/84 (52.3%)	BoneView			Six radiologists and six emergency physicians Sensitivity 86.9% (SD: 8.3); specificity 92.9% (SD 5.8)	Six radiologists and six emergency physicians Sensitivity 71.8% (SD: 13.6); specificity 88.0% (SD 9.9)		
Nguyen (2022)	30/60 (50%)	BoneView			Average across 5 radiology residents and 3 expert paediatric radiologists. Sensitivity 70.83 [51.47, 85.89]; specificity	Average across 5 radiology residents and 3 expert paediatric radiologists. Sensitivity 53.75 [34.71, 72.02] specificity		

Author (date).	Prevalence (reference standard)	Test	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear reader assisted	Mixed or unclear reader unassisted	Senior and expert readers assisted	Senior and experts unassisted
					86.25 [68.77, 96.02] TP 21 FP 4 TN 26 FN 9	85.83 [68.26, 95.79] TP 16 FP 4 TN 26 FN 14		
Oppenheimer (2023)	NR	BoneView			Resident radiologist. Sensitivity 88.57%	Resident radiologist. Sensitivity 82.86%		
Bousson (2023)	Foot: 56/186 (30.1%), ankle 42/232 (18.1%)	Rayvolve			Radiology residents (4 years of residency). Sensitivity Foot 90.8% (83.0, 95.0), Ankle 92.1% (82.2, 96.5) Foot TP 51 FP 49 TN 81 FN 5 Ankle TP 39 FP 53 TN 137 FN 3			
Bousson (2023)	Foot: 56/186 (30.1%), ankle 42/232 (18.1%)	Smarturgences			Radiology residents (4 years of residency). Sensitivity Foot 85.4% (73.1,			

Author (date).	Prevalence (reference standard)	Test	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear reader assisted	Mixed or unclear reader unassisted	Senior and expert readers assisted	Senior and experts unassisted
					90.2), Ankle 89.9% (76.0, 95.5) Foot TP 48 FP 12 TN 118 FN 8 Ankle TP 38 FP 15 TN 175 FN 4			
Hip								
Oppenheimer (2023)	NR	BoneView			Resident radiologist. Sensitivity 93.22%	Resident radiologist. Sensitivity 93.22%		
Jørgensen (S9 Abstract) (2023).	107/214 (50%)	RBFracture	.	.	Two radiographers, two medical interns and two consultants. Sensitivity 0.96 (CI: 0.95; 0.98); specificity 0.86 (CI: 0.83; 0.89); PPV NR; NPV NR TP 103 FP 15 TN 92 FN 4	Two radiographers, two medical interns and two consultants. Sensitivity 0.93 (CI: 0.91;0.95); specificity 0.87 (CI: 0.85; 0.90); PPV NR; NPV NR TP 100 FP 14 TN 93 FN 7	.	.

Author (date).	Prevalence (reference standard)	Test	Less experienced readers assisted	Less experienced readers unassisted	Mixed or unclear reader assisted	Mixed or unclear reader unassisted	Senior and expert readers assisted	Senior and experts unassisted
Radiobotics (2021); Bonde [unpublished]; confidential study report. Non-obvious proximal femur fractures, including the femoral neck and head	[redacted] Mixed or unclear readers 156/312 (50%)	RBFracture	[redacted]	[redacted]	Readers with >2 years' experience plus AI. Sensitivity [redacted]; specificity [redacted]; PPV [redacted]; NPV [redacted]	Readers with >2 years' experience without AI. Sensitivity [redacted]; specificity [redacted]; PPV [redacted] NPV [redacted]	.	.
Salter-Harris								
Nguyen (2022)	21/NR	BoneView			Average across 5 radiology residents and 3 expert paediatric radiologists. Sensitivity 92.26 [71.29, 99.4]; specificity NR;	Average across 5 radiology residents and 3 expert paediatric radiologists. Sensitivity 80.95 [57.41, 94.78]; specificity NR		

Notes: ^ fracture-wise data

5.1.4. X-ray reading time

X-ray reading time with and without AI assistance was available for three technologies: BoneView (4 studies[10 11 13 14]), RBFracture (2 studies[17 20], though only 1 study[20] [redacted]) and Rayvolve (1 study[21]). The data for all fracture types (i.e. not for specific fracture locations) is shown in Table 10. Consistent with the diagnostic accuracy data, the EAG reported these data separately according to the perceived level of experience of readers, as described in the publications. There were no noticeable differences in reading time across the staff groupings, which was surprising given that reading time would be expected to be shorter for more senior staff.

Studies reported that BoneView and Rayvolve were both associated with a reduction in x-ray reading time across all staff groups: a reduction of 7 seconds per X-ray between means with Rayvolve, while reductions between means ranged from 2.6 to 13 seconds per X-ray with BoneView. One study reported that RBFracture [redacted]. However, there were large standard deviations around reading time in all studies, which the EAG assumed may be due in part to the reading time varying widely according to the type and complexity of the fracture/injury. The EAG was also concerned about the reliability of how reading time would be measured in studies, and potential differences in the way this was defined and recorded between studies. This meant that the EAG had serious concerns about the reliability of the data and how to interpret differences between studies.

The EAG considered that reduced reading time for X-rays may have benefits for service resource, though were uncertain to what extent a difference of seconds per X-ray would mean to a service. However, the EAG also considered it plausible that increased reading times may be acceptable (and preferable) where the additional time translated to increased accuracy. Overall, based on the evidence available, the EAG tentatively concluded that BoneView may be associated with faster reading X-ray reading time, and that these findings should be considered alongside the results for its diagnostic accuracy.

Table 10: Reading time for assisted and unassisted X-rays (all fracture types)

	Less experienced staff		Mixed or unclear staff		Senior	
	With AI mean seconds per x-ray (SD)	**Without AI mean seconds per x-ray (SD)**	**With AI mean seconds per x-ray (SD)**	**Without AI mean seconds per x-ray (SD)**	**With AI mean seconds per x-ray (SD)**	**Without AI mean seconds per x-ray (SD)**
BoneView (Gleamer)	52.8 (17.7) [Duron] 27 (18); range 6, 114 [Canoni-Meynet] 44 (43) [Guermazi]	61.5 (24.8) [Duron] 39 (34); range 4, 313 [Canoni-Meynet] 57 (54) [Guermazi]	57.0 (49.4) [Duron] 49.2 (28.5) [Guermazi] 29.6 (19.8) [Meetschen]	67.0 (59.3) [Duron] 55.5 (32.6) [Guermazi] 32.2 (20.8) [Meetschen]	40 (24); range 7, 163 [Canoni-Meynet]	50 (28); range 10, 180 [Canoni-Meynet]
RBFracture (Radiobotics)	-	-	[redacted] [Yogendra] 46.4 (NR) [Bachmann]	[redacted]) [Yogendra]	[redacted] [Yogendra]	[redacted] [Yogendra]
Rayvolve (AZMed)	18 (NR) [Fu]	25 (NR) [Fu]	19 (NR) [Fu]	26.1 (NR) [Fu]	-	-

Abbreviations: NR, not reported; SD, standard deviation

5.2. Evidence synthesis

The EAG investigated whether it was possible to meta-analyse data from the included studies, for example to identify a pooled estimate of sensitivity and specificity for a particular technology. Having identified significant variation in the diagnostic accuracy results according to reader experience, an assessment of the feasibility of meta-analysis was conducted with study results grouped according to the level of experience of readers outlined in Section5.1, as well as by fracture and technology type. A threshold of six studies within each grouping was considered sufficient for meta-analysis. The assessment identified that there were sufficient studies for meta-analysis in three categories:

(1) accuracy of unassisted emergency department and less experienced clinicians in reading x-rays of any fracture type

(2) accuracy of unassisted mixed and unclear groups of clinicians in reading x-rays of any fracture type, and

(3) accuracy of BoneView when assisting a mixed or unclear group of clinicians reading X-rays of any fracture type.

On further investigation, however, there was unexplained heterogeneity in the results of the studies within each analysis, and meta-analysis was therefore not considered to be feasible. Specifically, plots of accuracy data varied significantly and there was a clear positive correlation between logit sensitivity and specificity, which suggested that at least one meaningful covariate was not included in the analysis[29]. Within the EVA, it was not feasible to investigate further the potential reasons for heterogeneity in the data, which might have included a meta-regression to investigate factors that influence the accuracy of assisted and unassisted diagnosis.

The EAG took two further steps to synthesise the evidence base: (1) data from the included studies within each grouping was summarised using median and ranges, to give a concise insight into the variability of results across studies and (2) conducted a narrative synthesis to identify patterns in the data that could be used to inform an understanding about the potential value of the technology for assisting in the diagnosis of fractures. Synthesised data from the included studies is split by fracture type (general/all fracture types and specific fracture locations). The EAG advises that the synthesised results provide the median and range of sensitivity and specificity values reported in the studies and do not include any variance (e.g. 95% confidence intervals) around the data reported in the studies. This approach was used for simplicity and, due to limitations in the evidence base as a whole, including unexplained heterogeneity across studies, the results are presented to provide an insight into potential patterns across the dataset, rather than to identify precise diagnostic accuracy data for the technologies.

5.2.1. Diagnostic accuracy of the technologies across studies

Median sensitivity and specificity for each of the technologies as reported across study groupings are presented in this section. For ease of interpretation, the results are also presented as the median proportion of missed fractures and false positives. Section 5.2.1.1 includes the full study population included in the studies, representing multiple fracture types within the case mix of each study, as well as results for specific types of fracture subgroups (hand/wrist, foot/ankle, hip, and non-obvious fractures). Section 5.2.1.2 includes results specific to children and young people.

5.2.1.1. Mixed and subgroup fractures

Results for unassisted readers are shown in Table 11. The EAG noted that the rate of missed fractures for clinicians reading X-rays without AI assistance was high across studies, even for the senor and expert reader grouping (which was intended to be consistent with consultant-level radiologists and reporting radiographers). Across studies, the proportion of fractures missed ranged from 20 – 30% and was consistent across reader experience, though more experienced readers gave very few false positive decisions. The EAG consulted with two radiographers (authors NG and RM) to enquire whether the accuracy of unassisted diagnosis reported in the papers would be consistent with their expectations in clinical practice. They advised that the rate of missed fractures reported for more experienced staff in the included studies was higher than they expected; they expected that consultant radiologists and radiographers would be expected to miss very few fractures using X-rays (10% or less), even where readers were unable to consult medical notes and the results of other imaging modalities, as was typical in the studies. This would also be consistent with guidelines from The Royal College of Radiologists[30]. The EAG sought robust data for the accuracy of X-ray for identifying fractures as used by readers of varying experience but was unable to identify this during the assessment. The EAG considered this to be a significant uncertainty in the evidence base, since uncertainty surrounding the generalisability of the data from the unassisted arms of the studies would also affect the way in which the results for the technology reported in the studies should be interpreted.

Sensitivity and specificity each varied significantly across studies though, in general, unassisted readers had higher specificity when identifying fractures than sensitivity. This resulted in a high median rate of missed fractures in all fracture analyses, though the rate of false positives was also over 10% in most studies in all reader groupings except senior and expert readers.

The accuracy of unassisted readers for detecting hip fractures was high. Sensitivity for detecting hand/wrist and foot/ankle fractures was lower than the mixed fracture analyses, and there was variability in specificity for detecting hand/wrist fractures across studies. As might be anticipated, there was greater variability in accuracy across studies reporting findings for groups of readers with mixed levels of experience, which likely represents the variable case mix in readers between studies. There was poorer sensitivity for identifying non-obvious fractures across all reader groupings.

Table 11: Diagnostic accuracy of unassisted diagnosis (no AI) across studies

Group (n studies)	Median sens (range)	Median spec (range)	Median % missed #s (range)	Median % over diagnosis (range)
All fracture analyses				
Unassisted, any staff grouping, all fractures (11)[10 11 13 14 16 17 20-22 25 27]	0.72 (0.31, 0.93)	0.89 (0.60, 1.00)	28.1% (7.1, 42.1)	13.3% (1.4, 40.0)
Unassisted, less experienced staff, all fractures (7)[11 13 14 17 21 22 27]	0.70 (0.58, 0.92)	0.87 (0.60, 0.97)	30.4% (7.9, 42.1)	12.9% (3.0, 40.0)
Unassisted, mixed or unclear staff, all fractures (9)[10 13 14 16 17 20 21 25 27]	0.73 (0.58, 0.87)	0.90 (0.77, 0.97)	26.7% (13.5, 42.0)	10.7% (2.9, 23.0)
Unassisted, senior and expert staff, all fractures (3)[11 20 22]	[redacted]	[redacted]	[redacted]	[redacted]
Subgroup fracture types				
Unassisted, mixed or unclear staff, hand/wrist (3) [13 16 25]	0.69 (0.60, 0.78)	0.88 (0.85, 0.96)	35.5% (30.0, 40.9)	14.5% (13.3, 15.6)
Unassisted senior and expert staff, hand/wrist (2)[9 11]	0.72 (0.68, 0.76)	0.96 (NA)	23.5% (NA)	3.6% (NA)
Unassisted mixed or unclear staff, foot/ankle (3)[13 16 25]	0.72 (0.54, 0.83)	0.88 (0.86, 0.88)	37.0% (27.3, 46.7)	12.9% (12.5, 13.3)
Unassisted senior and expert staff, foot/ankle (1)[11]	0.57 (NA)	NR	NR	NR
Unassisted less experienced staff, hip (1)[7]	[redacted]	[redacted]	[redacted]	[redacted]
Unassisted mixed or unclear staff, hip (3)[28]	[redacted]	[redacted]	[redacted]	[redacted]
Unassisted less experienced staff, non-obvious fractures (1)[24]	0.31 (NA)	0.89 (NA)	68.8% (NA)	10.5% (NA)
Unassisted mixed or unclear staff, non-obvious fractures (1)[16]	0.72 (NA)	NR	28.5% (NA)	100.0% (NA)
Unassisted senior or expert staff, non-obvious fractures (1)[24]	0.56 (NA)	0.89 (NA)	43.8% (NA)	10.5% (NA)

Abbreviations: NA, not applicable; sens, sensitivity; spec, specificity

Note: data are the median and range of sensitivity and specificity values reported in studies or calculated by the EAG within each grouping. No variance data around the data are provided, however the specific values should be considered to be uncertain.

The results for BoneView as evaluated across studies are shown in Table 12. BoneView showed high sensitivity and specificity, irrespective of the reader group. Nevertheless, median numbers of missed fractures (all fracture analyses) exceeded 15% for all readers except the senior and expert reader group. As with unassisted results, there was some variability in sensitivity and specificity values reported across studies, particularly in analyses with mixed and unclear readings. In general, BoneView had improved specificity relative to sensitivity, with fewer false positives than missed fractures.

Sensitivity for non-obvious fractures was improved compared to the results for unassisted, although the rate of missed fractures and false positives was still high (43.8% and 21.1%) respectively, suggesting that services may still wish to use precautionary policies to avoid the risk of missed fractures. There was an improvement in sensitivity and specificity for detecting hand/wrist and foot/ankle fractures relative to unassisted. There was very little evidence for the accuracy of BoneView for identifying hip fractures, which reported high accuracy (both sensitivity and specificity), though not conclusively different from some of the unassisted evidence.

Table 12: Diagnostic accuracy of BoneView across studies

Group (n studies)	Median sens (range)	Median spec (range)	Median % missed #s (range)	Median % over diagnosis (range)
All fracture analyses				
BoneView, any staff grouping, all fractures (7)[10 11 13 14 16 23 25]	0.83 (0.75, 0.91)	0.93 (0.65, 0.99)	17.3% (8.6, 25.0)	7.4% (0.6, 35.0)
BoneView, less experienced staff, all fractures (3)[11 13 14]	0.79 (0.74, 0.89)	0.97 (0.93, 0.98)	21.3% (11.2, 25.7)	3.3% (1.7, 7.4)
BoneView, mixed or unclear staff, all fractures (6)[10 13 14 16 23 25]	0.81 (0.75, 0.91)	0.90 (0.65, 0.97)	19.0%(8.6, 25.0)	9.8% (2.6, 35.0)
BoneView,senior or expert staff, all fractures (1)[11]	0.88 (NA)	0.99 (NA)	11.7% (NA)	0.64% (NA)
Subgroup fracture types				
BoneView, mixed or unclear staff,	0.89 (0.8, 0.96)	0.92 (0.88, 0.95)	13.3% (8.3, 20.5)	9.4% (7.8, 13.3)

Group (n studies)	Median sens (range)	Median spec (range)	Median % missed #s (range)	Median % over diagnosis (range)
hand/wrist (4)[13 16 23 25]				
BoneView senior or expert staff, hand/wrist (2)[9 11]	0.89 (0.88, 0.90)	0.92 (NA)	11.7% (NA)	7.4% (NA)
BoneView, mixed or unclear staff, foot/ankle (4)[13 16 23 25]	0.89 (0.71, 0.98)	0.93 (0.80, 0.96)	11.6% (1.8, 30.0)	10.4% (6.8, 20.8)
BoneView, senior or expert staff, foot/ankle (1)[11]	0.83 (NA)	NR	NR	NR
BoneView, mixed or unclear staff, hip (1)[16]	0.93 (NA)	0.99 (NA)	NR	NR
BoneView, less experienced staff, non-obvious fractures (1)[24]	0.56 (NA)	0.79 (NA)	43.8% (NA)	21.1% (NA)
BoneView, mixed or unclear staff, non-obvious fractures (1)[16]	0.83 (NA)	NR	16.7%	100%
BoneView, senior or expert staff, non-obvious fractures (1)[24]	0.81 (NA)	0.89 (NA)	18.8% (NA)	10.5% (NA)

Abbreviations: NA, not applicable; sens, sensitivity; spec, specificity

Results for RBFracture as reported across studies are shown in Table 13. The sensitivity of RBFracture was broadly comparable to that of BoneView, although there was a trend for specificity to be worse, with similar rates of false positives across reader groupings as were reported for unassisted readers. This may reflect a prioritisation of the technology threshold towards sensitivity; i.e. prioritising the avoidance of missed fractures over avoiding false positives. However, rates of missed fractures were still generally high. There were fewer studies that evaluated RBFracture, and there was substantial variability in outcomes between studies. This creates more uncertainty in the findings. Notably, one study in senior readers had poor sensitivity compared to other reader groups; although this study was conducted in a paediatric population, which may have influenced the findings (see Section 5.2.1.2).

Table 13: Diagnostic accuracy of RBFracture across studies

Group (n studies)	Median sens (range)	Median spec (range)	Median % missed #s (range)	Median % over diagnosis (range)
All fracture analyses				

Group (n studies)	Median sens (range)	Median spec (range)	Median % missed #s (range)	Median % over diagnosis (range)
RBFracture, any staff group, all fractures (3)[17 20 27]	[redacted]	[redacted]	[redacted]	[redacted]
RBFracture, less experienced staff, all (2)[17 27]	[redacted]	[redacted]	[redacted]	[redacted]
RBFracture, mixed, all (3)[17 20 27]	[redacted]	[redacted]	[redacted]	[redacted]
RBfracture, senior, all (1)[20]	[redacted]	[redacted]	[redacted]	[redacted]
Subgroup fracture types				
RBFracture, junior, hip (1)[7]	[redacted]	[redacted]	[redacted]	[redacted]
RBFracture, mixed, hip (2)[28]	[redacted]	[redacted]	[redacted]	[redacted]

Abbreviations: NA, not applicable; sens, sensitivity; spec, specificity

Results for Rayvolve are shown in Table 14. Two studies evaluated Rayvolve, both of which reported high sensitivity but poor specificity, particularly for hand/wrist and foot/ankle fractures. The EAG considered this was a feature of the technology algorithm, to prioritise missed fractures over false positives. Accordingly, specificity was comparable with unassisted diagnosis, while sensitivity was generally improved.

Table 14: Diagnostic accuracy of Rayvolve across studies

Group (n studies)	Median sens (range)	Median spec (range)	Median % missed #s (range)	Median % over diagnosis (range)
Rayvolve all staff all fractures (2)[21 23]	0.94 (0.93, 0.96)	0.83 (0.70, 0.85)	6.1% (4.4, 7.4)	16.9% (14.7, 29.6)
Rayvolve junior all fractures (1) [21]	0.94 (NA)	0.85 (NA)	6.1% (NA)	14.7% (NA)
Rayvolve mixed all fractures (2) [21 23]	0.94 (0.93, 0.96)	0.77 (0.70, 0.83)	5.9% (4.4, 7.4)	23.3% (16.9, 29.6)
Rayvolve mixed hand (1)[23]	0.98 (NA)	0.75 (NA)	2.1% (NA)	25.4%
Rayvolve mixed foot (1)[23]	0.91 (0.91, 0.92)	0.67 (0.63, 0.72)	8.0% (7.1, 8.9)	32.8% (27.9, 37.7)

Abbreviations: NA, not applicable; sens, sensitivity; spec, specificity

Results for TechCare Alert are shown in Table 15. Two studies evaluated TechCare Alert, with no crossover in the reader groupings, meaning that data was only available separately or in a group reporting results for any reader grouping. Both studies reported high sensitivity and specificity for TechCare Alert, with no rates of missed fractures and false positives. Results were improved compared to unassisted diagnosis. Results were similarly high for hand/wrist and foot/ankle fractures.

Table 15: Diagnostic accuracy of TechCare Alert across studies

Group (n studies)	Median sens (range)	Median spec (range)	Median % missed #s (range)	Median % over diagnosis (range)
TechCare Alert all staff all fractures (2)[22 23]*	0.93 (0.90, 0.95)	0.98 (0.93, 0.98)	7.1% (5.1, 9.8)	1.9% (1.9, 7.5)
TechCare Alert junior all fractures (1)[22]	0.95 (NA)	0.98 (NA)	5.1% (NA)	1.9% (NA)
TechCare Alert mixed all fractures (1)[23]	0.90 (NA)	0.93 (NA)	5.1% (NA)	1.9% (NA)
TechCare Alert senior all fractures (1)[22]	0.93 (NA)	0.98 (NA)	7.1%	1.9%
TechCare Alert mixed hand (1)[23]	0.94 (NA)	0.92 (NA)	6.2% (NA)	8.3% (NA)
TechCare Alert mixed foot (1)[23^]	0.88 (0.85, 0.90)	0.91 (0.90, 0.92)	11.9% (9.5, 14.3)	8.6% (7.9, 9.2)

*2 studies and 3 data points as 1 study reported 2 staff grades

5.2.1.2. Subgroup results (paediatric participants)

Median results across studies reporting diagnostic accuracy data in children and young people only are shown in Table 16. No evidence was available in a less experienced reader group only. Amongst mixed or unclear experience readers, the assistance of BoneView or RBFracture improved median sensitivity for detecting fractures, though made no clear difference to specificity. In real terms, this reduced the number of fractures that were missed but did not change the number of false positives. One study that reported data for highly experienced staff reported poor sensitivity both with and without assistance with RBFracture, which lacked face validity. In this group of readers, the assistance of RBFracture improved sensitivity but this

came with a cost to specificity; i.e. there was a meaningful reduction in missed fractures but a large increase in false positive diagnoses.

Table 16: Diagnostic accuracy of assisted and unassisted diagnosis in children and young people across studies

Group (n studies)	Tech	Staff	Fracture	Median sens (range)	Median spec (range)	Median % missed #s (range)	Median % over diagnosis (range)
No AI mixed all fractures (4)[16 17 20 25]	No AI	Mixed	All	0.78 (0.73, 1.0)	0.91 (0.77, 0.95)	22.6% (0.0, 26.7)	9.3% (5.6, 22.5)
BoneView mixed all fractures (2)[16 25]	BoneView	Mixed	All	0.91 (0.83, 1.0)	0.91 (0.30, 0.92)	8.7% (0, 17.3)	9.0% (8.0, 10.0)
RBFracture mixed all fractures (2)[17 20]	RBFracture	Mixed	All	[redacted]	[redacted]	[redacted]	[redacted]
No AI senior all fractures (1)[20]	No AI	Senior	All	[redacted]	[redacted]	[redacted]	[redacted]
RBFracture senior all fractures (1)[20]	RBFracture	Senior	All	[redacted]	[redacted]	[redacted]	[redacted]

5.3. Conclusions of the clinical, service and technological evidence

As an emerging technology, the evidence base for the clinical and service value of the technology for assisting with fracture diagnosis was expectedly limited. Nevertheless, within the context of an EVA, the EAG considered it notable that a total of 16 studies were identified and eligible for inclusion in the evidence review. Within the methods of the EVA, the EAG was appropriately broad in its inclusion criteria: including conference abstracts, non-peer reviewed reports, and manuscripts in preparation, all of which may not typically be considered within a NICE evaluation. The EAG also included a variety of study designs, including those with known limitations for determining reliable estimates for diagnostic accuracy. As the aims of an EVA include identifying an overview of the existing evidence base and key evidence gaps and research needs, the inclusion of this evidence was useful for these purposes. However, as stressed within the report, these studies are not robust for identifying reliable estimates of clinical and service outcomes and should be interpreted with caution.

The evidence base as a whole suggested that there is a need for further, high-quality research to evaluate the clinical and service outcomes associated with the technology (see Section 9). While several of the technologies had been evaluated in few studies available to date, there were nine and five included studies evaluating BoneView and RBFracture. For these technologies, the EAG considered that the time for a proof-of-concept of the technologies had passed, and that these technologies require evaluation within robust comparative study designs. These would include diagnostic randomised controlled trials and prospective, robustly sampled diagnostic studies, each set within the likely target settings and reader groups that would be expected to use the technology in clinical practice. To date, much of the evidence has focussed on the accuracy of the technology as a standalone tool (excluded from the evidence review as clinical standards would mean that this use would not be possible within the NHS) and as assistance to clinicians in radiology departments, including consultant-level staff, who would not be expected to require the use of the technology in clinical practice. In addition to the potential value of the technology for avoiding missed fractures, stakeholders to the assessment highlighted the potential value of the technology for service outcomes, such as time and resource savings, but this has not been a focus of any of the available studies to date. All this said, however, the EAG noted that the evaluation of the technology will be particularly complex in this field, as compared to evaluating diagnostic tools in other indications (see Section 9). In the same way as these complexities affected the ease of interpretation of the evidence base in this assessment, the EAG expected that some of these complexities would remain for any future NICE assessment, even with a more developed evidence base.

Given the limitations in the included studies, the EAG considered that the precise estimates of sensitivity and specificity reported in the studies were uncertain and may have limited generalisability to the outcomes that may be seen in clinical practice. However, across the evidence identified, there was a trend for the technologies to result in improvements in diagnostic sensitivity for identifying fractures relative to unassisted diagnosis, but minimal or no improvement in specificity for identifying where fractures were not present. This trend was seen across technologies and reader groups. Where accuracy of diagnosis unassisted by the technology was already high, such as in the identification of hip fractures and when used by senior readers with high accuracy, the additional value of the technology was evidently smaller. Generally speaking, the evidence showed that the use of the technology did not result in perfect sensitivity and specificity for any reader or any fracture type: while not unexpected, the implications of this were that the technology could not be relied upon to identify all fractures or

to never result in a false positive diagnosis. Within analyses of non-obvious fractures, where arguably the use of the technology could be greatest, the technology generally improved sensitivity but was still associated with a large number of false negatives that would require handling through additional service use (e.g. imaging, consultant review, precautionary tactics, patient recall). While studies mostly reported that use of the technology may reduce reading time for each X-ray by several seconds, advice from clinical experts was that this may not be preferable, as the technology should – using best practice – be an add-on step of the care pathway and thus result in overall increased time for diagnosis within services.

In conclusion, on the basis of the available evidence base, the EAG considered that there are early indications that the use of the technology to aid identification of fractures could have value for avoiding missed fractures. Further evidence is needed to evaluate whether this potential holds true for the specific ways in which the technology could be used within the NHS, including the target fracture types, readers, and the broader care pathway used within each of the target settings. There was some evidence that accuracy may vary across the technologies evaluated, however due to the lack of evidence for some technologies, and variation in study designs used to evaluate each technology, this was uncertain. There is also a need to determine the potential implications of the technology for broader service use outcomes, particularly to determine any potential trade-off between increased time to diagnosis with other potential time and resource savings. Further evidence is also needed to determine the clinical outcomes associated with any difference in diagnostic accuracy and service use.

6. ECONOMIC EVIDENCE SEARCHES AND SELECTION

6.1. Evidence search strategy and study selection

A single search was conducted to identify clinical, technological and economic evidence. Please see Section 4.1 for details of the evidence searches.

6.2. Included and excluded studies

A total of 1907 records were retrieved by the database searches plus 90 records identified by other sources. Of these, 654 records were excluded as duplicates, resulting in 1343 records to be screened by title and abstract. From this screening, 75 records were selected for full-text retrieval. In total, the review included 4 studies that informed health state costs and utilities: these are summarised in Table 17

A list of studies excluded along with the rationale for exclusion is provided in Appendix C. A PRISMA diagram of the search and screen process is provided in Appendix B.

Table 17: Key studies selected for the economic model

Study name, design and location	Intervention(s) and comparator	Participants and setting length of follow-up	Relevant outcomes and key findings	EAG comments
Low et al (2021) CEA Singapore https://doi.org/10.1016/j.jval.2021.06.005 Published article	**Intervention:** DECT supplementing SECT SECT alone **Comparator:** MRI	**Participants:** 70-year-old female (base case) **Setting:** ED **Follow-up:** Modelled time horizon: Lifetime	**Primary outcome:** • Utilities **Secondary outcomes:** • Sensitivity, Specificity • Costs	This study was used to inform TP,TN, FP and FN utilities for hip fracture in the economic model
Judge et al (2016) Service evaluation study with health economic analysis England (UK) https://www.ncbi.nlm.nih.gov/books/NBK385615/ Published article	**Intervention:** Changes to secondary prevention services at England hospitals between 2003 and 2012 (fracture liaison nurse (FLN) care and orthogeriatrician (OG) care) **Comparator:** Usual care (no OG or FLN)	**Participants:** 43 health professionals working in fracture prevention services in secondary care **Setting:** Acute hospitals in England **Follow-up (patient):** 2.6 years from index hip fracture Modelled time horizon in CEA: Lifetime	**Primary outcome:** • Costs **Secondary outcomes:** • Utilities	This study was used to inform TP,TN, FP and FN costs (UK) for hip fracture in the economic model

Study name, design and location	Intervention(s) and comparator	Participants and setting length of follow-up	Relevant outcomes and key findings	EAG comments
Nwankwo et al (2022) Trial based CEA UK https://doi.org/10.1302/2633-1462.36.BJO-2022-0036 Published article	**Intervention:** Removable brace **Comparator:** Cast	**Participants:** Patients presented to hospital with ankle fracture **Setting:** Acute hospital **Follow-up:** Modelled time horizon: 1 year	**Primary outcome:** • Utilities and costs	This study was used to inform TP,TN, FP and FN costs and utilities for ankle/foot fracture in the economic model
Rua et al (2020) Trial based CEA UK Published article	Intervention: Immediate MRI Comparator: No further imaging	Participants (n=132) were recruited from the ED at a hospital in central London. Setting: ED Follow-up: Modelled time horizon: 6 months	Primary outcome: • Utilities and costs	This study was used to inform TP,TN, FP and FN costs and utilities for Hand/wrist fracture in the economic model

7. EVIDENCE SUBMITTED BY COMPANIES

Three companies (Gleamer, Milvue and Radiobotics) submitted a RFI response to inform this assessment, no RFI was received from AZmed or Qure.ai. As part of the RFI responses, each company provided details about evaluations of their technologies that may be relevant for consideration within this assessment. These studies were screened using the same methods described in the protocol for the assessment, meaning that they were included where they met the inclusion criteria for the evidence reviews. This included two unpublished manuscripts, currently in preparation, submitted by Radiobotics. Studies that were excluded during screening are listed in Appendix C alongside the reasons for exclusion.

A number of studies described by companies in their RFI responses were identified by the EAG as plausibly relevant but there was insufficient information to determine their eligibility for the assessment. This included studies where:

- It was unclear whether the technology was evaluated as a standalone diagnostic tool or whether the technology was used in conjunction with clinician judgement.
- An incomplete citation was provided by the company, and it was not possible to identify the publication source.
- Results were not provided for any of the outcomes in the NICE scope or data were presented in an unusable format.

In all such cases, the EAG submitted a request for clarification from the companies who submitted RFIs (Gleamer, Milvue, and Radiobotics).

Similar uncertainties were identified for several studies identified by the EAG in its evidence review, in addition to uncertainties regarding the name of the technology evaluated. Requests for clarification were submitted to companies where they were listed as a study sponsor and/or where staff from the company were listed as a study sponsor (Gleamer, Milvue, AZmed, and Radiobotics).

Where no clarification response was received by the companies, any studies that did not clearly meet the review eligibility criteria, as outlined above, were excluded.

8. ECONOMIC EVALUATION

8.1. Quality appraisal of selected studies

Consistent with the methods for an EVA, no formal quality appraisal of included studies was undertaken.

8.2. Relevant economic models

Four papers relevant to the topic area were identified and used as a basis for this assessment, specifically providing information on the payoffs (cost and QALYs accrued) from the diagnostic outcomes of true and false positive and negative. An overview of these publications is provided in Table 17.

- Rua (2020)[31] conducted an assessment for the wrist, evaluating the cost-effectiveness of immediate magnetic resonance imaging (MRI) in managing patients with suspected scaphoid fractures. The clinical outcomes were sourced from the SMaRT trial, which recruited participants with negative initial radiograph findings at the emergency department of a central hospital in London.

- Nwankwo (2022)[32] calculated the incremental cost-effectiveness of a removable brace vs plaster cast in the management of adult patients with ankle fractures. The outcomes were derived from the Ankle Injury Rehabilitation (AIR) trial, a UK-based pragmatic multicentre randomized controlled trial (RCT). Eligible patients presented to the hospital with an ankle fracture, whether treated operatively or non-operatively, for which the clinician deemed a cast a reasonable management option.

- Low (2021)[33] analysed the utilities associated with hip fractures, comparing the costs and QALYs of different imaging strategies for diagnosing occult hip fractures. The study compared magnetic resonance imaging (MRI) with dual-energy computed tomography (DECT), single-energy computed tomography (SECT) supplemented with DECT, and SECT alone.

- Judge (2016)[34] evaluated the cost-effectiveness of three models of secondary fracture prevention for all patients with hip fractures admitted to NHS hospitals.

Further details on how the data were used are described from Sections 8.3.4 to 8.3.6.

8.3. Economic model

The EAG developed a *de novo* model to explore the potential cost-effectiveness of AI-assisted diagnosis of fracture compared with unassisted diagnosis of fracture in an urgent care setting from the perspective of the NHS and Personal Social Services (PSS).

Given the heterogeneous nature of the subsequent costs and consequences of different fracture locations, the EAG divided the analysis into three separate decision problems, focussing on diagnosis and treatment of (1) wrist and hand fractures, (2) ankle and foot, and (3) hip. These three fracture sites were chosen as being both where the EAG considered there to be opportunity for greatest benefit from AI-assisted diagnosis and where the costs and consequences of the fractures differed substantially, warranting separate modelling. The outputs of these were then weighted based on the case mix of a typical urgent care setting to estimate the overall change in costs and QALYs accrued within that setting attributable to AI-assisted diagnosis. Scenario analyses explored additional adjustment for use of AI in diagnosing fractures other than the three types explicitly modelled.

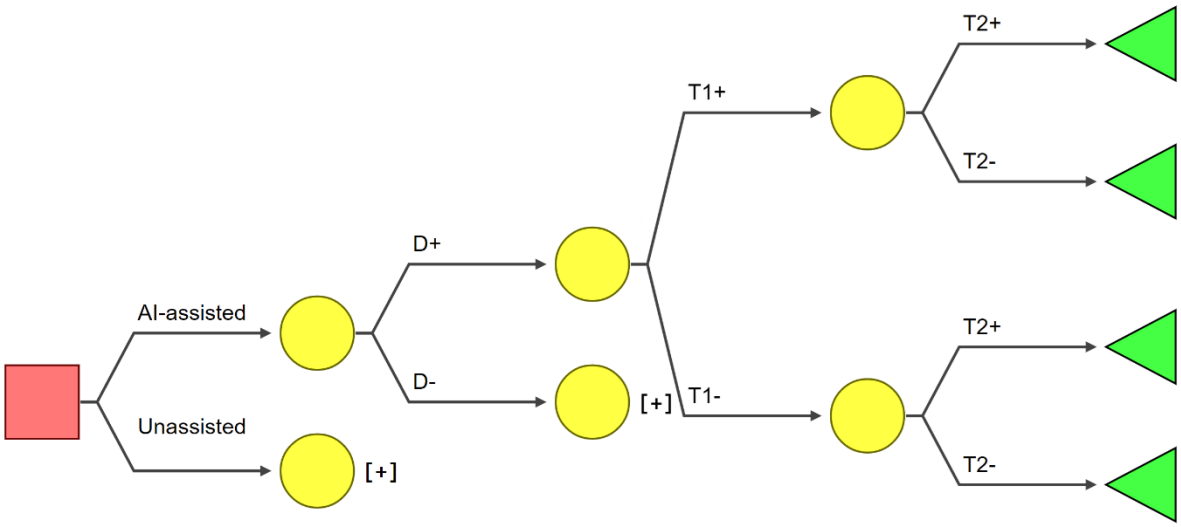
The model described below represents a very rapid overview of the likely costs and consequences associated with use of several commercial AI algorithms to assist in the diagnosis of fracture, and unassisted diagnosis, in an urgent care setting. The purpose of this analysis approach was to explore whether there was a *prima facie* plausible case for any of the technologies to represent value for money for the NHS / taxpayer within the context of a NICE EVA. The results should therefore not be used as a definitive estimate of the costs, effects and cost-effectiveness of the interventions: a more thorough analysis may lead to different conclusions.

8.3.1. Model structure

The model structure (Figure 1) comprised a decision tree incorporating the prevalence, sensitivity and specificity and cost per diagnosis of a strategy (i.e. AI-assisted or unassisted diagnosis). The model structure allowed up to two reviews of a radiograph, although the model could be simplified to just a single review (e.g. where input data either represent just a single view, or where accuracy data are presented for a double-reading process alone, without disaggregation by reader). Allowing for two reviews generated eight terminal nodes, two each for four possible states of the world: true positive, true negative, false positive and false negative. Payoffs in terms of costs and QALYs were attached to these, drawing on previous relevant studies in the literature. The prevalence in this case was defined as the proportion of

patients arriving in the urgent care setting with a suspected fracture who have a fracture. Payoffs attached to each terminal node are described in Sections 8.3.4 to 8.3.6. Details on how the three models were combined to estimate the overall impact on a 'typical' urgent care setting are in Section 8.3.7.

Figure 1: General structure of model



Square = decision node; Circle = chance node; Triangle = terminal node. D+ = prevalence (i.e. probability of a suspected fracture being a true fracture); D- = 1 - D+; T1+ = conditional probability of a positive result from first review of X-ray (in branch shown this is sensitivity); T2+ = conditional probability of a positive result from second review of X-ray. Costs and QALYs are assigned to terminal nodes. Image drawn with SilverDecisions (silverdecisions.pl)

8.3.2. Model assumptions: EAG base case and scenario analyses

The EAG base case assumed that patients present in an emergency department with a suspected fracture, which can be one of three or four types: ankle or foot, wrist or hand, hip or 'other'. The base case assumed that AI-assisted diagnosis was limited to ankle/foot, wrist/hand and hip fractures alone and so only considered these fracture types. A scenario analysis (scenario 8) assumed that the software was used to read radiographs for all suspected fractures (including 'other').

As the cost of the emergency department attendance and X-ray (including the grade of staff reading the x-ray) was common to all comparators in this analysis, these costs were excluded. The difference in cost between different technologies was therefore limited to the cost per scan.

This also meant that the cost of reading the scans were directly transferrable to other urgent care settings (i.e. UTC and MIU). Sensitivity and specificity of the diagnosis was assumed to depend on the grade of staff reading the scan, rather than the setting where it took place.

The decision problem addressed in the EAG base case was to explore the cost-effectiveness of AI-assisted diagnosis using four technologies (BoneView, Rayvolve, RBfracture and TechCare Alert) and unassisted diagnosis of fracture in the urgent care setting from the perspective of the NHS & PSS.

8.3.3. Sensitivity and specificity of diagnosis

Two of the studies identified in the literature review were selected as sources for prevalence and sensitivity and specificity of each diagnostic strategy. Whilst all studies had strengths and limitations, and a number of potentially relevant studies for the decision model were identified in Section 4.2.1, Bousson et al. (2023)[23] provided directly comparative data between BoneView, Rayvolve and TechCare Alert. Furthermore, data were disaggregated by foot, ankle and hand (but not hip). For the base case, the EAG estimated a mean sensitivity and specificity for foot and ankle, assumed hand applied equally to wrist, and assumed the sensitivity and specificity of hip fracture diagnosis was equal to that for 'all fractures'. Bachmann et al. (2024) compared RBFracture assisted diagnosis to unassisted diagnosis in a wide range of fracture types, reporting results by 'mixed' staff types, ED trainees and trauma care nurses. This was used as the source study for RBFracture and unassisted diagnosis, with diagnosis by "A&E trainees" considered the closest match to that in Bousson ("radiology residents" with four years' experience). Due to concerns about quality limitations in the studies, the EAG did not seek to adjust estimates for differences in the characteristics of these studies, and a full network meta-analysis of source studies is recommended in future work. Instead, the EAG explored optimistic and pessimistic scenarios in the analysis (see Section 8.3.10). As data were not disaggregated by fracture type, the base case assumed the same sensitivity and specificity for all fracture types for RBFracture and unassisted diagnosis.

Source studies estimating the sensitivity and specificity of diagnosis assumed a single read of a scan. To ensure as fair a comparison between technologies, the EAG base case assumed only a single read. A scenario analysis including a second read of all scans was included as the EAG considered that this was a more realistic approach for clinical practice. Base case input data (including parametric distributions for the probabilistic analysis) are summarised in Table 18.

Table 18: EAG Base case Prevalence, Sensitivity and Specificity

Fracture site	Prevalence	Technology	Sensitivity	Specificity	Source/notes
Ankle/Foot	0.241 β(50.38, 158.62)				Bousson 2023, crude mean of ankle and foot
		BoneView	0.94 β(196.46, 12.54)	0.86 β(179.22, 29.78)	Ibid
		Rayvolve	0.91 β(191.13, 17.87)	0.67 β(140.66, 68.34)	Ibid
		TechCare alert	0.88 β(183.19, 25.81)	0.91 β(190.61, 18.39)	Ibid
		RB Fracture	0.83 β(277.22, 56.78)	0.90 β(300.60, 33.40)	Bachmann 2024 (A&E trainee)
		Unassisted	0.74 β(247.16, 86.84)	0.87 β(290.58, 43.42)	Bachmann 2024 (A&E trainee)
Wrist/Hand	0.309 β(97.00 217.00)				Bousson 2023
		BoneView	0.915 β(287.31, 26.69)	0.921 β(289.19, 24.81)	Ibid
		Rayvolve	0.98 β(307.09, 6.91)	0.75 β(234.24, 79.76)	Ibid
		TechCare alert	0.94 β(293.90, 20.10)	0.92 β (287.94, 26.06)	Ibid
		RB Fracture	0.83 β(277.22, 56.78)	0.90 β (300.60, 33.40)	Bachmann 2024 (A&E trainee)
		Unassisted	0.74 β(247.16, 86.84	0.87 β (290.58, 43.42)	Bachmann 2024 (A&E trainee)
Hip	0.217 β (163.00, 587.00)				Bousson 2023
		BoneView	0.91 β (684.75, 65.25)	0.91 β (678.75, 71.25)	Ibid
		Rayvolve	0.93 β (694.50, 55.50)	0.70 β (528.00, 222.00)	Ibid
		TechCare alert	0.90 β (676.50, 73.50)	0.93 β (693.75, 56.25)	Ibid
		RB Fracture	0.83 β (277.22, 56.78)	0.90 β (300.60, 33.40)	Bachmann 2024 (A&E trainee)
		Unassisted	0.74 β (247.16, 86.84)	0.87 β (290.58, 43.42)	Bachmann 2024 (A&E trainee)

Notes: Table reports means and probability distribution and parameters used in analysis. Payoffs for each fracture type are described in the following sections (8.3.4 to 8.3.6).

8.3.4. Foot and Ankle

Based on the WHO Global Burden of Disease 2019 report,[35-37] Chen et al.[38] estimated that in the UK in 2019, the age standardised incidence of foot fracture was 202.64 per 100,000 (95%

uncertainty interval (UI) 142.23 to 278.04). Globally, foot fracture incidence exhibits a bimodal distribution by age, with peaks amongst the very elderly (80+) in both men and women and at around age 20-24 in males and 10-14 in females. Overall incidence is higher in men than in women, although is similar amongst the very elderly.

Two previous economic studies were identified as potential sources for data for this analysis. Nwankwo et al. (2023)[32] reported the results of a within-trial cost utility analysis comparing removable brace with cast in patients with ankle fractures aged 18+ over 12 months, whilst Baji et al. (2023)[39] reported the results of a within-trial economic analysis over 12 weeks. Health state utilities measured in Nwankwo et al. at 6 time points showed continuing improvement in both arms, suggesting healing continued over this time horizon. In the absence of data to the contrary, the foot and ankle fracture model assumed a time horizon of 12 months (i.e. implicitly assuming that there will be no difference in cost and outcomes between patients with and without fractures after this point). The Baji 2023 study was not considered further as a source for extracting payoffs due to its short time horizon.

In this analysis, patients were assumed to present to urgent care with a suspected fracture following a trauma. The two states of the world were for the ankle or foot to be broken, or for it to have sustained soft tissue injury alone. In the case of a soft tissue injury, the patient was assumed to remain in pain for a short while (e.g. 2 weeks), before returning to normal health.

The time horizon of the source study was one year. Payoffs in terms of costs and QALYs for the four possible outcomes are described below.

True positives

Nwankwo (2023)[32] reported health state utilities for patients experiencing ankle fractures over one year using the EQ-5D-5L instrument, yielding 0.723 QALYs for patients with brace and 0.720 for those treated with cast. The EAG assumed that 50% of patients would be treated with a brace and 50% with a cast. As allocation amongst the trial was almost exactly 50/50, the weighted average QALYs accrued was 0.722 (SE 0.06), which was assumed to be the QALYs associated with a true positive diagnosis of a foot or ankle fracture (Table 19).

Table 19 Calculation of QALYs associated with true positive detection of ankle / foot fracture

	Brace	Cast	Total
n	335	334	669
QALYs mean	0.723	0.72	0.722
QALYs SD	0.153	0.149	0.151
SE	0.008	0.008	0.006

Source: Nwankwo et al. 2023

NHS and personal social services costs over 12 months were estimated at £995.54 (SE £130.68) in the brace arm and £717.47 (SE £47.70) in the cast arm (2019/20 prices). This included ED visits, in-patient days, medication, community health services, outpatient visits, aids and adaptations to the home and personal social services (as described in Nwankwo 2023 supplementary material[32]). Following confirmation with the author, the ED visits measured in the study did not include the index attendance (i.e. initial presentation at ED with a suspected fracture). As per the assumptions for calculation of QALYs, the EAG assumed a 50/50 split between the two approaches and adjusted the price year with the NHSCII pay and prices index yielding a mean cost of £1,837.23 (SE £71.03) in 2022/23 prices (Table 20).

Table 20 Calculation of costs associated with true positive detection of ankle / foot fracture

12 m NHS+PSS cost	Brace	Cast	Mean	2022/23 prices
N	335	334	669	
Mean	£995.54	£717.47	£856.71	
SD	£2,391.74	£871.77	£1,632.89	£1,837.23
SE	£130.67	£47.70	£63.13	£71.03

Source: Nwankwo et al. 2023

True Negatives

A patient who is correctly diagnosed as not having fractured their ankle or foot was assumed to incur the cost of the index emergency department consultation and then be discharged (the index consultation is excluded from the analysis as is common to all arms).

To estimate QALYs, patients were assumed to endure the health state equivalent to a fracture for two weeks (representing pain and soft tissue bruising from the injury causing them to seek assistance), before reverting to the 'healed' health state utility for the remainder of the year. This led to a mean QALYs accrued over 1 year of 0.796 (Table 21). Utilities and durations were entered in the model with QALYs calculated dynamically from sampled values of both. Due to

lack of data, correlation between utility at the timepoints was assumed zero (i.e. independent). This will overestimate the distribution of sampled values of QALYs (i.e. overestimate uncertainty in QALYs). The utility associated with ankle/foot fracture (0.225) was considered to lack face validity for a soft tissue injury. Therefore, scenario analysis explored alternative assumptions.

Table 21 Calculation of QALYs associated with true negative ankle / foot fracture

	Brace		**Cast**		**Total**		
N	335		334		669		
	mean	SD	mean	SD	mean	SD	SE
Utility 1 (SD)	0.212	0.310	0.238	0.311	0.225	0.310	0.012
Duration 1					2		
Utility 2 (SD)	0.812	0.192	0.825	0.171	0.818	0.182	0.007
Duration 2					50		
QALYs					0.796		

Source: adapted from Nwankwo et al. 2023

False Positives

A false positive was assumed to accrue the same cost as a true positive patient and to accrue the same QALYs as a true negative.

False Negatives

Patients with a fracture who were wrongly diagnosed as not having a fracture were assumed to re-present in the ED after two weeks, whereupon additional investigations would be conducted, and a correct diagnosis made. Clinical advice to the EAG was that a patient would either present to their GP or represent in the ED. A plausible timeframe was considered to be within a month. They therefore incur the same costs as true positives (£1837.23, Table 20), plus a cost for additional examinations (£149.04, see Section **Error! Reference source not found.**), yielding a total of £1,986.27. To calculate QALYs, the patient was assumed to remain in the base health state for two weeks before following the utility trajectory of a true positive patient, trimmed to 52 weeks (Table 22).

Table 22: Calculation of QALYs associated with false negative ankle / foot fracture

	Brace		**Cast**		**Total**		
N	335		334		669		
	Mean	SD	mean	SD	mean	SD	SE
Utility 0	0.212	0.310	0.238	0.311	0.225	0.310	0.012
Utility 1	0.212	0.310	0.238	0.311	0.225	0.310	0.012
Timepoint 1					2		
Utility 2	0.534	0.258	0.497	0.272	0.516	0.265	0.010
Timepoint 2					8		
Utility 3	0.660	0.180	0.647	0.192	0.654	0.186	0.007
Timepoint 3					12		
Utility 4	0.73	0.177	0.702	0.198	0.716	0.187	0.007
Timepoint 4					18		
Utility 5	0.778	0.176	0.767	0.193	0.773	0.184	0.007
Timepoint 5					26		
Utility 6	0.809	0.192	0.821	0.171	0.815	0.182	0.007
Timepoint 6					52		
QALYs					0.697		

Source: adapted from Nwankwo et al. 2023

Summary of key assumptions

- Ankle fractures would be healed within 12 months.
- TN incur the cost of an ED attendance
- TN accrue QALYs based on 2 week's baseline utility from Nwankwo et al. followed by the remainder of the year at the 'healed' utility from Nwankwo et al.
- FP incur the same cost as TP, same QALYs as TN (2 weeks of pain then back to normal)

8.3.5. Hand and Wrist

Six previous economic evaluations[17 31 40-43] of interventions for hand and/or wrist fractures were identified, of which two were UK-based and therefore of most relevance to this study question (Rua et al. 2019[43] and Rua et al. 2020[31]). Both studies reported the results of the same randomised controlled trial of a second line diagnostic of immediate MRI vs no MRI in patients

with a suspected scaphoid fracture who had a negative X-ray (i.e. the sum of true and false negatives from the initial X-ray). The full cost-utility analysis (Rua et al. 2020) was used as the primary source for this analysis.

As Rua et al. 2020 compared immediate MRI with no MRI (control), the results were presented in an aggregate manner, without distinguishing between False Negatives and True Negatives. Consequently, the EAG drew on the control arm data to populate the costs and consequences of all four terminal nodes.

Utilities were measured at four timepoints to a time horizon of six months. Resource use comprised primary and secondary care contacts including fracture clinic appointments, subsequent ED visits, additional diagnostics, surgery, physiotherapy, splits and plaster casts.

The time horizon for the source study was six months. Payoffs in terms of costs and QALYs for the four possible outcomes are described below.

True Positives

The EAG assumed utilities for True Positive hand/wrist fractures were one standard deviation lower for three months, compared to those reported in Rua et al. 2020. This assumption was required as the aggregated results encompassed patients both with and without fractures (Table 23). Standard errors were assumed as per those reported in Rua et al. 2020.

Table 23 Health state utilities, Hand/Wrist True Positive

	Utilities*	SD*	SE**	EAG Base case
Baseline	0.786	0.158	0.020	0.628
1 month	0.747	0.238	0.030	0.509
3 months	0.843	0.227	0.028	0.843
6 months	0.843	0.211	0.026	0.843
QALYs for 6 months				**0.3463**

Source: * Rua et al 2020; ** EAG calculation

Abbreviations: QALY, quality-adjusted life-year; SD, standard deviation, SE, standard error

Table 24 Costs for the True positive population

	Control group	2022/23 prices
N	65	
Mean	£844.75	£986.34
SE	£42.29	£49.37

Abbreviations: N, sample; SE, standard error

True negatives

The utilities for True Negatives were assumed to be equal to those reported in Rua et al. 2020, as 89.6% of patients in the control arm confirmed the absence of fractures. The EAG noted that this resulted in a moderate underestimate of QALYs accrued due inclusion of 10.4% of fractures (Table 25).

Table 25 Utilities for True Negative population

Health state	Utility (mean)	SD	SE
Baseline	0.786	0.158	0.020
1 month	0.747	0.238	0.030
3 months	0.843	0.227	0.028
6 months	0.843	0.211	0.026
QALYs for 6 months	**0.393**		

Abbreviations: QALY, quality-adjusted life-year; SD, standard deviation, SE, standard error

Similarly to the utilities, the costs derived from the publication are assumed to align with a True Negative population.

Table 26 Costs for the True negative population

	Control group	2022/23 prices
N	65.00	
Mean	£559.79	£653.61
SE	£46.95	£54.82

Abbreviations: N, sample; SD, standard deviation; SE, standard error

False positives

As for true negatives, the utilities for false positives were assumed to be consistent with those reported in the publication, given 89.6% of patients in the control arm confirmed the absence of fractures. This may overestimate utility gains as it may be the case that treating false positives as positives may result in disutilities.

Table 27 Utilities for the False positive population

Health state	Utility (mean)	SD	SE
Baseline	0.786	0.158	0.020
1 month	0.747	0.238	0.030

Health state	Utility (mean)	SD	SE
3 months	0.843	0.227	0.028
6 months	0.843	0.211	0.026
QALYs for 6 months	**0.393**		

Abbreviations: QALY, quality-adjusted life-year; SD, standard deviation, SE, standard error

The costs associated with True Negatives and False Positives differ, as the latter group is treated similarly to those with fractures until a subsequent diagnostic test confirms otherwise. Therefore, the base case resource utilisation for this population is assumed to be 10% higher than that of the publication.

Table 28 Costs for the False positive population

	Control group	2022/23 prices
N	65.00	
Mean	£615.76	£718.97
SE	£51.65	£60.30

Abbreviations: N, sample size; SE, standard error

False Negatives

Regarding QoL, this population closely resembled that of true positives. However, it was assumed that patients in this group would be mistreated as negatives for the first two weeks following their presentation to urgent care, and therefore would not accrue QoL gains during this period. Furthermore, as observed in the true positive population, it was assumed that the utilities were lower by one standard deviation for up to three months compared to those reported in the publication. Finally, it was assumed that the QoL for patients after three and a half months of treatment aligned with the values reported in the publication.

Table 29 Utilities for the False negative population

Health state	Utility (mean)	SD	SE
Baseline	0.628	0.158	0.020
2 weeks	0.628	0.158	0.020
6 weeks	0.509	0.238	0.030
14 weeks	0.843	0.227	0.028
26 weeks	0.843	0.211	0.026
QALYs for 6 months	**0.329**		

Abbreviations: QALY, quality-adjusted life-year; SD, standard deviation, SE, standard error

Differences in resource use from the true positive population include an additional 5% of patients likely to undergo wrist surgery, as suggested by Rua (2020), and the inclusion of an extra visit to the emergency department.

Table 30 Costs for the False Negative population

	Control group	2022/23 prices
N	65.00	
Mean	£904.50	£1,056.10
SE	£53.43	£62.39

Abbreviations: N, sample size; SE, standard error

Summary of key assumptions

- All health states:
    - QoL after 3 months aligned with that of the publication as patients can be considered “cured”
    - 6 months is a long enough time horizon to capture all QoL and costs differences
    - The utilities and resource use from Rua et al. 2020 are representative of those of True negatives in the UK
- True positives
    - Utilities for the first 3 months are equal to one standard deviation below those reported in Rua et al 2020.
    - Resource use data
- False positives
    - The utilities of Rua et al. 2020 are representative of this population
    - Resource use associated to this health state is 10% higher than that of True negatives
- False negatives
    - Patients return to urgent care two weeks after the initial presentation following which their fracture is correctly diagnosed
    - Following from the above, the utility from Rua et al 2020 at baseline represent the utility for those 2 weeks, therefore not assuming disutilities in that period.

8.3.6. Hip

Six economic evaluations for hip fractures were identified. One study was UK based (Judge et al 2016[34]) and provided the lifetime costs of hip surgeries for usual care (UC), fracture liaison nurse (FLN) care and orthogeriatrician (OG) care models for delivery of secondary fracture prevention after index hip fracture. Other studies were conducted across different geographies (EU/Singapore/America) of which Low et al 2021[33], a Singapore based decision model comparing different diagnostic strategies for occult hip fractures provided utilities for hip fracture by age group and different time periods following the fracture (i.e. immediate, during first year and after first year following fracture) was used, as the study design was aligned with the decision problem scope and included a detailed model inputs table (along with the associated uncertainty parameters).

The time horizon for the source study was lifetime. Payoffs in terms of costs and utilities for the four possible diagnostic outcomes are described below.

True positives

Low et al 2021 reported utilities for hip fracture for 65–74-year-old and also at different time points following the index hip fracture (immediately following fracture, during first year following index hip fracture and beyond first year). Table 31 shows the mean utilities and uncertainty distributions inserted into the model based on Low et al.

Table 31. Hip fracture utilities as per Low et al 2021

	Mean	Distribution	param 1	param 2	Source
Utility immediate post fracture (delay/no surgery)	0.42	Beta	0.88	1.22	Abimanyi-Ochom[44]
Duration (weeks)	2				
Utility during the first year after hip fracture (index or first surgery)	0.59	Beta	2.65	1.8	Keating et al;[45] Jonsson et al[46]
Duration (weeks)	50				
Utility after the first year after hip fracture (secondary surgeries)	0.69	Beta	2.36	1.06	Keating et al;[45] Jonsson et al[46]
Duration (weeks)	884*				

Note: * Assuming lifetime horizon until 83 years of age (female life expectancy in line with Low et al 2021)

As the Judge et al. study was UK-based, the EAG considered it to be the preferred source for resource use and costs. This study comprised a lifetime Markov model with 1-year cycles that simulated the natural history of hip fractures, including progression, major non-hip fractures, and discharge to home or a care facility. The study further provided the intervention, hospital,

primary care and care home costs by male and female cohorts and by different care models namely usual care, FLN and OG care.

Table 32 below provides the mean discounted costs which is an average across all models of care with the assumption that patients were equally distributed across those models of care (namely UC, FLN and OG). This approach was taken as the population concerned was over 60 years typically having fragility related fractures, where FLN and OG care models could also be useful in addition to usual care. Standard error was not available; therefore, it was calculated from 95% CIs (upper and lower limits) provided (i.e. SE = upper limit – lower limit/3.92).

Table 32. Mean discounted costs across different models (usual care, FLN and OG) of secondary prevention care following hip fracture (2022/23 prices)

	Mean	SE (calculated using 95% CI)
Total Male costs	£40,628	£749
Total female costs	£52,050	£603
Average costs	£57,471	£834

Abbreviations: UC, usual care; FLN, fracture liaison nurse; OG, orthogeriatrician; LCI, lower confidence interval, UCI, upper confidence interval

True negatives

A patient who was correctly diagnosed as not having fractured their hip was assumed to incur the cost of an emergency room consultation and then be discharged. They were assumed to be experiencing health state equivalent to a post-fracture immediately for two weeks (to account for the pain and short-term impact of injury), before reverting to the baseline health utility (0.79 as shown in table below). Utilities and durations were used in the model to calculate QALYs dynamically from sampled values of both. Due to lack of data, the utility values at different time points were assumed to be independent.

Table 33. Utilities associated with true negatives for hip fractures

	Total			
N (observations)	187			
	Mean	**Distribution**	**Param 1**	**Param 2**
Immediate utility post fracture	0.42	Beta	0.88	1.22
Duration (weeks)	2			
Utility 65-74 years	0.79	Beta	2.84	0.76
Duration (weeks)	934*			

* Assuming lifetime horizon until 83 years of age (female life expectancy in line with Low et al 2021)

False positives

As per Low et al 2021, people with false positive results would be treated as having a hip fracture and would undergo surgery similar to true positives. Therefore, their costs would be the same as those of true positives. However, as they do not actually have the fracture, their utilities would be assumed to be the same as true negatives. The EAG noted that false positives were less likely to undergo surgery as the incorrect diagnosis was likely to be spotted before this point. The cost estimate for this analysis may therefore be an overestimate.

False negatives

Low et al. mentioned that people with false-negative results would be discharged only to return to ED within a month for hip surgery, following which they would have the same pathway as true positives.

In terms of costs, the same costs as true positives were incurred, plus costs for additional ED attendance and further imaging (CT was assumed), as given in Table 35.

Table 34. Utilities for False negative hip fractures

	Total			
N (observations)	187			
	Mean	**Distribution**	**Param 1**	**Param 2**
Immediate utility post fracture (applies to delay as well)	0.42	Beta	0.88	1.22
Duration (weeks)	4			
Utility first year following surgery	0.59	Beta	2.84	0.76
Duration (weeks)	48			
Utility beyond first year following surgery	0.69	Beta	2.84	0.76
Duration (weeks)	884*			

* Assuming lifetime horizon until 83 years of age (female life expectancy in line with Low et al 2021)

Table 35. Costs for false negative hip fractures

	Mean	**SE**	**Distribution**	**Param 1**	**Param 2**
Average lifetime costs (discounted) following hip fracture (based on three models of care)	£57,961	£918	Gamma	£3989	£15

Summary of key assumptions

- Long term costs and consequences for hip fractures were assumed for lifetime horizon (one index surgery followed by a second surgery assumed) as per Low et al 2021.

- False negatives were assumed to incur the same costs as true positives with additional costs of an ED attendance and any additional investigations (1 CT assumed). Time for return after discharge was assumed to be 1 month or 4 weeks for false negatives based on Low et al 2021.
- True negatives incurred the cost of an ED attendance and accrue QALYs based on immediate post-fracture utility applied for initial 2 weeks, following which they were assumed to return to baseline utility for 65-74 years.
- False positives incurred the same cost as true positives as per Low et al 2021 and same QALYs as true negatives.

8.3.7. Overall impact of AI-assisted diagnosis in an urgent care setting

Clinical advice to the EAG was that over 2022-23, there were approximately 25.3 million ED attendances in England and that fractures typically account for 2-4% of all visits, equating to between 506,000 and 1,012,000 attendances. Clinical advice suggested that around 12.5% of all fractures are ankle, 7.5% wrist and 12.5% are hip. Data specifically including ankle and foot, and wrist and hand were not available. However, the EAG assumed that these proportions represented the base case distribution to estimate the overall impact of diagnosis in an urgent care setting (Table 36). Scenario analysis included the use of the technology for all fractures.

These figures were multiplied by the number of patients per year expected in a 'typical' ED with 350-400 attendances per day (total attendances, not just fracture). This equated to 136,875 attendances per annum, of which 4,106 would be for fracture. Including just the base case fracture types, this equated to 1,334 scans per year.

The EAG's base case assumed a weighted average cost and outcomes by fracture type. This was scaled up to the number of attendances per year to estimate the overall difference in cost and QALYs accrued to patients attending the 'typical' ED. The three fracture subtypes had different time horizons. By merging the three together the assumption was that costs and QALYs accrued after 6 months for wrist/hand and 12 months for ankle/foot were identical across all arms, in other words complete healing has taken place by this time. If any differences remained after these time horizons, then the analysis may underestimate the cost-effectiveness of AI-assisted diagnosis (i.e. overestimate the incremental cost-effectiveness, underestimate the incremental net health benefit).

Table 36: Base case distribution of fracture types

Fracture type	Base case proportions	Number of attendances
Ankle/Foot	38.5%	513
Wrist/Hand	23%	308
Hip	38.5%	513
Total		1334

8.3.8. Cost of diagnosis & additional cost inputs

Cost of scans

Cost per scan was extracted from company RFIs where reported. Some companies operated volume-based pricing. The EAG's base case used a cost per scan based on 1,334 scans per annum. Base case prices are reported in **Error! Reference source not found.** and minimum and maximum prices are explored in Scenario analyses 3 and 4. Where no pricing data were supplied, the EAG inserted a notional cost per scan. The maximum economically justified price per scan for each software compared with unassisted diagnosis was also calculated.

Table 37 Cost per Scan (based on 1334 scans per annum)

Software	Cost per Scan	Notes
BoneView	£1.00	Notional cost
Rayvolve	£1.00	Notional cost
RBfracture	[redacted]	[redacted]
TechCare Alert	[redacted]	[redacted]
Unassisted	£0.00	By definition

Source: Company RFIs

Cost of A&E attendance

As the index presentation at A&E, and the X-ray, was common to all arms, the costs of these were excluded from the analysis. However, some cases (e.g. false negatives) were assumed to require an additional presentation after a period of time. This was costed on a mean cost of £149.04 (NHS Reference Costs 2022-23, code VB11Z: Emergency Medicine, no investigation, no significant treatment), and varied by +/-10% in a uniform distribution in the probabilistic analysis.

8.3.9. Approach to analysis

As there were multiple comparators, the EAG reports the incremental net health benefit of each technology compared with unassisted diagnosis at willingness to pay thresholds of £20,000 and £30,000 per QALY gained. This facilitates a fully incremental analysis as the option with the highest net health benefit is the option yielding an ICER at or below the threshold after taking into account dominated and extended dominated options; incremental comparisons between each technology can also be made by comparing the INHB directly. However, given the level of uncertainty and the rapid and approximate nature of the modelling for this assessment, the EAG cautions against comparing each AI software against each other, instead considering a 'class effect' for the software as a whole.

The EAG reports the results for each fracture location individually, followed by an overall weighted average for all three fracture subtypes based on case mix. This was multiplied up to show the expected impact on cost to the NHS and QALYs accrued to patients vs. unassisted for each software. Finally, the maximum economically justified price per scan was calculated as the cost per scan associated with an ICER of £20,000 per QALY and £30,000 per QALY.

8.3.10. Scenario analyses

The EAG planned to conduct scenario analyses on the use of the technology in different settings (e.g. ED vs UTC). However, the EAG considered that the grade of staff reading the radiograph was a more important determinant of the diagnostic accuracy than the setting, and any differences in resource use across the settings may largely be equal between all comparators. The EAG therefore conducted an optimistic and pessimistic scenario to represent different settings as described in Scenarios 1 & 2 below. Other scenarios are described below. An additional scenario is in Appendix D.

Scenarios 1 & 2: optimistic and pessimistic diagnostic accuracy

The EAG base case for this EVA used a naïve, unadjusted comparison of arms from two studies to inform the sensitivity and specificity of each technology. The EAG therefore explored an optimistic and pessimistic scenario, based on a review of all source studies (see Table 11 to Table 15). The optimistic scenario assumed the lowest sensitivity and specificity for unassisted diagnosis and highest for each technology, and the pessimistic the reverse (Table 38). Where a source suggested a lower or upper bound that was inside the EAG base case, the EAG base case was used as the estimate at the bound.

Table 38 Scenario analyses 1 & 2

Parameter	Base case	Scenario 1 (optimistic)	Scenario 2 (pessimistic)
Ankle/foot BoneView sensitivity	0.94	0.98	0.71
Ankle/foot BoneView specificity	0.86	0.96	0.80
Ankle/foot Rayvolve sensitivity	0.91	0.92	0.91
Ankle/foot Rayvolve specificity	0.67	0.72	0.63
Ankle/foot TechCare alert sensitivity	0.88	0.92	0.88
Ankle/foot TechCare alert specificity	0.91	0.91	0.85
Ankle/foot RB Fracture sensitivity	0.83	[redacted]	[redacted]
Ankle/foot RB Fracture specificity	0.90	[redacted]	[redacted]
Ankle/foot Unassisted sensitivity	0.74	0.54	0.83
Ankle/foot Unassisted specificity	0.87	0.86	0.88
Wrist/hand BoneView sensitivity	0.915	0.96	0.80
Wrist/hand BoneView specificity	0.921	0.95	0.88
Wrist/hand Rayvolve sensitivity	0.98	0.98	0.93
Wrist/hand Rayvolve specificity	0.75	0.85	0.7
Wrist/hand TechCare alert sensitivity	0.94	0.95	0.90
Wrist/hand TechCare alert specificity	0.92	0.98	0.92
Wrist/hand RB Fracture sensitivity	0.83	[redacted]	[redacted]
Wrist/hand RB Fracture specificity	0.90	[redacted]	[redacted]
Wrist/hand Unassisted sensitivity	0.74	0.60	0.78
Wrist/hand Unassisted specificity	0.87	0.60	0.78
Hip BoneView sensitivity	0.91	0.93	0.89
Hip BoneView specificity	0.91	0.99	0.83
Hip Rayvolve sensitivity	0.93	0.96	0.93
Hip Rayvolve specificity	0.70	0.85	0.70
Hip TechCare alert sensitivity	0.90	0.98	0.90
Hip TechCare alert specificity	0.93	0.95	0.93
Hip RB Fracture sensitivity	0.83	[redacted]	[redacted]
Hip RB Fracture specificity	0.90	[redacted]	[redacted]
Hip Unassisted sensitivity	0.74	0.74	[redacted]
Hip Unassisted specificity	0.87	[redacted]	[redacted]

Source: extracted from Table 11 to Table 16

Scenarios 3 & 4: cost per scan

Several companies price their software on the basis of annual volume. These scenarios explored a low volume (i.e. highest cost per scan) and high volume (lowest cost per scan) analysis for those technologies pricing based on volume ([redacted], Table 39).

Table 39: Scenarios 3&4 - cost per scan

Intervention	Base Case	Low volume/high cost	High volume/low cost
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]

Scenarios 5 & 6: Reduced time to interpret radiograph

Data on reduced reading time for X-rays are reported in Section 5.1.4 These studies generally reported time savings of between 7 and 13 seconds per scan assisted by the technology. For the purpose of this scenario analysis, the EAG assumed a notional 10 second reduction per X-ray and subtracted the cost of this from the cost of the AI-assisted strategies under two scenarios: scenario 5 assumed the radiograph was read by a junior / trainee radiologist and scenario 6 by a consultant level radiologist (Table 40). The EAG noted advice from clinical experts that best practice use of the technology may lead to increases in reading time, though considered the lack of evidence for this, the EAG did not explore this scenario in the analysis.

Table 40: Scenarios 5&6 - reduced time to read X-ray

Scenario	Staff grade	Unit cost (per hour)	Time reduction per radiograph	Cost reduction per radiograph	Source / Notes
Scenario 5	Registrar	£50	10 seconds	13.9p	PSSRU 2023, P95
Scenario 6	Consultant: medical	£109	10 seconds	30.3p	ibid

Scenario 7: Health state utility for negative ankle & foot fractures.

The base case analysis for an ankle or foot fracture assumed that a patient without a fracture experienced an equal health state utility to a fracture, but only for two weeks before resolving (i.e. assuming a soft tissue injury which healed in two weeks). However, the EAG considered the health state utility to be lower than was plausible for such an injury and therefore lacking in

face validity. The EAG therefore explored a scenario with a utility of 0.727, equivalent to EQ5D utility for a person with some mobility problems and some pain (overall profile 21121).

Scenario 8: Use of technology in all fractures.

The EAG base case assumed that the technology was only used in the three types of fractures considered in the analysis. However, the EAG considered it plausible that the software would be enabled for other fracture types. The EAG therefore conducted a scenario representing an additional cost associated with those other uses but assumed that zero benefit was gained from those reads. This therefore represents a pessimistic scenario of the broader use of the technology (Table 41).

Table 41: Casemix under scenario 8

	Casemix		**Attendances**	
Fracture type	Base case	Scenario	Base case	Scenario
Ankle/foot	38.5%	12.5%	513	513
Wrist/hand	23%	7.5%	308	308
Hip	38.5%	12.5%	513	513
Other		67.5%	0	2772

Source: adapted from expert opinion

Scenario 9: Second read of all scans

The EAG base case used data on diagnostic accuracy as extracted from the two source studies. These comprised a single interpretation of the X-ray, although a second review of diagnoses was considered to be more reflective of real practice. To represent this, the EAG conducted a scenario where all scans were read a second time. It was assumed that the diagnostic accuracy of the second reader was the same as the first. As this was common to all arms, the cost was excluded from analysis.

8.4. Results from the economic modelling

8.4.1. Base case results

Overall, the majority of the AI-assisted diagnostic algorithms were associated with a positive incremental net health benefit compared with unassisted diagnosis at £20,000 and £30,000 thresholds, although 95% credibility intervals in most cases crossed zero, both for all fracture types and when considered together (Table 42 to Table 47). As noted previously, due to data limitations, the EAG advises against comparisons between the individual algorithms. The

minimum economically justified prices were somewhat above the proposed per-scan prices proposed by the companies. However, the EAG also cautions against use of these data to inform pricing decisions, as the modelling was only considered suitable for an indicative 'signal' of cost-effectiveness. More detailed analysis would be required to estimate a suitable maximum price per scan, along with fully incremental analysis to compare the benefits and costs of all the algorithms against one another.

Table 42: Base Case: Ankle/foot

intervention	cost	cost_95CI	QALYs	QALYs_95CI	INHB20k	INHB20k_95CI	INHB30k	INHB30k_95CI
BoneView	637.52	518.33, 765.48	0.786	0.773, 0.798	0.001	-0.003, 0.006	0.002	-0.001, 0.004
Rayvolve	902.89	772.76, 1042.83	0.786	0.773, 0.798	-0.012	-0.018, -0.007	-0.008	-0.011, -0.004
RBfracture	[redacted]	[redacted]	0.785	0.772, 0.797	[redacted]	[redacted]	[redacted]	[redacted]
TechCare Alert	[redacted]	[redacted]	0.785	0.773, 0.798	[redacted]	[redacted]	[redacted]	[redacted]
Unassisted	633.74	519.00, 757.60	0.784	0.772, 0.797	0.000	0.000, 0.000	0.000	0.000, 0.000

Abbreviations: CI, confidence interval; INHB20k/INHB30k, incremental health benefit at willingness to pay threshold of £20/30k (versus unassisted); QALY, Quality Adjusted Life Year

Table 43: Base Case: Wrist/Hand

Intervention	cost	cost_95CI	QALYs	QALYs_95CI	INHB20k	INHB20k_95CI	INHB30k	INHB30k_95CI
BoneView	763.07	686.87, 841.19	0.398	0.386, 0.409	0.001	-0.001, 0.002	0.001	-0.001, 0.002
Rayvolve	769.61	700.57, 838.68	0.398	0.386, 0.409	0.001	-0.002, 0.002	0.001	-0.002, 0.002
RBfracture	[redacted]	[redacted]	0.397	0.386, 0.409	[redacted]	[redacted]	[redacted]	[redacted]
TechCare Alert	[redacted]	[redacted]	0.398	0.386, 0.409	[redacted]	[redacted]	[redacted]	[redacted]
Unassisted	768.11	695.97, 842.02	0.397	0.386, 0.408	0.000	0.000, 0.000	0.000	0.000, 0.000

Abbreviations: CI, confidence interval; INHB20k/INHB30k, incremental health benefit at willingness to pay threshold of £20/30k (versus unassisted); QALY, Quality Adjusted Life Year

Table 44: Base Case: Hip

intervention	cost	cost_95CI	QALYs	QALYs_95CI	INHB20k	INHB20k_95CI	INHB30k	INHB30k_95 CI
BoneView	16,761.80	14,993.44, 18,640.24	10.431	5.660, 13.075	0.08	-0.01, 0.179	0.054	-0.006, 0.119
Rayvolve	25,806.08	23,809.65, 27,845.00	10.431	5.660, 13.075	-0.372	-0.481, -0.259	-0.248	-0.321, -0.172
RBfracture	[redacted]	[redacted]	10.431	5.659, 13.075	[redacted]	[redacted]	[redacted]	[redacted]
TechCare Alert	[redacted]	[redacted]	10.431	5.660, 13.075	[redacted]	[redacted]	[redacted]	[redacted]
Unassisted	18,363.21	16,179.10, 20,612.27	10.431	5.659, 13.075	0.000	0.000, 0.000	0.000	0.000, 0.000

Abbreviations: CI, confidence interval; INHB20k/INHB30k, incremental health benefit at willingness to pay threshold of £20/30k (versus unassisted); QALY, Quality Adjusted Life Year

Table 45: Base Case: Overall

interventions	cost	cost_95CI	QALYs	QALYs_95CI	INHB20k	INHB20k_95CI	INHB30k	INHB30k_95CI
BoneView	6870.15	5078.67, 8825.15	4.408	2.344, 6.187	0.032	-0.003, 0.072	0.021	-0.002, 0.048
Rayvolve	10453.83	7826.39, 13261.71	4.408	2.344, 6.187	-0.148	-0.207, -0.096	-0.098	-0.137, -0.064
RBfracture	[redacted]	[redacted]	4.407	2.344, 6.187	[redacted]	[redacted]	[redacted]	[redacted]
TechCare Alert	[redacted]	[redacted]	4.408	2.344, 6.187	[redacted]	[redacted]	[redacted]	[redacted]
Unassisted	7485.85	5514.65, 9636.13	4.407	2.343, 6.186	0.000	0.000, 0.000	0.000	0.000, 0.000

Abbreviations: CI, confidence interval; INHB20k/INHB30k, incremental health benefit at willingness to pay threshold of £20/30k (versus unassisted); QALY, Quality Adjusted Life Year

Table 46: Population level results (point estimates, based on 1334 patients scanned)

interventions	Cost vs unassisted	QALYs vs Unassisted	INHB20k	INHB30k
BoneView	-821,343.80	1.334	42.69	28.01
Rayvolve	3,959,285.32	1.334	-197.43	-130.73
RBfracture	[redacted]	0	[redacted]	[redacted]
TechCare Alert	[redacted]	1.334	[redacted]	[redacted]
Unassisted	0.00	0	0	0

Abbreviations: INHB20k/INHB30k, incremental health benefit at willingness to pay threshold of £20/30k; QALY, Quality Adjusted Life Year

Table 47: Base Case Maximum Economically Justified Price

	BoneView	Rayvolve	RBfracture	TechCare Alert	Unassisted
combined_ejp20K	634.15	-2,952.34	[redacted]	[redacted]	0
combined_ejp30K	642.60	-2,944.27	[redacted]	[redacted]	0

Abbreviations: ejp, economically justifiable price

8.4.2. Scenario analysis results

Results are summarised in Table 48. Scenario analyses suggested that the decision was sensitive to the optimistic and pessimistic scenarios (scenarios 1 & 2), implying there was much uncertainty in the data. However, the results suggested that there was the potential for AI-assisted diagnosis of fracture to be cost-effective. Results were broadly insensitive to the low and high pricing scenarios for [redacted] (scenarios 3 & 4), when accounting for time taken to interpret X-rays (scenarios 5 & 6), adjusting for the health state utility of negative

ankle and foot fractures (scenario 7), use across all fractures rather than just the three locations analyses (scenario 8), and whether scans are read once or twice (scenario 9).

Whilst the scenario analyses of this EVA suggested that the only parameters to which the results were sensitive was the sensitivity and specificity of the diagnostic scans, the EAG advises caution in interpretation as more detailed modelling and analysis may lead to differing conclusions.

Table 48 Scenario analysis results (overall fractures)

Scenario	intervention	cost	cost_95CI	QALYs	QALYs_95CI	INHB20k	INHB20k_95 CI	INHB30k	INHB30k_95 CI
1	BoneView	[redacted]	[redacted]	4.414	2.360, 6.191	[redacted]	[redacted]	[redacted]	[redacted]
Optimistic	Rayvolve	[redacted]	[redacted]	4.414	2.360, 6.191	[redacted]	[redacted]	[redacted]	[redacted]
	RBfracture	[redacted]	[redacted]	4.414	2.359, 6.191	[redacted]	[redacted]	[redacted]	[redacted]
	TechCare Alert	[redacted]	[redacted]	4.414	2.360, 6.191	[redacted]	[redacted]	[redacted]	[redacted]
	Unassisted	[redacted]	[redacted]	4.412	2.359, 6.189	[redacted]	[redacted]	[redacted]	[redacted]
2	BoneView	[redacted]	[redacted]	4.400	2.330, 6.208	[redacted]	[redacted]	[redacted]	[redacted]
Pessimistic	Rayvolve	[redacted]	[redacted]	4.401	2.331, 6.208	[redacted]	[redacted]	[redacted]	[redacted]
	RBfracture	[redacted]	[redacted]	4.400	2.330, 6.207	[redacted]	[redacted]	[redacted]	[redacted]
	TechCare Alert	[redacted]	[redacted]	4.401	2.330, 6.208	[redacted]	[redacted]	[redacted]	[redacted]
	Unassisted	[redacted]	[redacted]	4.400	2.330, 6.208	[redacted]	[redacted]	[redacted]	[redacted]
3	BoneView	6,887.39	5,122.73, 8,840.88	4.423	2.411, 6.193	0.032	-0.004, 0.072	0.021	-0.002, 0.048
High cost	Rayvolve	10,479.79	7,871.55, 13,236.42	4.423	2.411, 6.193	-0.148	-0.206, -0.096	-0.098	-0.137, -0.064
	RBfracture	[redacted]	[redacted]	4.422	2.411, 6.192	[redacted]	[redacted]	[redacted]	[redacted]
	TechCare Alert	[redacted]	[redacted]	4.423	2.411, 6.193	[redacted]	[redacted]	[redacted]	[redacted]

Scenario	intervention	cost	cost_95CI	QALYs	QALYs_95CI	INHB20k	INHB20k_95 CI	INHB30k	INHB30k_95 CI
	Unassisted	7,504.71	5,546.24, 9,674.39	4.422	2.411, 6.192	0	0, 0	0	0, 0
4	BoneView	6,877.47	5,109.63, 8,888.22	4.415	2.391, 6.202	0.032	-0.004, 0.071	0.021	-0.002, 0.048
Low cost	Rayvolve	10,464.16	7,879.47, 13,280.19	4.415	2.391, 6.202	-0.148	-0.207, -0.097	-0.098	-0.138, -0.064
	RBfracture	[redacted]	[redacted]	4.414	2.391, 6.201	[redacted]	[redacted]	[redacted]	[redacted]
	TechCare Alert	[redacted]	[redacted]	4.415	2.391, 6.202	[redacted]	[redacted]	[redacted]	[redacted]
	Unassisted	7,492.89	5,542.73, 9,712.71	4.414	2.390, 6.201	0	0, 0	0	0, 0
5	BoneView	6,896.95	5,069.79, 8,933.91	4.427	2.370, 6.223	0.032	-0.004, 0.071	0.021	-0.002, 0.047
Time to interpret – junior	Rayvolve	10,489.87	7,824.94, 13,338.37	4.427	2.369, 6.222	-0.148	-0.207, -0.096	-0.098	-0.138, -0.064
	RBfracture	[redacted]	[redacted]	4.427	2.369, 6.222	[redacted]	[redacted]	[redacted]	[redacted]
	TechCare Alert	[redacted]	[redacted]	4.427	2.369, 6.222	[redacted]	[redacted]	[redacted]	[redacted]
	Unassisted	7,513.7	5,518.44, 9,725.16	4.426	2.369, 6.221	0	0, 0	0	0, 0
6	BoneView	6,872.48	5,091.02, 8,851.83	4.404	2.350, 6.194	0.031	-0.004, 0.072	0.021	-0.002, 0.048
Time to interpret - senior	Rayvolve	10,456.17	7,833.18, 13,288.60	4.404	2.350, 6.194	-0.148	-0.208, -0.094	-0.098	-0.138, -0.063
	RBfracture	[redacted]	[redacted]	4.403	2.350, 6.193	[redacted]	[redacted]	[redacted]	[redacted]
	TechCare Alert	[redacted]	[redacted]	4.404	2.350, 6.194	[redacted]	[redacted]	[redacted]	[redacted]

Scenario	intervention	cost	cost_95CI	QALYs	QALYs_95CI	INHB20k	INHB20k_95 CI	INHB30k	INHB30k_95 CI
	Unassisted	7,482.96	5,522.09, 9,663.16	4.403	2.350, 6.193	0	0, 0	0	0, 0
7	BoneView	6,864.76	5,042.82, 8,809.68	4.410	2.331, 6.186	0.031	-0.004, 0.071	0.021	-0.002, 0.048
Angle and foot negatives utility	Rayvolve	10,448.15	7,766.12, 13,280.91	4.410	2.331, 6.186	-0.148	-0.208, -0.095	-0.098	-0.138, -0.063
	RBfracture	[redacted]	[redacted]	4.409	2.331, 6.186	[redacted]	[redacted]	[redacted]	[redacted]
	TechCare Alert	[redacted]	[redacted]	4.410	2.331, 6.186	[redacted]	[redacted]	[redacted]	[redacted]
	Unassisted	7,474.09	5,447.64, 9,606.19	4.409	2.331, 6.185	0	0, 0	0	0, 0
8	BoneView	2,256.76	0.01, 15,714.09	1.456	0.000, 10.395	0.010	-0.001, 0.084	0.007	0.000, 0.056
Use in all fractures	Rayvolve	3,438.05	0.01, 23,809.06	1.456	0.000, 10.395	-0.049	-0.328, 0.000	-0.032	-0.218, 0.000
	RBfracture	[redacted]	[redacted]	1.455	0.000, 10.394	[redacted]	[redacted]	[redacted]	[redacted]
	TechCare Alert	[redacted]	[redacted]	1.456	0.000, 10.395	[redacted]	[redacted]	[redacted]	[redacted]
	Unassisted	2,456.79	0, 16,893.89	1.455	0.000, 10.394	0	0, 0	0	0, 0
9	BoneView	6,879.57	5,105.45, 8,881.25	4.416	2.360, 6.199	0.031	-0.004, 0.072	0.021	-0.003, 0.048
Second read	Rayvolve	10,464.26	7,840.36, 13,330.71	4.416	2.360, 6.199	-0.148	-0.207, -0.096	-0.098	-0.138, -0.064
	RBfracture	[redacted]	[redacted]	4.415	2.359, 6.199	[redacted]	[redacted]	[redacted]	[redacted]
	TechCare Alert	[redacted]	[redacted]	4.415	2.360, 6.199	[redacted]	[redacted]	[redacted]	[redacted]

Scenario	intervention	cost	cost_95CI	QALYs	QALYs_95CI	INHB20k	INHB20k_95 CI	INHB30k	INHB30k_95 CI
	Unassisted	7,489.17	5,504.42, 9,688.56	4.415	2.359, 6.198	0	0, 0	0	0, 0

Abbreviations: CI, confidence interval; INHB20k/30k, incremental health benefit at willingness to pay threshold of £20k/30k.

8.5. Summary and interpretation of the economic evidence

The early modelling for this EVA suggested that, at the proposed prices charged by the companies, AI assisted diagnosis had the potential to represent a value for money investment for the NHS at typical thresholds of £20,000 to £30,000 per QALY gained. However, this conclusion was considered highly uncertain at the present time. The EAG noted that the cost-effectiveness appeared to be driven by reductions in costs rather than a gain in QALYs. The EAG cautions against using this analysis to compare one AI algorithm against another due to data limitations, and instead to consider whether AI-assisted diagnosis as a class would likely be of value. For example, the optimistic and pessimistic scenarios may not fully capture uncertainty and therefore may bias in favour or against technologies with only one source study. Future work, however, will require comparison of one algorithm against another to ensure the varying diagnostic accuracy of the algorithms is matched to their prices in a fully incremental analysis.

Scenario analysis concluded that the only parameters to which the decision was sensitive were the diagnostic accuracy of the algorithms and unassisted diagnosis, with the ‘optimistic’ and ‘pessimistic’ scenarios generating diverging conclusions. Other scenarios, including cost and time savings from AI-assisted diagnosis, low and high price-per-scan scenarios, and varying assumptions around health state utilities, use of AI in all fractures and including a second read of all X-rays did not materially affect the results.

9. EVIDENCE GAPS AND RESEARCH RECOMMENDATIONS

The EAG conducted a broad evidence review to identify the existing evidence base for the use of the technology to assist with the identification of fractures in emergency care settings. The EAG identified:

- 16 studies that evaluated the diagnostic accuracy of the technology
- 7 studies that reported X-ray reading time
- 0 studies that reported other service outcomes
- 0 studies that reported clinical outcomes for people with suspected fractures
- 0 economic evaluations of the technology.

With respect to the specific technologies eligible for consideration, the majority of the evidence evaluated BoneView (9 studies) and RBFracture (5 studies), with 2 studies available for both Rayvolve and TechCare Alert, and 0 studies available for qMSK. The availability of a head-to-head comparison[23] of three of the technologies this early in development of the technologies was notable, although this study lacked the availability of an unassisted arm for comparison purposes. As discussed throughout the report, the evidence base to date was limited in quality, with various risks of bias to the results as well as concerns about the generalisability of the evidence to clinical settings. Due to these issues, the EAG was unable to draw firm conclusions about the potential value of the technology, or identify reliable patterns in the results, such as according to reader experience, case mix, fracture subgroup, or population.

During the assessment, including during interpretation of the evidence base and feedback from clinicians and stakeholders to the assessment, the EAG identified complexities in developing evidence to evaluate interventions for use in people with suspected fractures and the services that treat them. One of the most significant of these considerations was that the target population of people with suspected fractures is highly heterogeneous, comprising of people with a broad range of demographics, type, mechanism and location of injury, and broader health considerations. As the diagnostic accuracy of X-ray and the broader care pathway will vary across populations (including the diagnosing clinicians, use of additional imaging modalities, use of precautionary tactics, and ongoing treatments), this means that the potential value of the technology will vary according to the population in which its used. The evaluation of the

technology within a representative population will therefore be key to deriving reliable estimates for outcomes. However, the target population is also not static, and will vary between types of urgent care settings (ED, UTC, MIU), across geographical areas in the UK, and will vary even within the same centre according to the day of the week, time of day, or season. Each centre will also vary in the local policies that they use to diagnose certain types of fractures and fractures in certain subpopulations, including their use of precautionary tactics but also including the typical staff available to read X-rays and the length of time until a definitive diagnosis is made. To inform understanding of the potential value of the technology, it will therefore be important to understand the way in which outcomes for the technology change according to differences in the population case mix, reader, and care pathway.

The economic analysis contained within this report represented a very top-level overview of the likely costs and consequences of adopting AI-assisted diagnosis of fracture within urgent care settings. Due to the resource and time constraints of this EVA, the EAG was unable to explore a large number of issues and nuances apparent in the data and a substantive evidence synthesis project of 12-18 months' duration would, in the opinion of the EAG, provide a solid appraisal of the evidence to fully inform a decision as to whether AI-assisted diagnosis of fracture represented a value for money investment in the NHS. Particular issues this should consider include:

- **Evaluation of the diagnostic accuracy, clinical and service outcomes associated with the technology within robust study designs within settings comparable with the likely use of the technology in clinical practice.**

  This should include diagnostic randomised controlled trials and prospective, robustly sampled comparative studies (assisted vs. unassisted diagnosis) published in peer-reviewed formats. Given that the technology would be expected to influence both clinical and service outcomes, future studies are needed to evaluate outcomes across both of these domains.

- **Formal network meta-analysis of studies and/or head-to-head studies**

  A formal network meta-analysis was not considered feasible due to heterogeneity between studies (see section 5.2). However, further work is required to determine whether a less formal synthesis could be conducted, or else head-to-head studies are required for all relevant AI algorithms / software.

- **Studies to evaluate the factors that influence the value of the technology for identifying fractures**

  This may include studies designed to explore change in outcomes according to key factors that would inform use of the technology, such as reader experience, case mix, and determinants of patient outcomes, such as patient age, frailty, and prevalence of health conditions affecting bone health. When a suitable evidence base is available, meta-regression to explore factors that influence outcomes may help decision-makers to target the use of the technology in clinical practice.

- **Analysis of optimal cutoff points.**

  This analysis made use of the stated sensitivity and specificity estimates from the literature. However, manufacturers of the algorithms (as well as the NHS) should consider the most cost-effective cut off score to maximise the efficiency of diagnosis. For example, some algorithms will generate a propensity score or probability of an X-ray being a fracture. An internal setting will determine what score and above is defined as a positive. Varying this score varies the sensitivity and specificity (from which the ROC curve can be generated). Connecting this to a decision model allows estimation of the optimal cutoff score for an algorithm (Laking et al 2006[47])

- **Greater exploration of the likely longer-term costs and consequences of true and false positive and negative diagnoses.**

  This analysis relied on published studies to approximate long term costs and consequences of the four outcomes from a diagnosis. These varied in comparability, eg in terms of scope of resource use included and time horizon. A more comprehensive model breaking these items down in greater detail would enhance comparability between the different fracture types.

- **Impact of detecting multiple fractures.**

  The economic analysis did not differentiate between a single and multiple fracture in a single patient. A particular use case and benefit of AI-assisted diagnosis could be in identifying a less obvious additional fracture which may be more likely to be missed by the reader. Further research into the benefit of this is warranted.

- **Second read of only positives or only negatives**

Scenario 9 of the economic analysis only considered a second read / review of all X-rays. An alternative approach is to review all positive or all negative diagnoses alone. Additional modelling would facilitate exploring the cost-effectiveness of this which may assist in enhancing the efficiency of the diagnostic X-ray service in clinical practice.

- **Formal assessment of study quality**

  Consistent with methods for an EVA, formal quality assessment of the included studies was not conducted, and quality limitations of the included studies was conducted informally and discussed throughout the report. However, formal quality assessment of the current and future evidence base would be useful for characterising the strength of the evidence and identifying key weaknesses and their effect on outcomes. For the assessment of diagnostic accuracy studies, QUADAS-AI,[48] an extension to the original tool to account for the considerations specific to AI technologies, would be useful.

10. DISCUSSION

The EAG conducted a broad evidence review to identify the available evidence base for the value of AI as assistance to identifying fractures in urgent care settings. The assessment identified an emerging though limited evidence base for the technology, with meaningful gaps in evidence to inform decision-making on the use of the technology in clinical practice. Almost all of the evidence base identified evaluated the diagnostic accuracy of the technology for identifying fractures, though these analyses were largely not specific to emergency care settings or the staff that were anticipated to typically use the technology in clinical practice. There were also significant methodological limitations in the evidence base, which increased uncertainty in the findings. Aside from X-ray reading time, there was no evidence for the impact of the technology on service outcomes, such as the use of additional imaging, hospital appointments, and patient recalls, and no evidence for the health outcomes of people with suspected fractures. Overall, there was a paucity of evidence for determining the potential value of the technology for use within NHS settings.

In consideration of the limitations of the evidence base, the EAG tentatively concluded that there was early evidence that the technology may have value for reducing the risk of missed fractures but may not improve the avoidance of false positives. Due to uncertainty in the precise estimates reported by the studies, the EAG could not determine a reliable estimate for the reduction of missed fractures that could be avoided with the technology, though noted that the technology did not eradicate missed fractures entirely. Some studies, particularly those evaluating the accuracy of the technology in fracture types that were more difficult to identify, reported high rates of missed fractures that would be unacceptable in clinical practice. The implication of these findings was that while the technology may improve identification of fractures, its use would not remove the need for existing strategies used by urgent care settings to protect patients (such as further imaging, precautionary treatments, and fast turnaround times to definitive reports). As might be expected, there was less additional value of the technology when unassisted accuracy was already high, such as when used by senior staff and in fractures that were easier to diagnose. This would suggest that the technology may best be targeted towards the settings and populations where it may hold the greatest value; however, without evidence for the clinical and service outcomes that may be affected by using the technology, the EAG was unable to make this conclusion. For instance, the EAG considered it plausible that a small difference in sensitivity may nevertheless be meaningful if the additional fractures

identified would have otherwise resulted in significant health or resource implications. As noted in Section 9), further evidence is needed to explore this in order to inform decisions about if and how the technology could be used.

The current evidence base also did not allow for an understanding of how the potential evidence base might vary according to the target population. The majority of studies identified reported very few details about the study sample demographics, including the prevalence of people with frailty and health conditions that affect bone health. Very few studies reported outcomes specifically in children, where the identification of fractures can be particularly difficult. This wasn't notable in the results of the studies, however, where diagnostic accuracy both with and without the technology was not substantially different to results reported in adults alone. Given the limitations in the evidence base, the EAG considered that more evidence to evaluate the technology in children and across other key sub-populations would be important.

The economic analyses suggested that most of the AI assisted diagnostic algorithms were associated with a positive incremental net health benefit compared with unassisted diagnosis at NICE's lower and upper threshold band of £20,000 to £30,000 per QALY. The data were of insufficient quality to enable a robust comparison of one algorithm against another in a fully incremental analysis, which would be required to ensure the varying diagnostic accuracy of the different algorithms are matched to their prices, and to encourage a competitive market for the benefit of NHS patients. The results were largely insensitive to the different scenario analyses considered, except for diagnostic accuracy of the algorithms and unassisted diagnosis.

Overall, the results of this assessment must be considered within the context of the constraints of the EVA methods. Consistent with the aims of the EVA, the EAG adopted a pragmatic approach to the identification of evidence and exploration of this within its assessment. The EAG acknowledge a number of limitations to this approach, which included the use of single reviewer screening and data extraction, the broad inclusion criteria (allowing for the inclusion of lower quality evidence), and the lack of a formal quality assessment. The EAG was also unable to explore heterogeneity in the diagnostic accuracy results in more depth: while groupings of reader experience were made in order to aid interpretation of the results across studies, the EAG considered these to be unreliable, given difficulties in interpreting staff grades and experience as reported from publications of studies conducted in other countries. Given the variation in methods used across studies and the quality limitations of the evidence, further exploration of heterogeneity or re-grouping of studies may not have been meaningful, though

this is a limitation of the assessment. An exploratory economic analysis was developed, the objective of which was to establish whether there was a prima facie case for AI-assisted diagnosis of fracture to represent value for money for NHS patients (i.e. not to provide detailed guidance on and precise estimates of the cost-effectiveness of the different algorithms). It should therefore be considered an approximation placing plausible bounds on the likely costs and consequences of the algorithms and not a definitive estimate of the cost-effectiveness. A large number of gross assumptions were required to conduct the analysis within the assessment, for example the EAG was unable to consider the longer-term costs and consequences of false negatives and positives in anything but the most rudimentary manner. More detailed modelling in a full formal diagnostic assessment review is required to consider these issues and nuances, and how they are likely to impact the estimates of cost-effectiveness.

In conclusion, this assessment identified preliminary evidence that AI may have potential value for use within urgent care settings to aid with the identification of fractures. In order to inform a full evaluation of the technology, that could be used to inform decisions on routine commissioning, a significant body of further evidence is needed to establish the clinical, service and economic outcomes associated with the technology in settings relevant to urgent care within the NHS.

References

1. Hoppe BF, Rueckel J, Dikhtyar Y, et al. Implementing Artificial Intelligence for Emergency Radiology Impacts Physicians' Knowledge and Perception: A Prospective Pre- and Post-Analysis. *Investigative Radiology* 2024;59(5)
2. National Institute for Health and Care Excellence. Fractures (non-complex): assessment and management. NICE guideline [NG38]. Published: 17 February 2016. [Available from: https://www.nice.org.uk/guidance/ng38 accessed 1 July 2024.
3. NHS England. Diagnostic imaging reporting turnaround times 2023 [Available from: https://www.england.nhs.uk/long-read/diagnostic-imaging-reporting-turnaround-times/.
4. National Institute for Health and Care Excellence. Artificial intelligence-derived software to analyse chest X-rays for suspected lung cancer in primary care referrals: early value assessment, 2023.
5. Kuo RYL, Harrison C, Curran TA, et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. *Radiology* 2022;304(1):50-62. doi: https://dx.doi.org/10.1148/radiol.211785
6. Pauling C, Kanber B, Arthurs OJ, et al. Commercially available artificial intelligence tools for fracture detection: the evidence. *BJR|Open* 2024;6(1):tzad005. doi: 10.1093/bjro/tzad005
7. Radiobotics. RBFracture Improves the Diagnostic Accuracy of Emergency Care Professionals., 2021.
8. Bousson V, Attané G, Benoist N, et al. Artificial Intelligence for Detecting Acute Fractures in Patients Admitted to an Emergency Department: Real-Life Performance of Three Commercial Algorithms. *Academic Radiology* 2023;30(10):2118-39. doi: 10.1016/j.acra.2023.06.016
9. Cohen M, Puntonet J, Sanchez J, et al. Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs. *European radiology* 2023;33(6):3974-83. doi: https://dx.doi.org/10.1007/s00330-022-09349-3
10. Meetschen M, Salhofer L, Beck N, et al. AI-Assisted X-ray Fracture Detection in Residency Training: Evaluation in Pediatric and Adult Trauma Patients. *Diagnostics* 2024;14(6):596. doi: https://dx.doi.org/10.3390/diagnostics14060596
11. Canoni-Meynet L, Verdot P, Danner A, et al. Added value of an artificial intelligence solution for fracture detection in the radiologist's daily trauma emergencies workflow. *Diagnostic and interventional imaging* 2022;103(12):594-600. doi: https://dx.doi.org/10.1016/j.diii.2022.06.004
12. Dell'Aria A, Tack D, Saddiki N, et al. Radiographic Detection of Post-Traumatic Bone Fractures: Contribution of Artificial Intelligence Software to the Analysis of Senior and Junior Radiologists. *Journal of the Belgian Society of Radiology* 2024 doi: 10.5334/jbsr.3574
13. Duron L, Ducarouge A, Gillibert A, et al. Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study. *Radiology* 2021;300(1):120-29. doi: https://dx.doi.org/10.1148/radiol.2021203886
14. Guermazi A, Tannoury C, Kompel AJ, et al. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. *Radiology* 2022;302(3):627-36. doi: https://dx.doi.org/10.1148/radiol.210937
15. Nguyen T, Maarek R, Hermann AL, et al. Assessment of an artificial intelligence aid for the detection of appendicular skeletal fractures in children and young adults by senior and junior radiologists. *Pediatric Radiology* 2022;52(11):2215-26. doi: https://dx.doi.org/10.1007/s00247-022-05496-3

16. Oppenheimer J, Luken S, Hamm B, et al. A Prospective Approach to Integration of AI Fracture Detection Software in Radiographs into Clinical Workflow. *Life (Basel, Switzerland)* 2023;13(1) doi: https://dx.doi.org/10.3390/life13010223
17. Bachmann R, Gunes G, Hangaard S, et al. Improving traumatic fracture detection on radiographs with artificial intelligence support: a multi-reader study. *BJR open* 2024;6(1):tzae011. doi: https://dx.doi.org/10.1093/bjro/tzae011
18. Automatic Hip Fracture Detection and Anatomical localisation with Deep Learning. Research Day, Region Zealand; 2023.
19. Ruitenbeek HC, Oei EHG, Schmahl BL, et al. Towards clinical implementation of an AI-algorithm for detection of cervical spine fractures on computed tomography. *European journal of radiology* 2024;173:111375. doi: https://dx.doi.org/10.1016/j.ejrad.2024.111375
20. Yogendra PM, Wei Goh AG, TYing YS, et al. Accuracy of radiologists and radiology residents in detection of paediatric appendicular fractures with

and without the help of Artificial Intelligence. Unpublished (manuscript in preparation).
21. Fu T, Viswanathan V, Attia A, et al. Assessing the Potential of a Deep Learning Tool to Improve Fracture Detection by Radiologists and Emergency Physicians on Extremity Radiographs. *Academic radiology* 2024;31(5):1989-99. doi: https://dx.doi.org/10.1016/j.acra.2023.10.042
22. Parsy et al. MSK – AI - RETROSPECTIVE STUDY: Centre hospitalier de Valenciennes, 2020.
23. Bousson V, Attané G, Benoist N, et al. Artificial Intelligence for Detecting Acute Fractures in Patients Admitted to an Emergency Department: Real-Life Performance of Three Commercial Algorithms. *Acad Radiol* 2023;30(10):2118-39. doi: 10.1016/j.acra.2023.06.016 [published Online First: 20230718]
24. Dell'Aria A, Tack D, Saddiki N, et al. Radiographic Detection of Post-Traumatic Bone Fractures: Contribution of Artificial Intelligence Software to the Analysis of Senior and Junior Radiologists. *Journal of the Belgian Society of Radiology* 2024;108(1) doi: https://dx.doi.org/10.5334/jbsr.3574
25. Nguyen T, Maarek R, Hermann A-L, et al. Assessment of an artificial intelligence aid for the detection of appendicular skeletal fractures in children and young adults by senior and junior radiologists. *Pediatric Radiology* 2022;52(11):2215-26. doi: 10.1007/s00247-022-05496-3
26. Bonde N et al. Hip fracture detection on conventional radiographs using deep learning artificial intelligence as a clinical support tool: A decrease of false negative diagnoses. Unpublished
27. EVALUATING THE IMPACT OF ARTIFICIAL INTELLIGENCE ON FRACTURE DETECTION: A MULTI-CENTER, MULTI-COUNTRY RANDOMIZED CONTROLLED

CLINICAL TRIAL. RSNA 2024; 2024.
28. Radiobotics. TD 21-205 Retrospective CS Report RBfracture v.1 Rev.01 – signed.
29. Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? *Stat Med* 2009;28(21):2653-68. doi: 10.1002/sim.3631
30. The Royal College of Radiologists. Standards for the education and training of reporting practitioners in musculoskeletal plain radiographs 2022 [Available from: https://www.rcr.ac.uk/our-services/all-our-publications/clinical-radiology-publications/standards-for-the-education-and-training-of-reporting-practitioners-in-musculoskeletal-plain-radiographs/.
31. Rua T, Gidwani S, Malhotra B, et al. Cost-Effectiveness of Immediate Magnetic Resonance Imaging In the Management of Patients With Suspected Scaphoid Fracture: Results

From a Randomized Clinical Trial. *Value Health* 2020;23(11):1444-52. doi: 10.1016/j.jval.2020.05.020 [published Online First: 20200929]
32. Nwankwo H, Mason J, Costa ML, et al. Cost-utility analysis of cast compared to removable brace in the management of adult patients with ankle fractures. *Bone & Joint Open* 2022;3(6):455-62. doi: 10.1302/2633-1462.36.Bjo-2022-0036
33. Low YL, Finkelstein E. Cost-Effective Analysis of Dual-Energy Computed Tomography for the Diagnosis of Occult Hip Fractures Among Older Adults. *Value Health* 2021;24(12):1754-62. doi: 10.1016/j.jval.2021.06.005 [published Online First: 20210805]
34. Judge A, Javaid MK, Leal J. Models of care for the delivery of secondary fracture prevention after hip fracture: a health service cost, clinical outcomes and cost-effectiveness study within a region of England: NIHR Journals Library, 2016.
35. Ward JL, Azzopardi PS, Francis KL, et al. Global, regional, and national mortality among young people aged 10–24 years, 1950–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* 2021;398(10311):1593-618. doi: 10.1016/S0140-6736(21)01546-4
36. Cieza A, Causey K, Kamenov K, et al. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* 2020;396(10267):2006-17. doi: 10.1016/S0140-6736(20)32340-0
37. Wu A-M, Bisignano C, James SL, et al. Global, regional, and national burden of bone fractures in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019. *The Lancet Healthy Longevity* 2021;2(9):e580-e92. doi: 10.1016/S2666-7568(21)00172-0
38. Chen C, Lin J-R, Zhang Y, et al. A systematic analysis on global epidemiology and burden of foot fracture over three decades. *Chinese Journal of Traumatology* 2024 doi: https://doi.org/10.1016/j.cjtee.2024.03.001
39. Baji P, Barbosa EC, Heaslip V, et al. Use of removable support boot versus cast for early mobilisation after ankle fracture surgery: cost-effectiveness analysis and qualitative findings of the Ankle Recovery Trial (ART). *BMJ Open* 2024;14(1):e073542. doi: 10.1136/bmjopen-2023-073542 [published Online First: 20240111]
40. Howell M, Lawson A, Naylor J, et al. Surgical plating versus closed reduction for fractures in the distal radius in older patients: a cost-effectiveness analysis from the hospital perspective. *ANZ J Surg* 2022;92(12):3311-18. doi: 10.1111/ans.18134 [published Online First: 20221105]
41. Hannemann PF, Essers BA, Schots JP, et al. Functional outcome and cost-effectiveness of pulsed electromagnetic fields in the treatment of acute scaphoid fractures: a cost-utility analysis. *BMC Musculoskelet Disord* 2015;16:84. doi: 10.1186/s12891-015-0541-2 [published Online First: 20150411]
42. Faccioli N, Santi E, Foti G, et al. Cost-effectiveness of introducing cone-beam computed tomography (CBCT) in the management of complex phalangeal fractures: economic simulation. *Musculoskelet Surg* 2022;106(2):169-77. doi: 10.1007/s12306-020-00687-3 [published Online First: 20201119]
43. Rua T, Malhotra B, Vijayanathan S, et al. Clinical and cost implications of using immediate MRI in the management of patients with a suspected scaphoid fracture and negative radiographs results from the SMaRT trial. *Bone Joint J* 2019;101-b(8):984-94. doi: 10.1302/0301-620x.101b8.Bjj-2018-1590.R1
44. Abimanyi-Ochom J, Watts JJ, Borgström F, et al. Changes in quality of life associated with fragility fractures: Australian arm of the International Cost and Utility Related to Osteoporotic Fractures Study (AusICUROS). *Osteoporos Int* 2015;26(6):1781-90. doi: 10.1007/s00198-015-3088-z [published Online First: 20150320]

45. Keating JF, Grant A, Masson M, et al. Randomized comparison of reduction and fixation, bipolar hemiarthroplasty, and total hip arthroplasty. Treatment of displaced intracapsular hip fractures in healthy older patients. *J Bone Joint Surg Am* 2006;88(2):249-60. doi: 10.2106/jbjs.E.00215
46. Jonsson B, Kanis J, Dawson A, et al. Effect and offset of effect of treatments for hip fracture on health outcomes. *Osteoporos Int* 1999;10(3):193-9. doi: 10.1007/s001980050215
47. Laking G, Lord J, Fischer A. The economics of diagnosis. *Health Econ* 2006;15(10):1109-20. doi: 10.1002/hec.1114
48. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nature Medicine* 2021;27(10):1663-65. doi: 10.1038/s41591-021-01517-0

Appendix A – Search strategies

Date	Database Name[1]	Total Number of records retrieved
26/6/24	Medline ALL	745
26/6/24	Embase	920
26/6/24	The Cochrane Library	57
26/6/24	Web of Science	167
27/6/24	Company websites:	31
1/7/24	Guidelines	5
2/7/24	MHRA	0
2/7/24	FDA	0
2/7/24	Clinical Trials.gov	17
2/7/24	ICTRP	1
	HERC	0
	CEA Registry	15
	Company submissions	18
	Total	1976
	Duplicates	635
	Total to screen	1341

Search strategies

Embase <1974 to 2024 June 25>

1 exp artificial intelligence/ 106277

2 exp *Machine Learning/ 206084

3 ("deep learning" or "artificial neural network*" or "deep neural network*" or "convolutional neural network*").ti,ab,kf. 118642

4 ((machine or transfer or algorithmic) adj2 Learning).ti,ab,kf. 144555

5 ("AI" or "comput* Intelligence" or "comput* reasoning" or "machine Intelligence" or "artificial intelligence").ti,ab,kf. 114372

6 ("neural networks" or "natural language processing" or 'llm*1 or large language model*').ti,ab,kf. 79238

7 ("reinforcement learning" or "deep belief network*" or "recurrent neural network*" or "feedforward neural network*").ti,ab,kf. 14879

8 "feed forward neural network*".ti,ab,kf. 1080

9 ("boltzmann machine*" or "long short-term memory" or "gated recurrent unit*" or "rectified linear unit*" or autoencoder or "auto-encoder" or backpropagation or "multilayer perceptron" or "multi-layer perceptron" or convnet or "convolutional learning").ti,ab,kf. 18309

10 or/1-9 461984

11 "diagnostic imaging".ti,ab,kf. 30009

12 exp diagnostic imaging/ 275165

13 exp X-Ray/ 89150

14 (radiograph* or radiologist or radiogram or XR or x-ray or "radiological image*" or photographic or "digital image*" or radiology or roentgenogram or roentgenograph or "Rontgen ray*" or x-rayed or "x ray*").ti,ab,kf. 977193

15 11 or 12 or 13 or 14 1247290

16 exp fracture/ 370018

17 ((fractur* or break* or fissur* or shatter* or crack* or splinter* or broken or dislocat* or luxat* or subluxat* or trauma or disjoint* or displace*) adj2 (bone* or joint* or skeletal or skeleton)).ti,ab,kf. 41115

18 ((spiral or avulsion or compression or greenstick or "green stick" or intraarticular or "intra articular" or pathologic or stress or comminuted or dislocation or hairline or "hair line" or impacted or longitudinal or oblique or transverse or pathological or insufficiency or vertebral or elbow* or arm* or leg* or ankle* or wrist* or finger* or toe* or pelvis or pelvic or hip* or shoulder* or spine or spinal or chest or rib* or knee* or hand* or foot or feet or face or facial or microfracture or fatigue or macroscopic or periprosthetic) adj2 (fractur* or break* or fissur* or shatter* or crack* or splinter* or broken or injur*)).ti,ab,kf. 264517

19 (("long bone" or "short bone" or "flat bone" or sesamoid or irregular or epiphysis or physis or metaphysis or diaphysis or tubercle or epicondyle or complete or incomplete or displaced or non-displaced or "non displaced" or stable or unstable or simple or closed or segmental or bowing or buckle or oblique or complex or non-complex or "non complex" or salter-harris or "salter harris" or Lisfranc or "distal radial" or "growth plate" or suspect*) adj2 (fractur* or break* or fissur* or shatter* or crack* or splinter* or broken or injur*)).ti,ab,kf. 49830

20 or/16-19 544797

21 10 and 15 and 20 1020

22 (AZmed or "AZ med" or "AZ medical" or AZmedical or Gleamer or Radiobotics or Qure or Milvue).af. 206

23 (Rayvolve or Boneview or "Bone view" or RBfracture or "RB fracture" or qMSK or qXR or qER or "TechCare Alert" or "Tech Care Alert" or SmartUrgence or "Smart Urgence").af. 128

24 21 or 22 or 23 1307

25 limit 24 to (dd=20200701-20240626 or rd=20200701-20240626 or dc=20200701-20240626) 920

Ovid MEDLINE(R) ALL <1946 to June 25, 2024>

1 exp artificial intelligence/ 200607

2 exp Machine Learning/ 70860

3 ("deep learning" or "artificial neural network*" or "deep neural network*" or "convolutional neural network*").ti,ab,kf. 103348

4 ((machine or transfer or algorithmic) adj2 Learning).ti,ab,kf. 125093

5 ("AI" or "comput* Intelligence" or "comput* reasoning" or "machine Intelligence" or "artificial intelligence").ti,ab,kf. 91108

6 ("neural networks" or "natural language processing" or 'llm*1 or large language model*').ti,ab,kf. 66698

7 ("reinforcement learning" or "deep belief network*" or "recurrent neural network*" or "feedforward neural network*").ti,ab,kf. 13366

8 "feed forward neural network*".ti,ab,kf. 839

9 ("boltzmann machine*" or "long short-term memory" or "gated recurrent unit*" or "rectified linear unit*" or autoencoder or "auto-encoder" or backpropagation or "multilayer perceptron" or "multi-layer perceptron" or convnet or "convolutional learning").ti,ab,kf. 16957

10 or/1-9 387820

11 "diagnostic imaging".ti,ab,kf. 21096

12 exp diagnostic imaging/ 2977377

13 X-Rays/ 32478

14 (radiograph* or radiologist or radiogram or XR or x-ray or "radiological image*" or photographic or "digital image*" or radiology or roentgenogram or roentgenograph or "Rontgen ray*" or x-rayed or "x ray*").ti,ab,kf. 823590

15 11 or 12 or 13 or 14 3516695

16 exp fractures, bone/ 215213

17 ((fractur* or break* or fissur* or shatter* or crack* or splinter* or broken or dislocat* or luxat* or subluxat* or trauma or disjoint* or displace*) adj2 (bone* or joint* or skeletal or skeleton)).ti,ab,kf. 31804

18 ((spiral or avulsion or compression or greenstick or "green stick" or intraarticular or "intra articular" or pathologic or stress or comminuted or dislocation or hairline or "hair line" or impacted or longitudinal or oblique or transverse or pathological or insufficiency or vertebral or arm* or leg* or ankle* or wrist* or elbow* or finger* or toe* or pelvis or pelvic or hip* or shoulder* or spine or spinal or chest or rib* or knee* or hand* or foot or feet or face or facial or microfracture or fatigue or macroscopic or periprosthetic) adj2 (fractur* or break* or fissur* or shatter* or crack* or splinter* or broken or injur*)).ti,ab,kf. 211184

19 (("long bone" or "short bone" or "flat bone" or sesamoid or irregular or epiphysis or physis or metaphysis or diaphysis or tubercle or epicondyle or complete or incomplete or displaced or non-displaced or "non displaced" or stable or unstable or simple or closed or segmental or bowing or buckle or oblique or complex or non-complex or "non complex" or salter-harris or "salter harris" or Lisfranc or "distal radial" or "growth plate" or suspect*) adj2 (fractur* or break* or fissur* or shatter* or crack* or splinter* or broken or injur*)).ti,ab,kf. 42244

20 16 or 17 or 18 or 19 384408

21 10 and 15 and 20 885

22 (AZmed or "AZ med" or "AZ medical" or AZmedical or Gleamer or Radiobotics or Qure or Milvue).af. 146

23 (Rayvolve or Boneview or "Bone view" or RBfracture or "RB fracture" or qMSK or qXR or qER or "TechCare Alert" or "Tech Care Alert" or "Smart Urgence" or SmartUrgence).af. 69

24 21 or 22 or 23 1068

25 limit 24 to (ed=20200701-20240626 or dt=20200701-20240626) 745

The Cochrane Library

Date Run: 26/06/2024 16:12:22

ID Search Hits

#1 MeSH descriptor: [Artificial Intelligence] explode all trees 3198

#2 MeSH descriptor: [Machine Learning] explode all trees 986

#3 ("deep learning" or artificial NEXT neural NEXT network* or deep NEXT neural NEXT network* or convolutional NEXT neural NEXT network*):ti,ab,kw 1772

#4 ((machine or transfer or algorithmic) near/3 Learning):ti,ab,kw 3228

#5 ("AI" or comput* NEXT Intelligence or comput* NEXT reasoning or "machine Intelligence" or "artificial intelligence"):ti,ab,kw 7206

#6 ("neural networks" or "natural language processing" or llm*1 or large NEXT language NEXT model*):ti,ab,kw 1291

#7 ("reinforcement learning" or deep NEXT belief NEXT network* or recurrent NEXT neural NEXT network* or feedforward NEXT neural NEXT network*):ti,ab,kw 306

#8 feed NEXT forward NEXT neural NEXT network*:ti,ab,kw 23

#9 (boltzmann NEXT machine* or "long short-term memory" or gated NEXT recurrent NEXT unit* or rectified NEXT linear NEXT unit* or autoencoder or "auto-encoder" or backpropagation or "multilayer perceptron" or "multi-layer perceptron" or convnet or "convolutional learning"):ti,ab,kw 189

#10 #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 13275

#11 "diagnostic imaging":ti,ab,kw 44309

#12 MeSH descriptor: [Diagnostic Imaging] explode all trees 69276

#13 MeSH descriptor: [X-Rays] explode all trees 106

#14 (radiograph* or radiologist or radiogram or XR or x-ray or radiological NEXT image* or photographic or digital NEXT image* or radiology or roentgenogram or roentgenograph or Rontgen NEXT ray* or x-rayed or "x ray" or "x rayed"):ti,ab,kw 59279

#15 #11 or #12 or #13 or #14 116068

#16 MeSH descriptor: [Fractures, Bone] explode all trees 9509

#17 ((fractur* or break* or fissur* or shatter* or crack* or splinter* or broken or dislocat* or luxat* or subluxat* or trauma or disjoint* or displace*) near/3 (bone* or joint* or skeletal or skeleton)):ti,ab,kw 7070

#18 ((spiral or avulsion or compression or greenstick or "green stick" or intraarticular or "intra articular" or pathologic or stress or comminuted or dislocation or hairline or "hair line" or impacted or longitudinal or oblique or transverse or pathological or insufficiency or vertebral or elbow* or arm* or leg* or ankle* or wrist* or finger* or toe* or pelvis or pelvic or hip* or shoulder* or spine or spinal or chest or rib* or knee* or hand* or foot or feet or face or facial or microfracture or fatigue or macroscopic or periprosthetic) near/2 (fractur* or break* or fissur* or shatter* or crack* or splinter* or broken or injur*)):ti,ab,kw 25565

#19 (("long bone" or "short bone" or "flat bone" or sesamoid or irregular or epiphysis or physis or metaphysis or diaphysis or tubercle or epicondyle or complete or incomplete or displaced or non-displaced or "non displaced" or stable or unstable or simple or closed or segmental or bowing or buckle or oblique or complex or non-complex or "non complex" or salter-harris or "salter harris" or Lisfranc or "distal radial" or "growth plate" or suspect*) near/2 (fractur* or break* or fissur* or shatter* or crack* or splinter* or broken or injur*)):ti,ab,kw 3290

#20 #16 or #17 or #18 or #19 32637

#21 #10 and #15 and #20 53

#22 (AZmed or "AZ med" or "AZ medical" or AZmedical or Gleamer or Radiobotics or Qure or Milvue) 50

#23 (Rayvolve or Boneview or "Bone view" or RBfracture or "RB fracture" or qMSK or qXR or qER or "TechCare Alert" or "Tech Care Alert" or SmartUrgence or "Smart Urgence") 24

#24 #21 or #22 or #23 125

Limit to 2020-2024 57

Web of Science

#1 TS=("boltzmann machine*" or "long short-term memory" or "gated recurrent unit*" or "rectified linear unit*" or autoencoder or "auto-encoder" or backpropagation or "multilayer perceptron" or "multi-layer perceptron" or convnet or "convolutional learning") 109,743

#2 TS=("feed forward neural network*") 7,197

#3 TS=("reinforcement learning" or "deep belief network*" or "recurrent neural network*" or "feedforward neural network*") 102,380

#4 TS=("neural networks" or "natural language processing" or llm*1 or "large language model*") 413,303

#5 TS=(("AI" or "comput* Intelligence" or "comput* reasoning" or "machine Intelligence" or "artificial intelligence")) 263,773

#6 TS=(((machine or transfer or algorithmic) N2 Learning)) 218

#7 TS=("deep learning" or "artificial neural network*" or "deep neural network*" or "convolutional neural network*") 503,397

#8 #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 1,003,767

#9 TS=(("long bone" or "short bone" or "flat bone" or sesamoid or irregular or epiphysis or physis or metaphysis or diaphysis or tubercle or epicondyle or complete or incomplete or displaced or non-displaced or "non displaced" or stable or unstable or simple or closed or segmental or bowing or buckle or oblique or "distal radial" or Lisfranc or complex or non-complex or "non complex" or salter-harris or "salter harris" or "growth plate" or suspect*) N2 (crack* or splinter* or broken or injur*)) 406

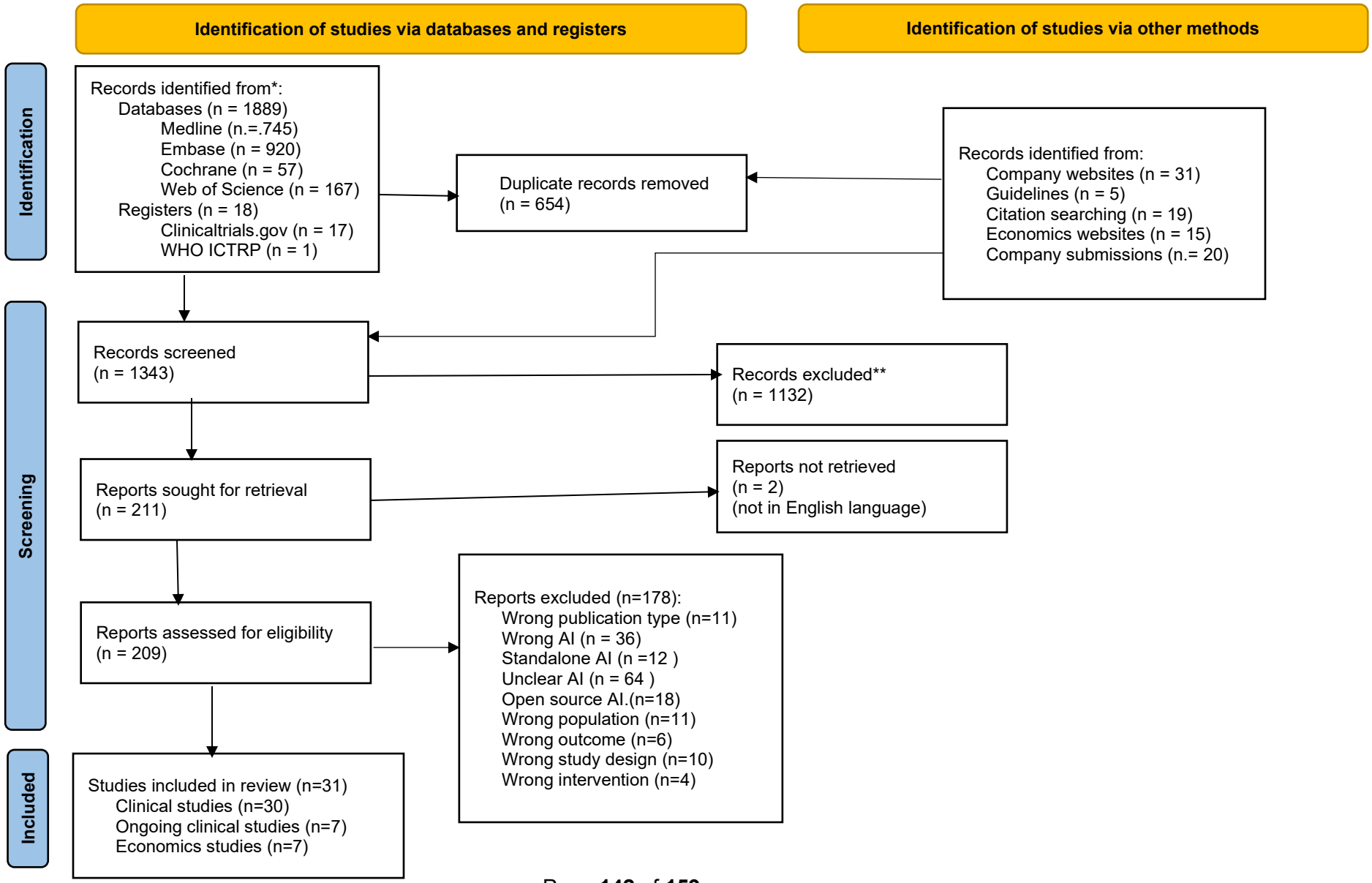
#10 TS=(("long bone" or "short bone" or "flat bone" or sesamoid or irregular or epiphysis or physis or metaphysis or diaphysis or tubercle or epicondyle or complete or incomplete or displaced or non-displaced or "non displaced" or stable or unstable or simple or closed or segmental or bowing or buckle or oblique or complex or non-complex or "non complex" or salter-harris or "salter harris" or Lisfranc or "distal radial" or "growth plate" or suspect*) N2 (fractur* or break* or fissur* or shatter*)) 467

#11 TS=((spiral or avulsion or compression or greenstick or "green stick" or intraarticular or "intra articular" or pathologic or stress or comminuted or dislocation or hairline or "hair line" or impacted or longitudinal or oblique or transverse or pathological or insufficiency or vertebral or arm* or leg* or ankle* or wrist* or pelvis or pelvic or hip* or shoulder* or spine or spinal or chest or rib* or knee* or hand* or elbow* or finger* or toe* or foot or feet or face or facial or microfracture or fatigue or macroscopic or periprosthetic) N2 (crack* or splinter* or broken or injur*)) 416

#12 TS=((spiral or avulsion or compression or greenstick or "green stick" or intraarticular or "intra articular" or pathologic or stress or comminuted or dislocation or hairline or "hair line" or impacted or longitudinal or oblique or transverse or pathological or insufficiency or vertebral or arm* or leg* or ankle* or wrist* or pelvis or pelvic or hip* or shoulder* or spine or spinal or chest or rib* or knee* or hand* or elbow* or foot or feet or finger* or toe* or face or facial or microfracture or fatigue or macroscopic or periprosthetic) N2 (fractur* or break* or fissur* or shatter*)) 427

#13 TS=((fractur* or break* or fissur* or shatter* or crack* or splinter* or broken or dislocat* or luxat* or subluxat* or trauma or disjoint* or displace*) N2 (bone* or joint* or skeletal or skeleton)	73
#14 #9 OR #10 OR #11 OR #12 OR #13	1,259
#15 #8 AND #14	7
#16 ALL=(AZmed or "AZ med" or "AZ medical" or AZmedical or Gleamer or Radiobotics or Qure or Milvue)	273
#17 ALL=(Rayvolve or Boneview or "Bone view" or RBfracture or "RB fracture" or qMSK or qXR or qER or "TechCare Alert" or "Tech Care Alert" or SmartUrgence or "Smart Urgence")	107
#18 #15 OR #16 OR #17 and 2020 or 2021 or 2022 or 2023 or 2024 (Publication Years)	167

Appendix B – PRISMA diagrams



*Consider, if feasible to do so, reporting the number of records identified from each database or register searched (rather than the total number across all databases/registers).
**If automation tools were used, indicate how many records were excluded by a human and how many were excluded by automation tools.

Source: Page MJ, et al. BMJ 2021;372:n71. doi: 10.1136/bmj.n71.

Appendix C – Excluded studies

Table 49: List of excluded English-language publications studies from company lists, with reasons

Excluded study	Reason for exclusion
Gleamer	
Altmann-Schneider 2023	Standalone AI
Hayashi 2022	Standalone AI
Pediatric FRACTURE study (https://fracturestudy.com)	Ongoing study. Unclear but appears as standalone AI.
Regnard et al 2022	Standalone AI
Russe et al. 2024	Standalone AI
Boginskis 2023	Wrong outcome (no scoped outcomes reported)
Jacques 2024	Wrong reference standard (CT)
Hoppe 2024	De-prioritised as low priority outcome (clinician acceptability)
Rosa 2023	Standalone AI
Milvue	
Fanni et al 2023	Wrong study design (literature review)
Parpaleix et al 2023	Standalone AI
Ranjan et al 2023	Wrong study design (narrative review)
Van Leeuwen 2024	Wrong population (bone age and lung nodule detection)
Zerouali et al 2023	Wrong population (estimating alignment)
Lind Plesner 2023	Wrong population (airspace disease, pneumothorax, pleural effusion)
Shelmerdine 2022	Standalone AI

Excluded study	Reason for exclusion
Radiobotics	
Radiobotics S3. Anyebe 2023	Standalone AI
Confidential doc 88 data description for verification Rev 10	No outcome data
Radiobotics. S4 confidential company submission. Post-Market Surveillance & Collection of Metrics	Wrong study design (details of data to be collected through post-marketing surveillance)
Radiobotics. S5 confidential company submission. Artificial Intelligence (AI) in a Singaporean Emergency Department: Detecting Fractures and Reducing Recalls	Standalone AI
Radiobotics. S6 confidential company submission. Deploying Artificial Intelligence in the Detection of Adult Appendicular Fractures in the Emergency Department After-hours: Efficacy, Cost-savings and Non-monetary Benefits.	Standalone AI
Radiobotics. S10 confidential company submission. Evaluation of generalizability of an ai tool for radiographic fracture detection in a multi-country performance study.	Standalone AI
Radiobotics. S12 confidential company submission. Is acute fracture assessment improved with AI support?	Wrong study design (summary of evidence for Radiobotics, no new data)
Radiobotics. S2. Reducing Missed Fractures: Radiobotics' RBfracture™ at Kettering General Hospital. 2024.	Standalone AI
Radiobotics. S13. One Laudos, BZ. RBfracture retrospective pilot study	Standalone AI

Table 50: List of excluded full-text publications from EAG evidence search, with reasons

Excluded study	Reason for exclusion
Clinical Validation of Boneview for FDA Submission: Evaluation of the Ability of the Artificial Intelligence Software, Boneview, to Improve Physicians' and Radiologists' Performances in Detecting Fractures on Bone X-Rays Radiographs 2020.	Wrong publication type
Retrospective Study Comparing Radiologist Diagnostic Performance Versus Artificial Intelligence (AI) for Hip Fracture Suspicion in Elderly Patients 2020.	Wrong publication type
Multicenter Validation Study of an Artificial Intelligence Tool for Automatic Classification of Chest X-rays 2021.	Unclear AI

Excluded study	Reason for exclusion
Assessment of AI Performance for the Detection of Bone Fractures in Children Aged Less Than 2 Years Old in Suspected Child Abuse Setting 2022.	Standalone AI
A Prospective Observational Study of Artificial Intelligence Morphometric Evaluation of Vertebral Fractures 2024.	Wrong AI
Altmann-Schneider I, Pistorius S, Saladin C, Schafer D, Fischer H, Arslan N, et al. Diagnostic performance of an artificial intelligence aid for the detection of pediatric appendicular skeletal fractures. Pediatric Radiology. 2023;53(Supplement 2):S178-S9.	Standalone AI
Ananda A, Ngan KH, Karabag C, Ter-Sarkisov A, Alonso E, Reyes-Aldasoro CC. Classification and Visualisation of Normal and Abnormal Radiographs; A Comparison between Eleven Convolutional Neural Network Architectures. Sensors (Basel, Switzerland). 2021;21(16).	Open source AI
Anderson PG, Baum GL, Keathley N, Sicular S, Venkatesh S, Sharma A, et al. Deep Learning Assistance Closes the Accuracy Gap in Fracture Detection Across Clinician Types. Clinical orthopaedics and related research. 2023;481(3):580-8.	Wrong AI
Anttila TT, Karjalainen TV, Makela TO, Waris EM, Lindfors NC, Leminen MM, et al. Detecting Distal Radius Fractures Using a Segmentation-Based Deep Learning Model. Journal of digital imaging. 2023;36(2):679-87.	Open source AI
Aryasomayajula S, Hing CB, Siebachmeyer M, Naeini FB, Ejindu V, Leitch P, et al. Developing an artificial intelligence diagnostic tool for paediatric distal radius fractures, a proof of concept study. Annals of the Royal College of Surgeons of England. 2023;105(8):721-8.	Unclear AI
Bartha E, Davidson T, Hommel A, Thorngren KG, Carlsson P, Kalman S. Cost-effectiveness analysis of goal-directed hemodynamic treatment of elderly hip fracture patients: before clinical research starts. Anesthesiology. 2012;117(3):519-30.	Wrong population
Beaupre LA, Lier D, Smith C, Evens L, Hanson HM, Juby AG, et al. A 3i hip fracture liaison service with nurse and physician co-management is cost-effective when implemented as a standard clinical program. Arch Osteoporos. 2020;15(1):113.	Wrong population
Bettinger H, Lenczner G, Guigui J, Rotenberg L, Zerbib E, Attia A, et al. Evaluation of the Performance of an Artificial Intelligence (AI) Algorithm in Detecting Thoracic Pathologies on Chest Radiographs. Diagnostics (Basel, Switzerland). 2024;14(11).	Wrong population
Beyaz S, Acici K, Sumer E. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. Joint diseases and related surgery. 2020;31(2):175-83.	Unclear AI
Bhandarkar AR, Onyedimma C, Jarrah R, Fu S, Liu H, Bydon M. An Integrated Voice Recognition and Natural Language Processing Platform to Automatically Extract Thoracolumbar Injury Classification Score (TLICS) Features from Radiology Reports. Clinical Neurosurgery. 2022;68:50-1.	Wrong population
Boginskis V, Zadoroznijs S, Cernavska I, Beikmane D, Sauka J. Artificial intelligence effectivity in fracture detection. Medicni perspektivi. 2023;28(3):68-78.	Wrong outcome (no scoped outcomes reported)

Excluded study	Reason for exclusion
Bulstra AEJ, Buijze GA, Cohen A, Colaris JW, Court-Brown CM, Doornberg JN, et al. A Machine Learning Algorithm to Estimate the Probability of a True Scaphoid Fracture After Wrist Trauma. Journal of Hand Surgery. 2022;47(8):709-18.	Unclear AI
Burkow J, Holste G, Otjen J, Perez F, Junewick J, Zbojniewicz A, et al. High sensitivity methods for automated rib fracture detection in pediatric radiographs. Scientific reports. 2024;14(1):8372.	Unclear AI
Burkow J, Holste G, Perez F, Junewick J, Zbojniewicz A, Frost J, et al. Rib fracture detection in pediatric radiographs via deep convolutional neural networks. Pediatric Radiology. 2021;51:S125.	Unclear AI
Chen C, Zhang Z, Tsai T, Kuo K. Artificial intelligence to improve osteoporosis screening on x-ray radiographs. Osteoporosis International. 2020;31:S352.	Unclear AI
Chen H-Y, Hsu BW-Y, Yin Y-K, Lin F-H, Yang T-H, Yang R-S, et al. Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs. PloS one. 2021;16(1):e0245992.	Unclear AI
Chen HY, Soong C, Lin FH, Yang TH, Chan DC, Chang CH, et al. Application of deep learning algorithm to detect and visualize vertebral fractures on plain radiographs. Osteoporosis and Sarcopenia. 2023;9(4):S3.	Unclear AI
Cheng C-T, Chen C-C, Cheng F-J, Chen H-W, Su Y-S, Yeh C-N, et al. A Human-Algorithm Integration System for Hip Fracture Detection on Plain Radiography: System Development and Validation Study. JMIR medical informatics. 2020;8(11):e19416.	Unclear AI
Cheng CT, Hsu CP, Ooyang CH, Chou CY, Lin NY, Lin JY, et al. Evaluation of ensemble strategy on the development of multiple view ankle fracture detection algorithm. British Journal of Radiology. 2023;96(1145):20220924.	Open source AI
Cheng CT, Wang Y, Chen HW, Hsiao PM, Yeh CN, Hsieh CH, et al. A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. Nature Communications. 2021;12(1):1066.	Wrong AI
Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. The Lancet. 2018;392(10162):2388-96.	Wrong population
Choi J, Hui JZ, Spain D, Su YS, Cheng CT, Liao CH. Practical computer vision application to detect hip fractures on pelvic X-rays: A bi-institutional study. Trauma Surgery and Acute Care Open. 2021;6(1):e000705.	Open source AI
Choi JW, Cho YJ, Ha JY, Lee YY, Koh SY, Seo JY, et al. Deep Learning-Assisted Diagnosis of Pediatric Skull Fractures on Plain Radiographs. Korean journal of radiology. 2022;23(3):343-54.	Wrong AI
Choi JW, Cho YJ, Lee S, Lee J, Lee S, Choi YH, et al. Using a Dual-Input Convolutional Neural Network for Automated Detection of Pediatric Supracondylar Fracture on Conventional Radiography. Investigative radiology. 2020;55(2):101-10.	Unclear AI
Curl PK, Jacob AMD, Bresnahan B, Cross NM, Jarvik JG. Cost-Effectiveness of Artificial Intelligence-Based Opportunistic Compression Fracture Screening of Existing Radiographs. Journal of the American College of Radiology : JACR. 2024.	Wrong AI, wrong study design
Dasegowda G, Sato JY, Elton DC, Garza-Frias E, Schultz T, Bridge CP, et al. No code machine learning: validating the approach on use-case for classifying clavicle fractures. Clinical imaging. 2024;112:110207.	Unclear AI
Dupuis, M et al. External validation of an artificial intelligence solution for the detection of elbow fractures and joint	Unclear if standalone

Excluded study	Reason for exclusion
effusions in children, Diagnostic and Interventional Imaging, 105(3), 2024, 104-109.	AI (no response from company clarification)
Dupuis M, Delbos L, Veil R, Adamsbaum C. External validation of a commercially available deep learning algorithm for fracture detection in children. Diagnostic and Interventional Imaging. 2022;103(3):151-9.	Unclear if standalone AI (no response from company clarification)
Elton DC, Dasegowda G, Sato JY, Frias EG, Bridge CP, Mamonov AB, et al. No-code machine learning in radiology: implementation and validation of a platform that allows clinicians to train their own models. medRxiv. 2024.	Unclear AI
Erdas CB. Automated fracture detection in the ulna and radius using deep learning on upper extremity radiographs. Joint diseases and related surgery. 2023;34(3):598-604.	Open source AI
Food US, Administration D. (FDA) USFaDA. Artificial Intelligence and Machine Learning (AI/ML)- Enabled Medical Devices2023 2023.	Wrong publication type
Futurist TM. FDA Approved AI Based Algorithms.	Wrong publication type
Gao Y, Soh NYT, Liu N, Lim G, Ting D, Cheng LT-E, et al. Application of a deep learning algorithm in the detection of hip fractures. iScience. 2023;26(8):107350.	Unclear AI
Ghosh A, Patton D, Bose S, Henry MK, Ouyang M, Huang H, et al. A Patch-Based Deep Learning Approach for Detecting Rib Fractures on Frontal Radiographs in Young Children. Journal of digital imaging. 2023;36(4):1302-13.	Open source AI
Ghosh A, Patton D, Bose S, Ouyang M, Huang H, Sze R, et al. A Patch-based Convolutional Neural Network Approach for the Detection of Rib Fractures on Frontal Radiographs in Young Children. Pediatric Radiology. 2022;52:S48.	Unclear AI
Gipson J, Tang V, Seah J, Kavnoudias H, Zia A, Lee R, et al. Diagnostic accuracy of a commercially available deep-learning algorithm in supine chest radiographs following trauma. The British journal of radiology. 2022;95(1134):20210979.	Wrong AI
Govindarajan A, Govindarajan A, Tanamala S, Chattoraj S, Reddy B, Agrawal R, et al. Role of an Automated Deep Learning Algorithm for Reliable Screening of Abnormality in Chest Radiographs: A Prospective Multicenter Quality Improvement Study. Diagnostics (Basel, Switzerland). 2022;12(11).	Wrong population
Guo J, Mu Y, Xue D, Li H, Chen J, Yan H, et al. Automatic analysis system of calcaneus radiograph: Rotation-invariant landmark detection for calcaneal angle measurement, fracture identification and fracture region segmentation. Computer methods and programs in biomedicine. 2021;206:106124.	Unclear AI
Guo L, Zhou C, Xu J, Huang C, Yu Y, Lu G. Deep Learning for Chest X-ray Diagnosis: Competition Between Radiologists with or Without Artificial Intelligence Assistance. Journal of imaging informatics in medicine. 2024;37(3):922-34.	Unclear AI
Hansen V, Jensen J, Kusk MW, Gerke O, Tromborg HB, Lysdahlgaard S. Deep learning performance compared to healthcare experts in detecting wrist fractures from radiographs: A systematic review and meta-analysis. European journal of radiology. 2024;174:111399.	Wrong study design

Excluded study	Reason for exclusion
Hayashi D, Kompel AJ, Ventre J, Ducarouge A, Nguyen T, Regnard N-E, et al. Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning. Skeletal radiology. 2022;51(11):2129-39.	Standalone AI
NICE. Hip fracture: management. Clinical guideline [CG124]2011 2011-6-22.	Wrong publication type
NICE. Osteoporosis: assessing the risk of fragility fracture. Clinical guideline [CG146]2012 2012-8-8.	Wrong publication type
NICE. Fractures (complex): assessment and management. NICE guideline [NG37]2016 2016-2-17.	Wrong publication type
NICE. Fractures (non-complex): assessment and management. NICE guideline [NG38]2016 2016-2-17.	Wrong publication type
Hendrix N, Hendrix W, van Dijke K, Maresch B, Maas M, Bollen S, et al. Musculoskeletal radiologist-level performance by using deep learning for detection of scaphoid fractures on conventional multi-view radiographs of hand and wrist. European radiology. 2023;33(3):1575-88.	Open source AI
Hendrix N, Scholten E, Vernhout B, Bruijnen S, Maresch B, de Jong M, et al. Development and validation of a convolutional neural network for automated detection of scaphoid fractures on conventional radiographs. Radiology: Artificial Intelligence. 2021;3(4):e200260.	Unclear AI
Hong N, Cho SW, Shin S, Lee S, Jang SA, Roh S, et al. Deep-Learning-Based Detection of Vertebral Fracture and Osteoporosis Using Lateral Spine X-Ray Radiography. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research. 2023;38(6):887-95.	Open source AI
Hoppe BF, Rueckel J, Dikhtyar Y, Heimer M, Fink N, Sabel BO, et al. Implementing Artificial Intelligence for Emergency Radiology Impacts Physicians' Knowledge and Perception: A Prospective Pre- and Post-Analysis. Investigative Radiology. 2024;59(5).	De-prioritised for inclusion as a low priority outcome (clinician acceptability)
Huhtanen JT, Nyman M, Doncenco D, Hamedian M, Kawalya D, Salminen L, et al. Deep learning accurately classifies elbow joint effusion in adult and pediatric radiographs. Scientific reports. 2022;12(1):11803.	Unclear AI
Oppenheimer. An overview of the performance of AI in fracture detection in lumbar and thoracic spine radiographs on a per vertebra basis. Skeletal radiology. 2024;53(8):1563-71.	Standalone AI
Janisch M, Apfaltrer G, Hrzic F, Castellani C, Mittl B, Singer G, et al. Pediatric radius torus fractures in x-rays-how computer vision could render lateral projections obsolete. Frontiers in Pediatrics. 2022;10:1005099.	Unclear AI
Jeong TS, Yee GT, Kim KG, Kim YJ, Lee SG, Kim WK. Automatically Diagnosing Skull Fractures Using an Object Detection Method and Deep Learning Algorithm in Plain Radiography Images. Journal of Korean Neurosurgical Society. 2022;66(1):53-62.	Unclear AI

Excluded study	Reason for exclusion
Jimenez-Sanchez A, Kazi A, Albarqouni S, Kirchhoff C, Biberthaler P, Navab N, et al. Precise proximal femur fracture classification for interactive training and surgical planning. International journal of computer assisted radiology and surgery. 2020;15(5):847-57.	Unclear AI
Jones CM, Danaher L, Milne MR, Tang C, Seah J, Oakden-Rayner L, et al. Assessment of the effect of a comprehensive chest radiograph deep learning model on radiologist reports and patient outcomes: a real-world observational study. BMJ Open. 2021;11(12):e052902.	Wrong population
Jones RM, Sharma A, Hotchkiss R, Sperling JW, Hamburger J, Ledig C, et al. Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. npj Digital Medicine. 2020;3(1):144.	Unclear AI
Jonsson B, Kanis J, Dawson A, Oden A, Johnell O. Effect and offset of effect of treatments for hip fracture on health outcomes. Osteoporos Int. 1999;10(3):193-9.	Wrong population
Kandel I, Castelli M. Improving convolutional neural networks performance for image classification using test time augmentation: a case study using MURA dataset. Health information science and systems. 2021;9(1):33.	Unclear AI
Kaviani P, Kalra MK, Digumarthy SR, Gupta RV, Dasegowda G, Jagirdar A, et al. Frequency of Missed Findings on Chest Radiographs (CXRs) in an International, Multicenter Study: Application of AI to Reduce Missed Findings. Diagnostics (Basel, Switzerland). 2022;12(10).	Wrong AI
Kaya O, Tasci B. A Pyramid Deep Feature Extraction Model for the Automatic Classification of Upper Extremity Fractures. Diagnostics. 2023;13(21):3317.	Unclear AI
Kekatpure A, Deshpande S, Srivastava S. Development of a diagnostic support system for distal humerus fracture using artificial intelligence. International Orthopaedics. 2024.	Open source AI
Kim MW, Jung J, Park SJ, Park YS, Yi JH, Yang WS, et al. Application of convolutional neural networks for distal radio-ulnar fracture detection on plain radiographs in the emergency room. Clinical and Experimental Emergency Medicine. 2021;8(2):120-7.	Unclear AI
Kim S, Rebmann P, Tran PH, Kellner E, Reisert M, Steybe D, et al. Multiclass datasets expand neural network utility: an example on ankle radiographs. International journal of computer assisted radiology and surgery. 2023;18(5):819-26.	Open source AI
Kim T, Goh TS, Lee JS, Lee JH, Kim H, Jung ID. Transfer learning-based ensemble convolutional neural network for accelerated diagnosis of foot fractures. Physical and engineering sciences in medicine. 2023;46(1):265-77.	Wrong AI
Kitamura G, Chung CY, Moore BE, 2nd. Ankle Fracture Detection Utilizing a Convolutional Neural Network Ensemble Implemented with a Small Sample, De Novo Training, and Multiview Incorporation. Journal of digital imaging. 2019;32(4):672-7.	Wrong AI
Kong SH, Kim JH, Lee JW, Bae BU, Sung JK, Jung KH, et al. Development of Spine Radiography-Based Fracture Prediction Model Using Convolutional Neural Network. Journal of Bone and Mineral Research. 2020;35:207.	Unclear AI
Koska OI, Cilengir AH, Uluc ME, Yucel A, Tosun O. All-star approach to a small medical imaging dataset: combined deep,	Unclear AI

Excluded study	Reason for exclusion
transfer, and classical machine learning approaches for the determination of radial head fractures. Acta radiologica (Stockholm, Sweden : 1987). 2023;64(4):1476-83.	
Krogue JD, Cheng KV, Hwang KM, Toogood P, Meinberg EG, Geiger EJ, et al. Automatic Hip Fracture Identification and Functional Subclassification with Deep Learning. Radiology Artificial intelligence. 2020;2(2):e190023.	Unclear AI
Kruger N, Abramowitz S, Nitschke G. A197: Machine learning in diagnosing cervical spine injuries. Global Spine Journal. 2022;12(3):113S.	Unclear AI
Langerhuizen DWG, Bulstra AEJ, Janssen SJ, Ring D, Kerkhoffs GMMJ, Jaarsma RL, et al. Is Deep Learning On Par with Human Observers for Detection of Radiographically Visible and Occult Fractures of the Scaphoid? Clinical orthopaedics and related research. 2020;478(11):2653-9.	Unclear AI
Langerhuizen DWG, Janssen SJ, Mallee WH, van den Bekerom MPJ, Ring D, Kerkhoffs G, et al. What Are the Applications and Limitations of Artificial Intelligence for Fracture Detection and Classification in Orthopaedic Trauma Imaging? A Systematic Review. Clin Orthop Relat Res. 2019;477(11):2482-91.	Wrong study design
Lassalle L, Regnard NE, Ventre J, Marty V, Clovis L, Zhang Z, et al. Automated weight-bearing foot measurements using an artificial intelligence-based software. Skeletal Radiology. 2024.	Wrong population
OneLaudos, Radiobotics. RBfracture retrospective pilot study.	Standalone AI
Lee KC, Choi IC, Kang CH, Ahn KS, Yoon H, Lee JJ, et al. Clinical Validation of an Artificial Intelligence Model for Detecting Distal Radius, Ulnar Styloid, and Scaphoid Fractures on Conventional Wrist Radiographs. Diagnostics. 2023;13(9):1657.	Wrong AI
Lee S, Kim KG, Kim YJ, Jeon JS, Lee GP, Kim K-C, et al. Automatic Segmentation and Radiologic Measurement of Distal Radius Fractures Using Deep Learning. Clinics in orthopedic surgery. 2024;16(1):113-24.	Unclear AI
Li W, Chui TKH, Tiu KL, Lee KB, Lai KC, Li KK. AUTOMATED DETECTION AND LOCALIZATION OF VERTEBRAL COMPRESSION FRACTURES FOR EARLY IDENTIFICATION AND INITIATION OF MANAGEMENT. Aging Clinical and Experimental Research. 2023;35:S289-S90.	Unclear AI
Liu X, Li K, Luo Y, Bai S, Wu J, Chen W, et al. A deep-learning model for identifying fresh vertebral compression fractures on digital radiography. European Radiology. 2022;32(3):1496-505.	Wrong AI
Lu X, Chang EY, Du J, Yan A, McAuley J, Gentili A, et al. Robust Multi-View Fracture Detection in the Presence of Other Abnormalities Using HAMIL-Net. Military medicine. 2023;188(6):590-7.	Wrong AI
Luo J, Kitamura G, Arefan D, Doganay E, Panigrahy A, Wu S. Knowledge-Guided Multiview Deep Curriculum Learning for Elbow Fracture Classification. Machine learning in medical imaging MLMI (Workshop). 2021;12966:555-64.	Open source AI
Lysdahlgaard S. Utilizing heat maps as explainable artificial intelligence for detecting abnormalities on wrist and elbow radiographs. Radiography (London, England : 1995). 2023;29(6):1132-8.	Wrong AI
Ma Y, Luo Y. Bone fracture detection through the two-stage system of Crack-Sensitive Convolutional Neural Network. Informatics in Medicine Unlocked. 2021;22:100452.	Wrong AI

Excluded study	Reason for exclusion
Maarek R, Hermann AL, Kamoun A, Marchi A, Khelifi R, Collin M, et al. Assessment of an AI aid in detection of paediatric appendicular skeletal fractures by senior and junior radiologists. Insights into Imaging. 2022;14:36-7.	Unclear AI
Mert S, Stoerzer P, Brauer J, Fuchs B, Haas-Lützenberger EM, Demmer W, et al. Diagnostic power of ChatGPT 4 in distal radius fracture detection through wrist radiographs. Arch Orthop Trauma Surg. 2024;144(5):2461-7.	Standalone AI
Michelson JD. Using decision analysis to assess comparative clinical efficacy of surgical treatment of unstable ankle fractures. J Orthop Trauma. 2013;27(11):642-8.	Incorrect population
Min H, Rabi Y, Wadhawan A, Bourgeat P, Dowling J, White J, et al. Automatic classification of distal radius fracture using a two-stage ensemble deep learning framework. Physical and engineering sciences in medicine. 2023;46(2):877-86.	Open source AI
Moghaddam SA, Yadekar M, Vahdat AS, Esmaeili F. AUTOMATIC DETECTION OF MANDIBULAR FRACTURES ON PANORAMIC RADIOGRAPHS USING THE CONVOLUTIONAL NEURAL NETWORK. Russian Electronic Journal of Radiology. 2023;13(3):5-13.	Unclear AI
Mosquera C, Diaz FN, Binder F, Rabellino JM, Benitez SE, Beresnak AD, et al. Chest x-ray automated triage: A semiologic approach designed for clinical implementation, exploiting different types of labels through a combination of four Deep Learning architectures. Computer Methods and Programs in Biomedicine. 2021;206:106130.	Unclear AI
Murphy EA, Ehrhardt B, Gregson CL, von Arx OA, Hartley A, Whitehouse MR, et al. Machine learning outperforms clinical experts in classification of hip fractures. Scientific reports. 2022;12(1):2058.	Open source AI
Naguib SM, Hamza HM, Hosny KM, Saleh MK, Kassem MA. Classification of Cervical Spine Fracture and Dislocation Using Refined Pre-Trained Deep Model and Saliency Map. Diagnostics (Basel, Switzerland). 2023;13(7).	Unclear AI
Nagy E, Marterer R, Hrzic F, Sorantin E, Tschauner S. Learning rate of students detecting and annotating pediatric wrist fractures in supervised artificial intelligence dataset preparations. PloS one. 2022;17(10):e0276503.	Wrong AI
Nam Y, Choi Y, Kang J, Seo M, Heo SJ, Lee MK. Diagnosis of nasal bone fractures on plain radiographs via convolutional neural networks. Scientific reports. 2022;12(1):21510.	Open source AI
Nassiri F, Davy G, Wright C, Ingrand P. Artificial Intelligence Software: Can it Improve the MSK Decision Making of Radiographers? A Pilot Study in France. Journal of Medical Imaging and Radiation Sciences. 2022;53(4):S11.	Unclear AI
Ngan E, Nguyen HT, Cano M, Jones L, Annapragada A, Kan JH, et al. Children's Long Bones Fracture Subtype Identification and Localization on Plain Radiographs of Using Neural Network. Pediatric Radiology. 2023;53:S58.	Open source AI
Nguyen NH, Nguyen HQ, Nguyen NT, Nguyen TV, Pham HH, Nguyen TNM. Deployment and validation of an AI system for detecting abnormal chest radiographs in clinical settings. Frontiers in Digital Health. 2022;4:890759.	Wrong AI
Nishiyama M, Ishibashi K, Ariji Y, Fukuda M, Nishiyama W, Umemura M, et al. Performance of deep learning models constructed using panoramic radiographs from two hospitals to diagnose fractures of the mandibular condyle. Dento maxillo facial radiology. 2021;50(7):20200611.	Unclear AI
Nowroozi A, Salehi MA, Shobeiri P, Agahi S, Momtazmanesh S, Kaviani P, et al. Artificial intelligence diagnostic accuracy	Wrong study design

Excluded study	Reason for exclusion
in fracture detection from plain radiographs and comparing it with clinicians: a systematic review and meta-analysis. Clinical radiology. 2024.	
Oakden-Rayner L, Gale W, Bonham TA, Lungren MP, Carneiro G, Bradley AP, et al. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. The Lancet Digital health. 2022;4(5):e351-e8.	Unclear AI
Oka K, Shiode R, Yoshii Y, Tanaka H, Iwahashi T, Murase T. Artificial intelligence to diagnosis distal radius fracture using biplane plain X-rays. Journal of orthopaedic surgery and research. 2021;16(1):694.	Unclear AI
Olczak J, Emilson F, Razavian A, Antonsson T, Stark A, Gordon M. Ankle fracture classification using deep learning: automating detailed AO Foundation/Orthopedic Trauma Association (AO/OTA) 2018 malleolar fracture identification reaches a high degree of correct classification. Acta orthopaedica. 2021;92(1):102-8.	Wrong AI
Ozkaya E, Topal FE, Bulut T, Gursoy M, Ozuysal M, Karakaya Z. Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography. European journal of trauma and emergency surgery : official publication of the European Trauma Society. 2022;48(1):585-92.	Unclear AI
Park JY, Lee SH, Kim YJ, Lee GJ, Kim KG. Machine learning model based on radiomics features for AO/OTA classification of pelvic fractures on pelvic radiographs. PLoS ONE. 2024;19(5):e0304350.	Open source AI
Parpaleix A, Parsy C, Codari M, Mejdoubi M. Added value of artificial intelligence in traumatic radiographic findings detection in emergency settings. Insights into Imaging. 2022;14(Supplement 4):202.	Unclear if standalone AI
Pauling C, Thomas K, Evans E, Laidlow-Singh H, Garbera D, Fernando R, et al. ACCURACY OF ARTIFICIAL INTELLIGENCE FOR FRACTURE DETECTION IN OSTEOGENESIS IMPERFECTA. BMJ Paediatrics Open. 2023;7:A12-A3.	Unclear AI
Pourchot A, Bailly K, Ducarouge A, Sigaud O, Ieee, Elect Engineers IE, et al. NEURAL ARCHITECTURE SEARCH FOR FRACTURE CLASSIFICATION. Bordeaux, FRANCE2022 2022-10-16. 3226-30 p.	Wrong outcome
Pouvreau M, Delmas J, Chateil JF. EVALUATION OF ARTIFICIAL INTELLIGENCE IN FRACTURE DETECTION IN CHILDREN: PRELIMINARY RESULTS. Pediatric Radiology. 2022;52:S156.	Standalone AI
Radiobotics. S4 confidential company submission.	Wrong intervention
Radiobotics. S5 confidential company submission.	Wrong outcome
Radiobotics. S6 confidential company submission.	Wrong intervention
Radiobotics. S10 confidential company submission.	Standalone AI
Radiobotics. S12 confidential company submission.	Abstract only
Rashid T, Zia MS, Najam Ur R, Meraj T, Rauf HT, Kadry S. A Minority Class Balanced Approach Using the DCNN-LSTM Method to Detect Human Wrist Fracture. Life (Basel, Switzerland). 2023;13(1).	Unclear AI
Rayan JC, Reddy N, Herman Kan J, Zhang W, Annapragada A. Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making. Radiology: Artificial Intelligence.	Unclear AI

Excluded study	Reason for exclusion
2019;1(1):e180015.	
Rayscape. Rayscape Medical Whitepaper2022 2022.	Wrong AI
Reichert G, Bellamine A, Fontaine M, Naipeanu B, Altar A, Mejean E, et al. How Can a Deep Learning Algorithm Improve Fracture Detection on X-rays in the Emergency Room? Journal of Imaging [Internet]. 2021; 7(7).	Unclear if standalone AI
Regnard NE, Lanseur B, Lassalle L, Lambert A, Dallaudiere B, Feydy A. Performances of a deep learning algorithm for the detection of fracture, dislocation, elbow joint effusion, focal bone lesions on trauma x-rays. Insights into Imaging. 2022;14:227.	Standalone AI
Regnard NE, Lanseur B, Ventre J, Ducarouge A, Clovis L, Lassalle L, et al. Assessment of performances of a deep learning algorithm for the detection of limbs and pelvic fractures, dislocations, focal bone lesions, and elbow effusions on trauma X-rays. European Journal of Radiology. 2022;154:110447.	Standalone AI
Rezaei Z, Eslami B, Komleh HE, Jahromi KD. Abnormality detection in musculoskeletal radiographs by densenet and inception-v3. Iranian Journal of Radiology. 2019;16:S14-S5.	Unclear AI
Rosa F, Buccicardi D, Romano A, Borda F, D'Auria MC, Gastaldo A. Artificial intelligence and pelvic fracture diagnosis on X-rays: a preliminary study on performance, workflow integration and radiologists' feedback assessment in a spoke emergency hospital. European journal of radiology open. 2023;11:100504.	Standalone AI
Rosenberg GS, Cina A, Schiro GR, Giorgi PD, Gueorguiev B, Alini M, et al. Artificial Intelligence Accurately Detects Traumatic Thoracolumbar Fractures on Sagittal Radiographs. Swiss Medical Weekly. 2023;153:16S.	Unclear AI
Ruitenbeek H, Egnell L, Ziegeler K, Brejnebol MW, Nybing JU, Lensskjold A, et al. Protocol for the AutoRayValid-RBfracture Study: Evaluating the efficacy of an AI fracture detection system. medRxiv. 2023.	Wrong publication type
Russe MF, Rebmann P, Tran PH, Kellner E, Reisert M, Bamberg F, et al. AI-based X-ray fracture analysis of the distal radius: accuracy between representative classification, detection and segmentation deep learning models for clinical practice. BMJ open. 2024;14(1):e076954.	Standalone AI
Rutledge M, Yap M, Chai K. Plain film mandibular fracture detection using machine learning - Model development. Advances in Oral and Maxillofacial Surgery. 2023;11:100436.	Wrong AI
Sanchez M, Alford K, Krishna V, Huynh TM, Nguyen CDT, Lungren MP, et al. AI-clinician collaboration via disagreement prediction: A decision pipeline and retrospective analysis of real-world radiologist-AI interactions. Cell Reports Medicine. 2023;4(10):101207.	Wrong AI
Sato Y, Takegami Y, Asamoto T, Ono Y, Hidetoshi T, Goto R, et al. Artificial intelligence improves the accuracy of residents in the diagnosis of hip fractures: a multicenter study. BMC musculoskeletal disorders. 2021;22(1):407.	Unclear AI
Seah JCY, Tang CHM, Buchlak QD, Holt XG, Wardman JB, Aimoldin A, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. Lancet Digit Health. 2021;3(8):e496-e506.	Wrong AI

Excluded study	Reason for exclusion
Shahnavazi M, Mohamadrahimi H. The application of artificial neural networks in the detection of mandibular fractures using panoramic radiography. Dental Research Journal. 2023;20(1):27.	Unclear AI
Shaik A, Larsen K, Lane NE, Zhao C, Su K-J, Keyak JH, et al. A Staged Approach using Machine Learning and Uncertainty Quantification to Predict the Risk of Hip Fracture. ArXiv. 2024.	Wrong AI, wrong population
Shen L, Gao C, Hu S, Kang D, Zhang Z, Xia D, et al. Using Artificial Intelligence to Diagnose Osteoporotic Vertebral Fractures on Plain Radiographs. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research. 2023;38(9):1278-87.	Wrong AI
Shim JH, Kim WS, Kim KG, Yee GT, Kim YJ, Jeong TS. Automated Segmentation and Diagnostic Measurement for the Evaluation of Cervical Spine Injuries Using X-Rays. Journal of imaging informatics in medicine. 2024.	Open source AI
SIGN, Scotland HI. Management of osteoporosis and the prevention of fragility fractures2021 2021.	Wrong publication type
Silberstein J, Sun Z. A Novel AI Tool for Automated Detection Of Osteoporotic Vertebral Fractures On Routine Chest Radiographs. Australasian Medical Journal. 2023;16(3):550-2.	Wrong AI
Silberstein J, Wee C, Gupta A, Seymour H, Ghotra SS, Sa Dos Reis C, et al. Artificial Intelligence-Assisted Detection of Osteoporotic Vertebral Fractures on Lateral Chest Radiographs in Post-Menopausal Women. Journal of clinical medicine. 2023;12(24).	Wrong AI
Singh D, Nagaraj S, Mashouri P, Drysdale E, Fischer J, Goldenberg A, et al. Assessment of Machine Learning-Based Medical Directives to Expedite Care in Pediatric Emergency Medicine. JAMA Network Open. 2022;5(3):e222599.	Wrong AI
Soong C, Lin FH, Yang RS, Yang TH, Chen HY. APPLICATION OF DEEP LEARNING ALGORITHM TO DETECT AND VISUALIZE VERTEBRAL FRACTURES ON PLAIN FRONTAL RADIOGRAPHS. Aging Clinical and Experimental Research. 2023;35:S137-S8.	Unclear AI
Stanborough RO, Garner HW. Beyond the AJR: Validation and Algorithmic Audit of a Deep Learning System to Detect Hip Fractures Radiographically. AJR American journal of roentgenology. 2023;220(1):150.	Wrong publication type
Sun H, Wang X, Li Z, Liu A, Xu S, Jiang Q, et al. Automated Rib Fracture Detection on Chest X-Ray Using Contrastive Learning. Journal of digital imaging. 2023;36(5):2138-47.	Unclear AI
Tan H, Xu H, Yu N, Yu Y, Duan H, Fan Q, et al. The value of deep learning-based computer aided diagnostic system in improving diagnostic performance of rib fractures in acute blunt trauma. BMC medical imaging. 2023;23(1):55.	Wrong population
Tanamala S CS, Maniparambil M, Rao P, Biviji M. Clinical Context Improves the Performance of AI models for Cranial Fracture Detection2019 2019.	Wrong population
Tobler P, Cyriac J, Kovacs BK, Hofmann V, Sexauer R, Paciolla F, et al. AI-based detection and classification of distal radius fractures using low-effort data labeling: evaluation of applicability and effect of training set size. European radiology. 2021;31(9):6816-24.	Unclear AI
Tu E, Burkow J, Tsai A, Junewick J, Perez FA, Otjen J, et al. Near-pair patch generative adversarial network for data	Unclear AI

Excluded study	Reason for exclusion
augmentation of focal pathology object detection models. Journal of medical imaging (Bellingham, Wash). 2024;11(3):034505.	
Twinprai N, Boonrod A, Boonrod A, Chindaprasirt J, Sirithanaphol W, Chindaprasirt P, et al. Artificial intelligence (AI) vs. human in hip fracture detection. Heliyon. 2022;8(11):e11266.	Wrong AI
Ureten K, Sevinc HF, Igdeli U, Onay A, Maras Y. Use of deep learning methods for hand fracture detection from plain hand radiographs. Duz el radyografilerinden el kiriklarinin tespiti icin derin ogrenme yontemlerinin kullanilmasi. 2022;28(2):196-201.	Wrong AI
van Leeuwen KG, Schalekamp S, Rutten M, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol. 2021;31(6):3797-804.	Wrong study design
Ventre J, Regnard NE, Lanseur B, Lassalle L, Lambert A, Dallaudiere B, et al. Performances of a deep learning algorithm for the detection of fractures, dislocations, elbow joint effusions, focal bone lesions on trauma X-rays. Insights into Imaging. 2022;13:9.	Standalone AI
Wadhawan A, Min H. Fracture classification using deep learning algorithms. Journal of Medical Imaging and Radiation Oncology. 2022;66:11.	Wrong AI
Wang Y, Li Y, Lin G, Zhang Q, Zhong J, Zhang Y, et al. Lower-extremity fatigue fracture detection and grading based on deep learning models of radiographs. European radiology. 2023;33(1):555-65.	Unclear AI
Warin K, Limprasert W, Suebnukarn S, Inglam S, Jantana P, Vicharueang S. Assessment of deep convolutional neural network models for mandibular fracture detection in panoramic radiographs. International journal of oral and maxillofacial surgery. 2022;51(11):1488-94.	Unclear AI
Wee C, Sun Z. AI-assisted automated detection: The future tool of detecting missed osteoporotic vertebral fractures on chest radiographs. Journal of Medical Radiation Sciences. 2023;70:53.	Wrong AI
Wei D, Wu Q, Wang X, Tian M, Li B. Accurate Instance Segmentation in Pediatric Elbow Radiographs. Sensors (Basel, Switzerland). 2021;21(23).	Wrong AI
Xiao BH, Zhu MSY, Du EZ, Liu WH, Ma JB, Huang H, et al. A software program for automated compressive vertebral fracture detection on elderly women's lateral chest radiograph: Ofeye 1.0. Quantitative Imaging in Medicine and Surgery. 2022;12(8):4259-71.	Wrong AI
Xie Y, Li X, Chen F, Wen R, Jing Y, Liu C, et al. Artificial intelligence diagnostic model for multi-site fracture X-ray images of extremities based on deep convolutional neural networks. Quantitative imaging in medicine and surgery. 2024;14(2):1930-43.	Unclear AI
Xu F, Xiong Y, Ye G, Liang Y, Guo W, Deng Q, et al. Deep learning-based artificial intelligence model for classification of vertebral compression fractures: A multicenter diagnostic study. Frontiers in endocrinology. 2023;14:1025749.	Open source AI
Yang TH, Horng MH, Li RS, Sun YN. Scaphoid Fracture Detection by Using Convolutional Neural Network. Diagnostics.	Wrong AI

Excluded study	Reason for exclusion
2022;12(4):895.	
Yari A, Fasih P, Hosseini Hooshiar M, Goodarzi A, Fattahi SF. Detection and classification of mandibular fractures in panoramic radiography using artificial intelligence. Dento maxillo facial radiology. 2024.	Unclear AI
Yildiz Potter I, Yeritsyan D, Mahar S, Kheir N, Putman M, Rodriguez EK, et al. Proximal femur fracture detection on plain radiography via feature pyramid networks. Scientific reports. 2024;14(1):12046.	Wrong AI
Yoon AP, Chung WT, Wang C-W, Kuo C-F, Lin C, Chung KC. Can a Deep Learning Algorithm Improve Detection of Occult Scaphoid Fractures in Plain Radiographs? A Clinical Validation Study. Clinical orthopaedics and related research. 2023;481(9):1828-35.	Unclear AI
Yoon AP, Lee Y-L, Kane RL, Kuo C-F, Lin C, Chung KC. Development and Validation of a Deep Learning Model Using Convolutional Neural Networks to Identify Scaphoid Fractures in Radiographs. JAMA network open. 2021;4(5):e216096.	Unclear AI
Yu AC, Mohajer B, Eng J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. Radiol Artif Intell. 2022;4(3):e210064.	Wrong study design
Yu JS, Yu SM, Erdal BS, Demirer M, Gupta V, Bigelow M, et al. Detection and localisation of hip fractures on anteroposterior radiographs with artificial intelligence: proof of concept. Clinical radiology. 2020;75(3):237.e1-.e9.	Unclear AI
Zech JR, Carotenuto G, Igbinoba Z, Tran CV, Insley E, Baccarella A, et al. Detecting pediatric wrist fractures using deep-learning-based object detection. Pediatric radiology. 2023;53(6):1125-34.	Unclear AI
Zech JR, Ezuma CO, Patel S, Edwards CR, Posner R, Hannon E, et al. Artificial intelligence improves resident detection of pediatric and young adult upper extremity fractures. Skeletal Radiology. 2024.	Wrong AI
Zech JR, Jaramillo D, Altosaar J, Popkin CA, Wong TT. Artificial intelligence to identify fractures on pediatric and young adult upper extremity radiographs. Pediatric radiology. 2023;53(12):2386-97.	Wrong AI
Zhang H, Xu R, Guo X, Zhou D, Xu T, Zhong X, et al. Deep learning-based automated high-accuracy location and identification of fresh vertebral compression fractures from spinal radiographs: a multicenter cohort study. Frontiers in bioengineering and biotechnology. 2024;12:1397003.	Unclear AI
Zhang J, Xia L, Liu J, Niu X, Tang J, Xia J, et al. Exploring deep learning radiomics for classifying osteoporotic vertebral fractures in X-ray images. Frontiers in endocrinology. 2024;15:1370838.	Unclear AI

Appendix D – Additional Scenario Analyses

This appendix contains additional analysis requested by NICE and specialist committee members to provide further information prior to the appraisal committee meeting.

Some companies stated that their software was associated with setup costs. Feedback from an SCM suggested that there are also costs incurred within the NHS to support set up of a new technology, specifically NHS IT time and fees from PACS providers to ensure the new technology works correctly and does not cause any issues with the existing radiology workflow. These may be up to £50,000 per site. The EAG therefore presented an analysis adding a notional one-off set up fee of £50,000, and conducted a threshold analysis stating the maximum set-up fee for AI detection to be cost-effective. This is equivalent to the incremental net monetary benefit at a site (rather than individual patient) level.

Similarly, the EAG assumed a notional five-year life for the software. That is, the fixed costs were apportioned on a per-scan basis over five years. The EAG conducted a scenario here assuming they are apportioned over two years.

Method

Aggregate figures reported in Table 46 correspond to one year's use of the algorithms. These are multiplied by five to approximate a five year lifespan of the software. The INHB and INMB are calculated including a £50,000 one-off set up cost.

Results

The results are broadly insensitive to a £50,000 setup cost: most of the interventions are associated with a cost saving in excess of this when considered over a 5 year lifetime of the algorithm (Table 51). The EAG noted that even over a one year lifespan of the algorithm, a £50,000 will still not offset the savings associated use of most of the algorithms. The maximum cost of installation for the algorithms to still represent value for money is equal to £50,000 plus the INMB at the chosen threshold. For example, the maximum cost for installation, assuming a threshold of £20,000 / QALY for BoneView is £4.240m (£4.190m + £50k). The EAG noted that this is likely far above a plausible estimate of the cost for this.

Table 51: Population level results: scenario analysis

interventions	Setup cost	Cost vs unassisted (£000s)	Total (£000s)	QALYs vs Unassisted	INHB20k	INHB30k	INMB20K (£000s)	INMB30K (£000k)
BoneView	50,000	-4,107K	-4,057K	6.670	209.51	141.89	4,190K	4,257K
Rayvolve	50,000	19,796K	19,846K	6.670	-985.65	-654.88	-19,713K	-19,646K
RBfracture	50,000	[redacted]	[redacted]	0	[redacted]	[redacted]	[redacted]	[redacted]
TechCare Alert	50,000	[redacted]	[redacted]	6.670	[redacted]	[redacted]	[redacted]	[redacted]
Unassisted		0		0	0	0	0	0

Abbreviations: INHB20k/INHB30k, incremental health benefit at willingness to pay threshold of £20/30k; INMB20k/INMB£30k incremental net monetary benefit at each threshold; QALY, Quality Adjusted Life Year

NICE National Institute for Health and Care Excellence

Artificial intelligence software to help detect fractures on X-rays in urgent care

External Assessment Report (EAR) and economic model - Comments

External Assessment Report - Comments

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
Gleamer	1	113	8.3.10.	The scenario analysis presented in Table 37 is not reproducible due to volume of literature and differing datasets, leading to a bias towards those technologies that have results from only 1 study.	The EAG selected lowest and highest plausible values for the optimistic and pessimistic scenarios from the literature available. They represent the highest and lowest figures reported in Tables 11 to 16 of the report. The EAG agrees that this may bias for or against technologies that only have one study. The EAG stresses in several places that comparisons between technologies should preferably be avoided. Comment added to P117: "For example, the optimistic and pessimistic scenarios may not fully capture uncertainty and therefore may bias in favour or against technologies with only one source study."
Gleamer	2	113	8.3.10.	Algorithm versioning is an important consideration. From the study, which version of the algorithm was used and are the performances up to date based on the latest version?	The sensitivity and specificity of each algorithm was obtained from the relevant source studies and thus the version that was used in those is the implied version in the model. We note on P31 that versioning was not reported throughout the evidence base, and that this is an important reporting consideration for future studies. No edit made to report.
Gleamer	3	113	8.3.10.	Why are the unassisted optimistic and pessimistic scenarios reversed?	This was to provide a most optimistic and pessimistic scenario for the algorithms vs unassisted, rather than for all diagnosis together. No edit made to report

Artificial intelligence software to help detect fractures on X-rays in urgent care

External Assessment Report (EAR) and economic model - Comments