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

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]
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

Correspondence to Caroline Farmer, c.farmer@exeter.ac.uk

Date completed 20/06/24

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.

Source of funding This report was commissioned by the NIHR Evidence Synthesis Programme as project number NIHR136024

Acknowledgments The authors acknowledge the administrative support provided by Mrs Sue Whiffin and Ms Jenny Lowe (both PenTAG).

This report should be referenced as follows: Farmer, Coelho, Muthukumar, Robinson, Meertens, Ukoumunne, Melendez-Torres, Wilson. Artificial intelligence software to help detect fractures in urgent care: Assessment protocol. Peninsula Technology Assessment Group (PenTAG), 2024.

The views expressed in this report are those of the authors and not necessarily those of the NIHR Evidence Synthesis Programme. Any errors are the responsibility of the authors. Any errors are the responsibility of the authors.

Table of Contents

List of Tables

List of Figures

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). Clinicians initially review X-rays to decide on further imaging and treatment, with radiologists or radiographers subsequently confirming the diagnosis. While some fractures are easily visible on X-rays, others can be subtle or located in difficult-to-spot areas, leading to potential missed diagnoses. Missed fractures, though rare, can result in delayed healing, more severe injuries, and increased costs for the NHS. Occasionally, patients are treated as if they have a fracture even if one is not detected on the X-ray, as a precaution to avoid complications from undiagnosed fractures.

Artificial Intelligence (AI) technologies have been developed to assist in identifying fractures on X-rays. NICE is investigating the potential benefits of implementing AI in NHS settings to reduce missed fractures and unnecessary cautionary treatments. AI's effectiveness might vary with the clinician's experience, with greater benefits when used by less experienced staff. However, there are concerns that AI might also identify fractures that do not require treatment, adding unnecessary costs. According to NHS standards, AI would be used as an aid but would not replace clinicians in diagnosing fractures.

NICE has commissioned an Early Value Assessment (EVA) to evaluate licensed AI technologies for fracture detection in urgent care. This assessment, conducted by PenTAG, includes a comprehensive review of available evidence and confidential data from AI companies. It will evaluate AI's accuracy in aiding clinicians compared to clinician-only diagnoses, its impact on patient health, effects on NHS services, and whether the costs of AI are appropriate to the benefits it offers. The assessment will also identify future research recommendations to optimise AI use in the NHS.

1. INTRODUCTION AND BACKGROUND

1.1. Introduction

This assessment protocol outlines what the external assessment group (EAG) will do during the assessment. This protocol was produced in response to the scope for this assessment[1].

1.2. Appraisal decision problem

Table 1 summarises the decision problem to be addressed in this assessment. Further detail on each item can be found in the published scope and the following sections.

Table 1: Summary table of the decision problem

Item	Description	EAG comment
Population	People presenting to urgent care (the emergency department [ED], minor injury unit [MIU] or urgent treatment centre [UTC]) with a suspected fracture for which X-ray is requested.	The EAG considered that the technology would preferentially be used across all suspected fractures in urgent care (rather than only used for certain suspected fracture types or patient groups). However, the EAG was aware that some of the technologies are not suitable for use for some types of fractures, and therefore the use of these technologies would require targeting towards appropriate groups. The EAG noted that the NICE scope also included suspected dislocations that are assessed using x-ray. However, healed fractures (which may be used as an indication of past abuse) were not included in the assessment. Stakeholders to the assessment also highlighted specific fractures that would not typically be assessed using X-ray and therefore would not be relevant for the assessment.
Subgroups	Depending on the availability of evidence, the following subpopulations may be included: • Children and young people (0 to 16 years of age) • Older people • People who are clinically frail • People with conditions affecting bone health (for example, osteoporosis and osteogenesis imperfecta) Depending on the availability of evidence, the following fracture site subgroups may be included: • Hip • Hand (including wrist) • Foot (including ankle)	The EAG will include data for these subpopulations if sufficient evidence is available. Contrary to the scope, the EAG will not consider evidence in an older person subgroup. This is because the EAG received advice that age is not a reliable predictor of a change in outcome risk, and that indicators of frailty or the presence of conditions affecting bone health were better considerations. These subgroups were therefore considered to be most useful to the assessment within the timeframe of the EVA. The EAG considered that fracture site subgroups including the hip, hand (including wrist) and foot (including ankle) were to be considered in adults only. Overall, the EAG considered that the fracture site subgroups selected were those where stakeholders considered that the technology may be most useful. This means that there are no

Item	Description	EAG comment
	• Fractures including the growth plate (Salter-Harris) in children • Fractures of the elbow in children	subpopulations of the assessment where the technology would be considered to have little or no benefit.
Interventions	AI used as a decision aid for X-ray image interpretation and fracture assessment prior to radiology review, using any of the following software/platforms: • BoneView (Gleamer) • Rayvolve (AZmed) • RBfracture (Radiobotics) • qMSK (Qure.ai) • TechCare alert (Milvue)	The EAG considered that evidence may be available that evaluates technologies for detecting fractures independently from clinical review and interpretation. As explained in the NICE scope, regulations require for a trained person to interpret X-rays and this would therefore not represent the proposed clinical pathway, and such evidence will be excluded. Where feasible within the timeline of the assessment, the EAG will compile a record of all studies excluded from the assessment for this reason. The EAG considered the importance of evaluating AI technologies as a decision aid for any relevant health professional (e.g. specialist nurses, junior doctors). Given the reality of clinical practice, such evidence will be included where available.
Comparators	ED clinician or healthcare professional interpretation of X-ray radiograph without AI assistance. Reference standard or ground truth based on consultant radiologist or reporting radiographer interpretation and report.	The EAG will not consider outcomes associated with other AI technologies not listed on the scope, including open-source AI technologies (given that these are unlicenced). The EAG considered that the outcomes of using AI technology may vary according to the experience of staff reading and interpreting the findings within urgent care. In practice, there is variation in the staff that read X-rays within urgent care. The EAG therefore considered that comparisons of outcomes of a technology between different staff grades may be useful for understanding the potential role and value of AI technologies for detecting fractures in the NHS. However, although different healthcare professionals, at different grades, may evaluate X-rays in clinical practice, the EAG considered that the reference standard or ground truth should always include interpretation by a consultant radiologist or radiographer.

Item	Description	EAG comment
		The EAG further notes that ground truth is implicit in the derivation of sensitivity and specificity and therefore is not a comparator in the economic evaluation per se.
Outcomes eligible for inclusion	Intermediate outcomes • Measures of diagnostic accuracy to detect fractures • Accuracy when used by different healthcare professionals (emergency nurse practitioners, advanced clinical practitioners, urgent care doctors, diagnostic radiographers) • 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 (plaster casts, surgical procedures, physiotherapy appointments) and extent of treatments (complexity of surgery, length of physiotherapy course) • Number of hospital appointments/visits, including referrals to fracture clinics and orthopaedic assessment • Number of hospital admissions	In line with EVA methods, the EAG will prioritise scoped outcomes for consideration in the evidence review, i.e. in the event of large amounts of available evidence, the EAG will focus on the prioritised outcomes only. Prioritised outcomes are detailed in this protocol (Section 2.1) and were those expected to be most influential to decision-making for the NICE committee.

Item	Description	EAG comment
	• 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 technology for detecting fractures Clinical outcomes • Morbidity • Mortality Patient-reported outcomes • Health-related quality of life Cost outcomes • Cost of AI software • Staff costs for X-ray image interpretation • Training costs • Costs of additional medical appointments (including confirmatory imaging) • Costs of treatment • Costs of physiotherapy	
Healthcare setting	Emergency department (ED), urgent treatment centre (UTC) or minor injury unit (MIU).	Depending on the amount of available evidence, the EAG may make a pragmatic decision to include evidence where the setting is unclear. If this is the case, evidence clearly

Item	Description	EAG comment
		based in an ED, UTC or MIU will be highlighted and prioritised in the analysis.
Economic analysis	Costs will be considered from an NHS and Personal Social Services (PSS) perspective. The cost-effectiveness of interventions should be expressed in terms of incremental cost per quality-adjusted life year. The time horizon for estimating clinical and cost effectiveness should be sufficiently long to reflect any differences in costs or outcomes between the technologies being compared.	Relevant costs from NHS/PSS perspective would be considered, however, inclusion in modelling depends on relevant data availability. Similarly, expressing cost-effectiveness in terms of incremental cost per quality adjusted life year would depend on availability of relevant utility estimates from literature.

Abbreviations: AI, artificial intelligence; EAG, External Assessment Group; EVA, Early Value Assessment

1.2.1. Population

The decision problem includes people of all ages with a fracture, which is consistent with the population that would be seen in urgent care (ED/A&E, UTC or MIU) and therefore may be subject to the technology in practice. The EAG understood that, if used, the technology would preferentially be used as a decision-aid for any suspected fracture or dislocation in urgent care, and not limited for use for specific fracture or dislocation types or certain subgroups of the population. However, as some of the included technologies are not indicated for use in some fractures, the use of these technologies may require a change in the target population.

At the scoping workshop, stakeholders noted that bones will have reached maturity by age 16 years, after which time, bones age at different rates. This means that there may be no accepted older age population, but that presence of frailty or conditions that affect bone health may be indicators of a more at-risk older population. The EAG considered that outcomes of the technology may vary across age groups; for example, children may be more likely to experience more subtle fractures (e.g. green stick) while frailty or the presence of osteoporotic disease would increase the risk of fractures and poorer clinical outcomes. The EAG therefore considered that a recommendation to use the technology in all age groups would best be informed by evidence using a mix of age groups consistent with the proportions seen in NHS clinical practice. As stated above, the EAG considered it unlikely that the technology would be targeted towards certain populations, such as based on age (children) or the presence of frailty or conditions affecting bone health. However, where feasible within the timeframe of the assessment and where evidence is available, the EAG will consider evidence specific for those subgroups.

Stakeholders highlighted types of fractures that would typically not require the technology, either because X-ray is typically not used (e.g. suspected multiple fractures from a trauma) or because the fracture is obvious (e.g. open fracture). Stakeholders also suggested that the technology may have limited benefits for some areas of the body where initial interpretation of the X-ray is already highly accurate. They also suggested that, in some cases, the technology may lead to overdiagnosis and unnecessary treatment of injuries. For the technology to be beneficial, therefore, it would need to have sufficient benefit for other fractures to be of an overall net positive to the service.

The EAG noted that the objective of the assessment as stated in the NICE scope was to evaluate the use of the technology to diagnose fractures, and interpreted this to mean that it was outside of the scope to evaluate the benefits of using the technology to inform treatment planning only. For example, the EAG was aware that X-ray may be used in circumstances where the presence of a fracture is clear without imaging and X-ray would only be used to determine whether surgery was required. While the EAG determined this to be outside the scope of the assessment, it considered that it would likely be difficult to identify and exclude evidence from these circumstances from the evidence base unless studies were designed specifically for this purpose only. Where this is not clear, the EAG will take a pragmatic approach to include the evidence and will consider this issue in its interpretation of the evidence.

Stakeholders to the NICE scope commented on the likely fractures where the use of the technology would have greatest impact. This includes fractures where the technology may increase identification by clinician assessment alone and would therefore have benefits for onwards management (e.g. hip fractures, wrist fractures, ankle fractures). Stakeholders highlighted scaphoid fractures in the wrist as being particularly complex to identify, and while NICE guidance specifies that MRI should be considered for assessing a suspected scaphoid fracture, stakeholders noted that this may not always happen due to limited access to MRI.

As the potential outcomes of using the technology would be expected to vary across the population, and in order to capture some outcomes in a meaningful way, the EAG will consider evidence specific to a small number of suspected fracture types as subgroups to the assessment. The fracture types selected will depend on evidence availability but will likely prioritise those fractures where stakeholders considered the potential value of the technology to be greatest, such as hip, foot (including ankle) and hand (including wrist) fractures in adults, and elbow and Salter-Harris fractures in children.

In general, the EAG will prioritise the inclusion of studies that evaluate the technology in a mixed population (including a variety of suspected fracture types that would be seen in urgent care) or in specific fracture types specified as subgroups in the protocol. However, studies in other specific fracture types will be included and considered where feasible.

The EAG was aware that the type of fracture and the way this would typically be managed in the NHS would therefore be important for interpreting the outcomes of the technology in published evidence. The EAG therefore intends to draw upon stakeholder input in the interpretation (as well as the selection) of evidence for this EVA (see Section 5). Due to the need for pragmatism in EVA methods, if there is a lack of available evidence, the focus on these fractures as subgroups may need to change. If this is the case, the EAG will highlight any evidence gaps.

1.2.2. Interventions

The technologies included in the NICE scope[1] (Section 2.2) are shown in Table 2. These technologies cover a variety of fracture types, regulatory status and other features.

AI technologies evolve, usually in the form of periodic updates. The EAG will aim to report the version of AI technology used in any evidence presented. If sufficient information is available, the EAG will consider how the version of the technology might impact outcomes and use available evidence from the most recent version of the technology wherever possible. As noted in Section 1.2.1, some of the included technologies are limited for use to certain areas of the body or fracture type.

At the scoping workshop, stakeholders highlighted that the level of certainty in AI decisions would impact clinical decision-making; i.e. if the technology identifies a fracture with higher or lower confidence. The EAG considered that AI technologies may approach certainty in decision-making differently and will consider this in the assessment.

The EAG considered that the outcomes in this EVA may vary according to who is operating and interpreting the technology. This is particularly the case because the aim of these technologies is to assist rather than replace human judgement. The EAG was also aware that in clinical practice, different healthcare professionals may be evaluating X-rays. Due to this, the intervention will not be limited by who is operating the AI technology and interpreting the X-rays, providing they would reasonably do so in UK clinical practice. At the scoping workshop for this assessment, stakeholders noted that the technology may have most benefit when used by staff less trained in reading x-rays, but that in these circumstances there could be a risk that staff feel less confident in overruling a decision from the technology that they disagree with. Where evidence is sufficient, the EAG will consider different healthcare professionals/grades as important subgroups.

It is expected that the availability of AI technology may change the care pathway for suspected fractures, including the way in which the definitive radiographer/radiologist review is conducted (e.g. by changing the grade of professionals that evaluate the X-rays or influencing which X-rays receive a second review) or downstream service use (e.g. number of follow up appointments, number of people sent to virtual fracture clinics). The EAG evidence review aims to capture any variation in the care pathway reported by included studies, though consider that stakeholder input may best be able to consider the relevance of the included studies for how the technology would likely be used in the NHS.

Table 2: Summary of AI technologies included in the assessment

Technology	Fracture types covered	Regulatory status	Other
BoneView (Gleamer)	The company states that the software identifies fractures across entire appendicular skeleton, ribs and thoracic-lumbar spine. In addition, joint effusions, bone lesions and joint effusions. Suitable for people aged 2 years and over.	Class IIa CE marked	Uses X-ray radiographs in DICOM format. The results appear as bounding boxes around detected abnormalities.
Rayvolve (AZmed)	Fractures across appendicular skeleton and ribs. In addition, dislocations, joint effusions and chest pathologies. Suitable for adults only.	Class IIa CE marked	The company stated that Rayvolve identifies fractures and presents the results directly into the clinicians' interpretation console in the existing DICOM series. The tool is integrated into hospitals' existing radiology workflows using Wellbeing's AI Connect gateway.
RBfracture (Radiobotics)	The company states that RBfracture detects fractures across the entire appendicular skeleton and ribs. In addition, effusion of the knee and elbow, lipohaemarthrosis of the knee, and periprosthetic fractures. RBfracture is not intended to detect chronic or healed fractures. Suitable for people aged 2 years and over.	Class IIa CE marked	Healthcare professionals view the outputs of the software in their existing PACS/DICOM viewer. RBfracture returns a summary report overview including a red dot to indicate if a fracture or other finding has been detected. It also provides annotated radiographs with bounding boxes around areas of interest and a summary field with the analysis results. Bounding boxes with a dashed line indicate findings with a low confidence score, and a solid line indicates those with a high confidence score. RBfracture provides information in the PACS worklist about whether or not a supported lesion (fracture, effusion, lipohaemarthrosis) is detected.
qMSK (Qure.ai)	The company states that qMSK can detect fractures in the appendicular skeleton and ribs. Suitable for adults only.	Class IIb CE marked	
TechCare alert (Milvue)	Fractures across entire appendicular skeleton and ribs. In addition, effusion of the elbow, dislocations and chest pathologies. Suitable for adults and children without age limit.	Class IIa CE marked	TechCare alert is a special configuration of the Milvue Suite.

Abbreviations: CE, Conformité Européenne; DICOM, Digital Imaging and Communications in Medicine; PACS, picture archiving and communication system

1.2.3. Comparators

The primary comparator in the NICE scope is initial clinician interpretation of X-rays without AI assistance.

Given the variation of staff used to interpret X-rays in urgent care settings (see Section 1.3), the EAG will consider an additional comparator in the evidence review, which is a comparison of the same AI technology as a decision aid with different healthcare professionals (e.g. staff with different training/grade). Understanding of variation in outcomes according to staff use will be important for understanding the potential value of the technology in practice.

As this is an EVA, the EAG assumed that there will not be evidence that has compared outcomes of AI technologies in a head-to-head comparison, however the EAG will include these comparisons if identified.

The EAG was aware that similar open-source technologies are available, but these technologies are not licensed, and were therefore not considered to be suitable comparators for this EVA.

The EAG considered that, for assessing the diagnostic accuracy of AI technologies, the reference standard or ground truth should be a definitive report from a radiology department (including interpretation by consultant radiologists or reporting radiographers). The EAG received advice that the definitive review of X-rays in a radiology department was not necessarily 100% accurate for detecting fractures, particularly those fractures that are more difficult to detect (e.g. scaphoid, radial head). However, as this process would be used in clinical practice, this was judged to be a suitable reference standard for the assessment. The implications of using a reference standard that is not 100% accurate will be considered in the evidence review.

1.2.4. Outcomes

The EAG considered that all outcomes listed in the NICE scope were relevant to the assessment objectives. Given the recent development of the technology, the EAG expected that there may not yet be evidence for many of the scoped outcomes. However, if a large evidence base is identified, the EAG has marked the outcomes that will be included preferentially within the timeframe of the EAG assessment, as is consistent with EVA methods. The outcomes prioritised are those that most fundamentally assess the potential benefits and harms of the technology and those that are expected to be required for the economic analysis.

The EAG had the following considerations with regard to understanding the potential clinical and cost effectiveness of the technology:

- A potential benefit of the technology could be to avoid a delayed definitive diagnosis that could impact on patient health and on service-level outcomes (e.g. changes to delays in discharge, or in the need for recall if discharge occurs before a definitive diagnosis). The EAG considered that not all people with a delayed diagnosis would experience negative outcomes, as either the management of the injury would be the same regardless of whether a fracture was present or not (e.g. rib fracture) or a short delay to the correct management would be unlikely to have a negative impact on clinical outcomes (though may require additional healthcare resource use). However, for some fracture types and situations, a delayed diagnosis may have implications for outcomes, such as incorrect healing of the fracture or the risk of complications. A faster diagnosis may therefore be more clinically and cost effective in some cases than in others.

- A potential benefit of the technology could be to avoid a missed fracture. The EAG was uncertain about the number of true fractures that would be missed at both initial and definitive review, though considered that this would be a small proportion of those reviewed. Fewer still fractures may be missed if more stringent methods for review were used, such as definitive review by two consultant radiographers or radiologists. The EAG considered it plausible that fractures missed during both the initial assessment and the definitive review could include fractures with no or limited long-term consequences (e.g. subtle, hairline fractures) as well as those with potential serious consequences (e.g. scaphoid, facial fractures). In the latter instances, depending on local protocols, a missed fracture which might have implications for long-term health and healthcare resource use.

- A potential benefit of the technology could be to improve the efficiency of subsequent imaging. For example, a diagnosis of hip fracture using the technology may reduce the need for CT or MRI imaging, thus reducing healthcare resource utilisation and demand for these other imaging techniques. The EAG considered that this benefit might be limited in centres where CT or MRI aren't immediately available.

- A potential harm of the technology could be to increase the number of fractures detected that would not have required special management (i.e. over-diagnosis), such as hairline fractures. Similarly, if the technology detects previously healed fractures this might lead to over-diagnosis. The technology could therefore increase the use of unnecessary healthcare resource.

Stakeholders at the scoping workshop considered that avoiding a delayed diagnosis due to fractures being missed during the initial assessment was the most important outcome for determining the value of the technology.

1.2.5. Economic analysis

The economic analysis will be performed in accordance with NICE reference case[2] for the decision problem (as described in Table 1). Relevant costs from NHS and PSS perspective will be considered, with the following considerations:

- Based on SCM opinion at the NICE scoping workshop, the EAG noted that radiology reporting costs could vary substantially between in-house, out of hours, and outsourced radiology reporting services. EAG analyses will attempt to capture this difference in costs via scenarios if there is relevant data availability in the literature.

- Staff training costs and extra staff time costs due to initial expected delay as staff learn to use AI technologies would also be considered where possible and relevant.

Further, expressing cost-effectiveness in terms of cost per quality adjusted life years would depend on availability of suitable utility estimates in the literature. Further details about the proposed analysis are provided in Section 3.

1.3. Care pathway

An overview of the care pathway in the NHS is provided in the NICE scope. The EAG considered that variation in the care pathway in the NHS and in the way that the technology may be used may affect the clinical and cost effectiveness of the technology. The EAG will consult with stakeholders during its assessment on the typical care pathway used in NHS services. In preparing the protocol, the EAG considered the following:

- There is variation both within and across centres in which staff are involved in the initial and definitive interpretation of X-ray. Initial interpretation may be done by an emergency nurse practitioner (ENP), advanced clinical practitioner (ACP), or doctor, and in some cases initial interpretation may include support from a diagnostic radiographer. Definitive reporting may be done by a consultant radiologist or reporting radiographer. The EAG understood that the diagnostic accuracy of AI technology for interpreting X-rays may vary according to the staff using it (Kuo et al. 2022[3]), which would lead to variation in outcomes. The EAG considered that definitive report by either a radiographer or a radiologist would be equally accurate, though noted that this may affect the associated staff costs. The EAG heard that definitive review by a radiographer or radiologist would not necessarily be 100% accurate, due to the difficulty in assessing some fracture types.

- The EAG understood that the definitive review of X-rays will typically be conducted on the same day (including before discharge, which is known as ‘hot reporting’) or within a couple of days of presentation, though stakeholders to the appraisal suggested that in a minority of centres, it may be weeks before a person receives a definitive diagnosis. The time between presentation and definitive diagnosis will influence outcomes of a fracture missed in the original assessment, particularly for some fracture types. The EAG received feedback that definitive review typically occurs in order of presentation, and that no triaging or prioritisation of certain X-rays takes place. However, the EAG also heard that, where radiographer and radiologist capacity is limited, definitive review of some X-rays may not take place. For example, definitive review may only take place for X-rays where no abnormality has been identified in the initial assessment; i.e. priority will be given to ensuring that a fracture has not been missed. The EAG also heard that some centres will use approaches to avoid the risk of a fracture missed at both the initial assessment and definitive review, such as using alternative imaging (e.g. MRI) or assuming the presence of a fracture despite negative findings on the X-ray. Variation in how centres approach definitive reporting may therefore affect the accuracy of definitive reporting and the impact of inaccurate assessments at initial presentation.

- While NICE guidelines may recommend specific imaging pathways for the identification of certain suspected fractures, the EAG heard from stakeholders that not all centres may follow NICE guidance due to broader pressure on facilities. For example, NICE guidance (NG38[4]) suggests that clinicians consider using MRI to assess suspected scaphoid fractures, but this may not always be feasible due to high demand for MRI, and therefore X-ray may be used instead of, or to triage for, MRI.

- At the scoping workshop, it was noted that no fracture being identified on initial assessment may not necessarily result in the person being discharged without ongoing tests and management. There are instances where patients may be placed in a cast as a precaution while awaiting subsequent definitive review.

- The EAG considered it plausible that the availability of the technology could change the care pathway in some instances. For example, if the technology were to increase identification of fractures and there was trust in the use of the technology, this could lead some centres to rely on lower grade staff to interpret X-rays on presentation or allow some centres to have longer delays before a definitive diagnosis.

1.4. Objectives

There were several objectives for the assessment outlined in the NICE scope:

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?
2. What evidence is available to support the value proposition outlined in the scope?
    a. Improve the accuracy of fracture detection from X-rays in the emergency department, urgent treatment centre or minor injuries unit
    b. Service delivery and workflow improvements, for example, reduced waiting times, fewer people being recalled, and a reduction in unnecessary fracture clinic referrals and medical appointments.
3. What are the evidence gaps?

The EAG assessment will address these research questions by conducting a review of the current evidence base for the technology, including assessment of any unpublished data provided by the companies. The EAG will seek to characterise and map the available evidence in terms of how well it meets the NICE scope and will identify any key gaps and related research recommendations. The EAG will also aim to develop a conceptual economic model structure that would be appropriate for providing an early view of the plausible cost effectiveness of the technology in accordance with the NICE scope and EVA methodology. Where feasible, based on the availability of evidence, preliminary cost effectiveness results will be produced.

The EAG considered that the target population for the assessment was all people with a suspected fracture or dislocation where the technology would be used to aid diagnosis of a fracture during the initial assessment in urgent care. The assessment will therefore seek to identify outcomes for the technology in a mixed population of fracture types and consider the applicability of these to clinical practice in the NHS. Given that the outcomes of the technology are likely to vary across the population, evidence in a mixed population may have limited applicability to the NHS where the case mix in the studies is not representative of the case mix seen within the NHS. As the EAG expects that case mix also varies across services and areas of the UK, according to the demographic characteristics of the population, nuanced interpretation of evidence from a mixed population will be needed. Where feasible, the EAG will explore the impact of variations in prevalence rates in the analysis.

To support decision-making, and in addition to evidence in the mixed population, the EAG will also seek to include evidence that evaluates the outcomes of the technology in population subgroups with specific fractures (as described in Section 1.2.1). The fractures selected are those where stakeholders considered that the potential benefit of the technology would be greatest, either because identification at initial assessment is challenging and/or there are significant consequences of an inaccurate diagnosis during initial assessment. This approach essentially evaluates the technology on a ‘proof of concept’ basis, i.e., if the technology is not beneficial in these suspected fracture types, it will likely not be beneficial across the broader population.

2. EVIDENCE REVIEW

A single search strategy will be used to identify all evidence types for the review (see Section 2.2). Evidence will then be filtered according to its relevance for the synthesis and economic analysis.

Consistent with methods for technologies appraised within the NICE EVA process, the evidence review will use a pragmatic approach to identifying and synthesising the evidence base. This will consist of a systematic and comprehensive literature search to identify the relevant evidence base followed by pragmatic methods to select, appraisal and analyse this for the purposes of the assessment.

The pragmatic methods used will be guided by the availability of evidence; for example, where a high volume of evidence is identified, the EAG will prioritise the inclusion of a subset of evidence that best meets the assessment objectives.

2.1. Inclusion criteria

The inclusion criteria for the EAG's assessment are shown in Table 3. At the time of writing, the likely quantity of evidence that would be identified by the review was unclear. Scoping searches conducted by the EAG revealed that a high volume of research evaluating AI technologies for the detection of fractures has been published in the past several years, however as the abstracts of published studies rarely specify the name of the technology under evaluation, the EAG remains unsure what proportion of the published evidence evaluated technologies relevant for inclusion in the assessment.

To ensure that the assessment can be completed within the timeframe for the EVA, and as described at the start of this section, if a large evidence base is identified, the EAG will seek to prioritise evidence that it expects will be most influential to decision-making. With regard to the inclusion criteria, this means that certain evidence included during full-text screening will be prioritised for inclusion in the review: these are highlighted with * in Table 3.

Studies that do not meet the inclusion criteria will be excluded from the review. The EAG expects to identify research that partly but not wholly matches the inclusion criteria for the assessment, e.g., mixed populations including some fractures not eligible for inclusion, eligible technologies operated by clinical staff not relevant to NHS practice, or outcomes that are similar but use a broader or narrower definition than the outcome used in the inclusion criteria. In these situations, the EAG will consider the relevance of the evidence to the decision problem on a case-by-case basis to determine whether the results would be useful to decision-makers and will make a judgement call on inclusion. Where required, the EAG will discuss these decisions with the NICE team. In the case of mixed patient or operator populations, the EAG will typically include studies where ≥80% of the study population is consistent with the inclusion criteria. A list of studies excluded at full-text with the reasons for exclusion will be provided in the EAG report (see Section 2.3).

The evidence review will include diagnostic, clinical, patient and service outcomes from comparative studies only. This includes studies that compare the diagnostic accuracy of an included technology with the reference standard and studies that report other outcomes in an included technology compared to one of the comparators. Non-comparative studies are uninformative for the assessment as they cannot provide diagnostic accuracy data and other outcomes are typically subject to a high level of bias. The inclusion of non-comparative studies also significantly increases the resource needed during literature screening, which would not be feasible within the timeline of this EVA. However, given that the evidence base for the technology may be at an immature stage, the EAG will consider non-comparative studies of the technology submitted by companies during the assessment where there is paucity of studies for that technology identified in the evidence review. This is discussed further in Section 4.

Table 3: Inclusion and exclusion criteria

	Inclusion criteria	**Exclusion criteria**
Population and target condition	Children and adults presenting to an urgent care setting (ED, UTC, MIU) with a suspected fracture or dislocation, for which an X-ray is requested. Studies including a mixed population (e.g. representing a typical case mix in urgent care) will be prioritised for inclusion*	Children and adults with a confirmed fracture where X-ray is requested to inform treatment planning only.
Subgroups	• Children and young people (0 –16 years) • People demonstrating frailty (as indicated by a frailty measure)*	

	Inclusion criteria	**Exclusion criteria**
	• People with conditions affecting bone health (for example, osteoporosis and osteogenesis imperfecta)* • Adults with suspected: o Hip fractures o Foot (including ankle) fractures o Hand (including wrist) fractures • Children with suspected: o Elbow fractures o Salter-Harris fractures	
Intervention	AI used as a decision aid for X-ray image interpretation and fracture assessment prior to radiology review by clinical staff that would typically interpret X-rays in the included settings, using any of the following software/platforms: • BoneView (Gleamer) • Rayvolve (AZmed) • RBfracture (Radiobotics) • qMSK (Qure.ai) • TechCare Alert (Milvue)	• Other AI software/platforms • AI software/platforms used to interpret X-ray images for purposes other than for fracture or dislocation diagnosis • AI software/platforms used to interpret other imaging types (e.g. CT) • AI software/platforms not used as a decision aid (i.e. used alone without clinical judgement)
Comparators	• The same clinical staff without the use of the technology* • One of the included AI software platforms also used as a decision aid (i.e. a 'head-to-head' comparison) • The same AI software/platform used as a decision aid but operated by a different clinical staff grade or group*	• Different clinical staff without the use of the technology (e.g. different staff grade/speciality) • A head-to-head comparison with an included AI software/platform where one or both technologies are not being used as a decision aid (i.e. not in addition to clinical judgement) • A head-to-head comparison with AI software/platform not on the included studies list

	Inclusion criteria	Exclusion criteria
Reference standard/ground truth	Definitive report from a radiology department (including assessment by a consultant radiologist or reporting radiographer)	Definitive report not including assessment by a consultant radiologist or reporting radiographer
Outcomes	Diagnostic accuracy • TP, TN, FP, FN • sensitivity, specificity* • PPV, NPV Clinical and patient outcomes • Prevalence rate of fractures* • Rate of missed fracture-related further injury • Number of treatments and extent of treatments • Morbidity • Mortality* • Health-related quality of life* Service outcomes • Healthcare professional X-ray reading time • Certainty in diagnosis / diagnostic confidence (as defined in sources) • Time to diagnosis/definitive radiology report* • Time spent in urgent care • Time to treatment • Proportion of people requiring further imaging • Number of people recalled following discharge or following definitive radiology report • Number of further imaging events	Other outcomes

	Inclusion criteria	Exclusion criteria
	• Number of referrals to fracture clinic/orthopaedic assessment • Number of hospital appointments/visits • Number of hospital admissions • Length of stay in hospital • Failure rate or rate of inconclusive AI reports* • Healthcare professional user acceptability Economic outcomes • Cost of AI software* • Staff costs for X-ray image interpretation* • Training costs* • Costs of additional medical appointments • Costs of further imaging • Costs of treatment • Costs of physiotherapy • Costs of radiology reporting (in-house versus outsourced) • Other rehabilitation costs (e.g. for hip fractures)	
Study design	Diagnostic accuracy • Single-gate or two-gate study designs that provide TP, TN, FP and FN data for at least one included technology against the reference standard or ground truth (providing that all of these data and/or both sensitivity and specificity are reported) Clinical, patient and service outcomes	Single-arm studies (i.e. where there is no comparator or comparison with a reference standard/ground truth)

	Inclusion criteria	Exclusion criteria
	• Experimental designs (e.g. RCTs) comparing the intervention with a comparator or with the reference standard/ground truth • Comparative observational studies comparing the intervention with a comparator or with the reference standard/ground truth Economic analysis • Cost-effectiveness (including cost-utility) or Cost-consequences analysis comparing the intervention with a comparator or with the reference standard • Costing studies reporting the costs of relevant interventions or reference standard (from health system perspective) • Modelling studies reporting relevant costs or cost-effectiveness estimates	
Setting	Any emergency healthcare setting where people first present with a suspected fracture (ED, UTC, MIU, or similar in a non-UK country) All countries will be relevant for inclusion, though variations in healthcare systems will be considered when interpreting the findings.	

Abbreviations: AI, artificial intelligence; CT, computed tomography; ED, emergency department; FN, false negative; FP, false positive; MIU, minor injuries unit; NPV, negative predictive value; PPV, positive predictive value; RCT, randomised controlled trial; TN, true negative; TP, true positive; UTC, urgent care centre

2.2. Search strategy

Searches for clinical and cost-effectiveness will be conducted in one strategy, without any study type filters. The searches will be an update and extension of Kuo 2022 and will therefore be for papers published from July 2020 onwards. An exemplar search strategy for MEDLINE is provided in Appendix A.

The search process will include searching the following sources:

- Electronic databases, including MEDLINE (inc In-Process and PubMed-not-MEDLINE records), EMBASE and Cochrane.
- Economics sources, such as NHS EED and CEA Registry.
- Manufacturer websites.
- The WHO International Clinical Trials Registry Platform (ICTRP) and the US National Library of Medicines registry at clinicaltrials.gov.
- MHRA field safety notices and the MAUDE database will be searched for adverse events.
- Any industry submissions to NICE, as well as any relevant systematic reviews identified by the search strategy, will be scrutinised to identify additional relevant studies.
- Relevant clinical guidelines from NICE, SIGN and INAHTA, especially for economic modelling

In addition to the above searches, a targeted search of the broader literature may be undertaken if necessary to identify the evidence base in additional areas, e.g., HRQoL (health state utility values), resource use and costs for treatment and side-effects (UK studies only if available), and the methods available for the modelling of AI for fractures to inform cost-effectiveness analyses. The search strategies employed will be fully reported and described.

2.3. Study selection

Screening of evidence identified by the literature search will be carried out in Rayyan[5]. Three levels of screening will be used to select evidence for the review, these are described below. Screening will be conducted by a single reviewer. Studies where eligibility is uncertain will be categorised as uncertain and discussed with a senior reviewer and/or the full team.

Title and abstract screening: the titles and abstracts of studies identified in the search will be screened using population, intervention, comparator, reference standard, study design and setting criteria.

Full-text screening: the full texts of studies included at title and abstract screening will be screened using the full inclusion criteria. Reasons for exclusion will be documented.

Priority screening: this level of screening will be conducted only if it will not be feasible for the EAG to consider all of the evidence identified at full-text screening. If this is the case, for example where a large number of studies are identified, the EAG will prioritise the inclusion of studies that meet the priority criteria shown in Table 3. Reasons for prioritisation will be documented.

2.4. Data extraction strategy

Data from the included studies will be extracted into two data extraction tables (DETs) developed in Microsoft Excel *a priori*: one for the diagnostic accuracy, clinical, patient and service-level outcomes; one for the economic outcomes. The DETs will be piloted using six included studies (two diagnostic accuracy studies, two economic evaluations, two clinical and/or service evaluations) and revised before extraction of the remaining studies. Data extraction will be conducted by a single reviewer. Where indicated, complex areas of data extraction will be flagged by the reviewer for quality assurance by a second reviewer.

2.5. Quality assessment strategy

Consistent with the methods for an EVA, no formal quality assessment of the included evidence will be included as standard. However, the EAG may conduct a formal assessment for pivotal evidence included in the review (i.e. a subset of studies that are influential in the EAG model or the EAG conclusions).

While a formal quality assessment will not typically be conducted for the evidence base, the EAG will nevertheless ensure that studies included in the review are of sufficient quality to inform decision-making. Studies where the methodological design severely limits the EAG's confidence in the reliability or validity of the data will not be included. Such studies can be misleading to the committee and may undermine confidence in other, more robust, studies. A key quality consideration with studies evaluating the diagnostic accuracy of a technology is whether the findings can reasonably be generalised beyond the sample and methods used in the study. Diagnostic accuracy can be highly influenced by variation in how the technology is applied and interpreted, and by the case mix of the study population, and it is particularly important that a suitable reference standard/ground truth is used. The EAG will therefore strictly apply the *a priori* inclusion criteria for the review and afterward seek to consider whether the findings are relevant to the objectives of the review and prioritise studies accordingly.

2.6. Methods of synthesis/analysis

Diagnostic test accuracy outcomes will be tabulated and described in a narrative synthesis. Where possible, and required, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and prevalence will be calculated and reported for each study. Similarly, data for clinical, patient and service outcomes will be tabulated and described in a narrative synthesis.

Where sufficient studies of reasonable similarity are identified, data will be pooled in meta-analyses. Data will be pooled for each AI technology, as well as for important subgroups (specified fracture types, paediatrics, frailty and bone conditions), only where sufficient data are identified for each technology to allow for meaningful pooling and interpretation.

2.7. Reporting of the evidence base

A tabulated overview of the evidence landscape will be constructed to represent the evidence available and where there are meaningful gaps for the decision problem.

Evidence gaps identified pertaining to the intermediate and final outcomes from the scope and those pertaining to the economic modelling will be summarised in tabular and narrative form. If appropriate, a ‘traffic light’ scheme will be used to highlight relative importance of the gap. Key areas for evidence generation will be summarised in tabular form. Narrative text will also address missing clinical evidence for other parts of the scope, such as population (including key subgroups), setting and comparators. The overall relevance and validity of the included evidence will be considered in the evidence landscape and gap map.

3. ECONOMIC ANALYSIS

Decision analytic modelling will, conditional on relevant data availability, aim to assess the cost-effectiveness of AI used as a decision aid for fracture detection compared to unassisted diagnosis in urgent care. However, if there is no relevant evidence found in the AI technology studies or literature to derive model inputs, then a conceptual model would be produced which could potentially inform a full economic evaluation once further evidence becomes available. An early iteration of the potential conceptual model pending clinical expert validation has been presented in Appendix B.

The details of a potential approach to modelling are described below in Section **Error! Reference source not found.**. Please note that this approach would likely evolve with more understanding of the underlying evidence.

3.1. Proposed modelling approach

Diagnostic performance of AI used as a decision aid for fracture detection compared to unassisted diagnosis will be modelled primarily through prevalence, sensitivity, specificity and cost per scan, with estimates of cost and QALYs attributable to true and false positives and negatives. As described in Section 1.4, depending on data availability, the EAG will either model the entire decision problem as one (i.e. based on all suspected fractures presenting in a given setting) or, given the heterogeneous nature of the target population and care pathway, the EAG may focus on priority fractures (identified in the NICE scope and inclusion criteria as fracture sites of interest: hand (including wrist), and foot (including ankle)) in adults and elbow and Salter Harris fractures in children). This would evaluate the technology on a 'proof of concept' basis (as noted in Section 1.4), though the EAG would consider inclusion of other type of fractures typically seen in urgent care where feasible.

The analysis will be conducted via a short-term decision tree (suitable time horizon will be decided upon validation of conceptual model with the clinical experts) and any other relevant longer-term costs and consequences will be applied as one-off derived from literature contingent upon its availability (see Figure 1 for possible structure/conceptual model). Such a model could be run for each type of fracture under consideration at any given time and a weighted average of the costs and consequences could be calculated based on the case mix for a typical urgent care setting. Depending on the set-up and funding model in a given setting, it may be impractical for the AI algorithm to be limited to the specific suspected fracture types analysed. Therefore, the EAG will explore scenarios under the assumption of zero added benefit in the detection of other fractures, with added costs according to the licensing arrangements. For example, if the license is a site license there will be zero added cost for these extra scans. If the license is fee-per-scan then these costs will be factored in.

In summary, the EAG anticipates to conduct either a single decision model estimating the expected cost and outcomes associated with AI-assisted diagnosis vs unassisted diagnosis, or replication of the same (or similar) model structures exploring key fracture locations with a subsequent analysis estimating the total cost and health impact for a ‘typical’ urgent care setting based on case mix. Relevant subgroups will be considered as detailed in Table 1, where there is suitable evidence. A discount rate of 3.5% for both costs and consequences would be applied.

This approach is similar to that of a study by Curl et al. (2024)[6] evaluating the cost-effectiveness of AI based opportunistic compression fracture screening, using a hybrid decision tree plus Markov model, comparing strategies of opportunistic screening for osteoporotic vertebral compression fractures (OVCFs) against usual care.

Utility values from literature or clinical expert opinion will be included, based on which, QALYs would be calculated in the economic model.

Costs and resource use from healthcare system perspective would be informed primarily by NHS reference costs, unit costs from Personal Social Services Research Unit (PSSRU), discussions with clinical experts and relevant documentation from technology manufacturers. Other literature sources would be consulted as necessary.

If the model is non-linear (e.g. Markov model), probabilistic analysis will be performed by drawing samples from appropriate input parameter distributions at each simulation. Probabilistic results will be presented as expected costs and outcomes, with uncertainty represented using cost-effectiveness planes and/or cost-effectiveness acceptability curves (CEACs). Sensitivity and scenario analyses will be performed to access the uncertainty around point estimates used, as necessary.

Model outputs will primarily be expected to be cost and QALYs accrued from each strategy and incremental analysis, but if appropriate other intermediate outcomes such as diagnostic accuracy and how that changes with different grades of healthcare professionals using the AI technologies, timeliness of diagnosis and number of missed fractures with and without AI decision aid, rate of inconclusive AI reports, mortality, HRQoL, total costs of AI software (including staff and training costs) and costs of further imaging (CT/MRI etc.) will be reported.

The EAG noted that at the NICE scoping workshop, SCMs mentioned that estimates of the certainty of the AI detection would be helpful and would have implications for diagnostic confidence of AI algorithms in the long-term. This would require the algorithm to generate a probability of fracture or some similar propensity score as its output rather than a binary yes/no. If such an output is available, and can be mapped to sensitivity/specificity, the EAG will use the resulting Receiver Operating Characteristic (ROC) curve to identify the Optimal Operating Point (OOP)[7,8] based on the most cost-effective cut-off. At a point where the sensitivity and specificity corresponds to OOP, most accurate cost-effectiveness results could be obtained.[9]

4. HANDLING INFORMATION FROM THE COMPANIES

Within the NICE EVA process, companies are invited to submit information about their products, including any unpublished data that they have on file that may be useful to the assessment. Any data provided by companies will be screened for relevance using the criteria outlined in Section 2.1 and included in the analysis where eligible. As noted in the same section, diagnostic, clinical, patient and service outcomes from non-comparative studies are not being sought from the evidence review due to the resource requirements for this and the expectation that such evidence would be low quality for the assessment. However, in acknowledgement that the evidence for the technology may be at an immature level, where there is a paucity of evidence for a technology, the EAG will consider any non-comparative studies submitted by companies. These data are unlikely to be included in any analyses, due to risk of bias, however may be considered within the narrative synthesis and the landscape assessment.

Where required, the EAG will ask the companies to provide further detail or clarification on the information that they have submitted. Given the fast timelines of the EAG assessment, it is likely that the EAG will request that information from companies is received within a certain timeframe to ensure that it can be considered within the assessment. The EAG plans to send companies any such questions by 12th July 2024 and require any response to be received by 22nd July at the latest in order to consider these within the assessment.

Any confidential information received from companies will be protected and, if used to inform the assessment, will be highlighted appropriately for redaction from the published EAG report.

5. ADDITIONAL INFORMATION SOURCES

NICE will recruit experts and SCMs for this assessment, who will be consulted by the EAG during its assessment. In addition, the EAG plans to recruit additional clinical experts to advise with its assessment and who will be named as co-authors on the report.

6. TIMETABLE/MILESTONES

Milestone	**Date to be completed**
Submission of final protocol	28th June 2024
Submission of progress report	30th July 2024
Submission of draft Diagnostic Assessment Report	13th August 2024
Submission of final Diagnostic Assessment Report	28th August 2024

References

1. National Institute for Health and Care Excellence. Artificial intelligence software to help detect fractures in urgent care: final scope. June 2024.

2. National Institute for Health and Care Excellence. 5. The reference case. 04 April 2013. In: Guide to the methods of technology appraisal 2013 NICE process and methods [PMG9] [Internet]. Available from: https://www.nice.org.uk/process/pmg9/chapter/the-reference-case.

3. Kuo RYL, Harrison C, Curran T-A, Jones B, Freethy A, Cussons D, et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. Radiology. 2022;304(1):50-62.10.1148/radiol.211785

4. National Institute for Health and Care Excellence. Fractures (non-complex): assessment and management. NICE guideline [NG38]. Published: 17 February 2016.

5. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210.10.1186/s13643-016-0384-4

6. Curl PK, Jacob A, 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. 2024.https://doi.org/10.1016/j.jacr.2023.11.029

7. Sanghera S, Orlando R, Roberts T. Economic evaluations and diagnostic testing: an illustrative case study approach. Int J Technol Assess Health Care. 2013;29(1):53-60.10.1017/s0266462312000682

8. Sutton AJ, Cooper NJ, Goodacre S, Stevenson M. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. Med Decis Making. 2008;28(5):650-67.10.1177/0272989x08324036

9. Rautenberg T, Gerritsen A, Downes M. Health Economic Decision Tree Models of Diagnostics for Dummies: A Pictorial Primer. Diagnostics (Basel). 2020;10(3).10.3390/diagnostics10030158

Appendix A: Example search strategy

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

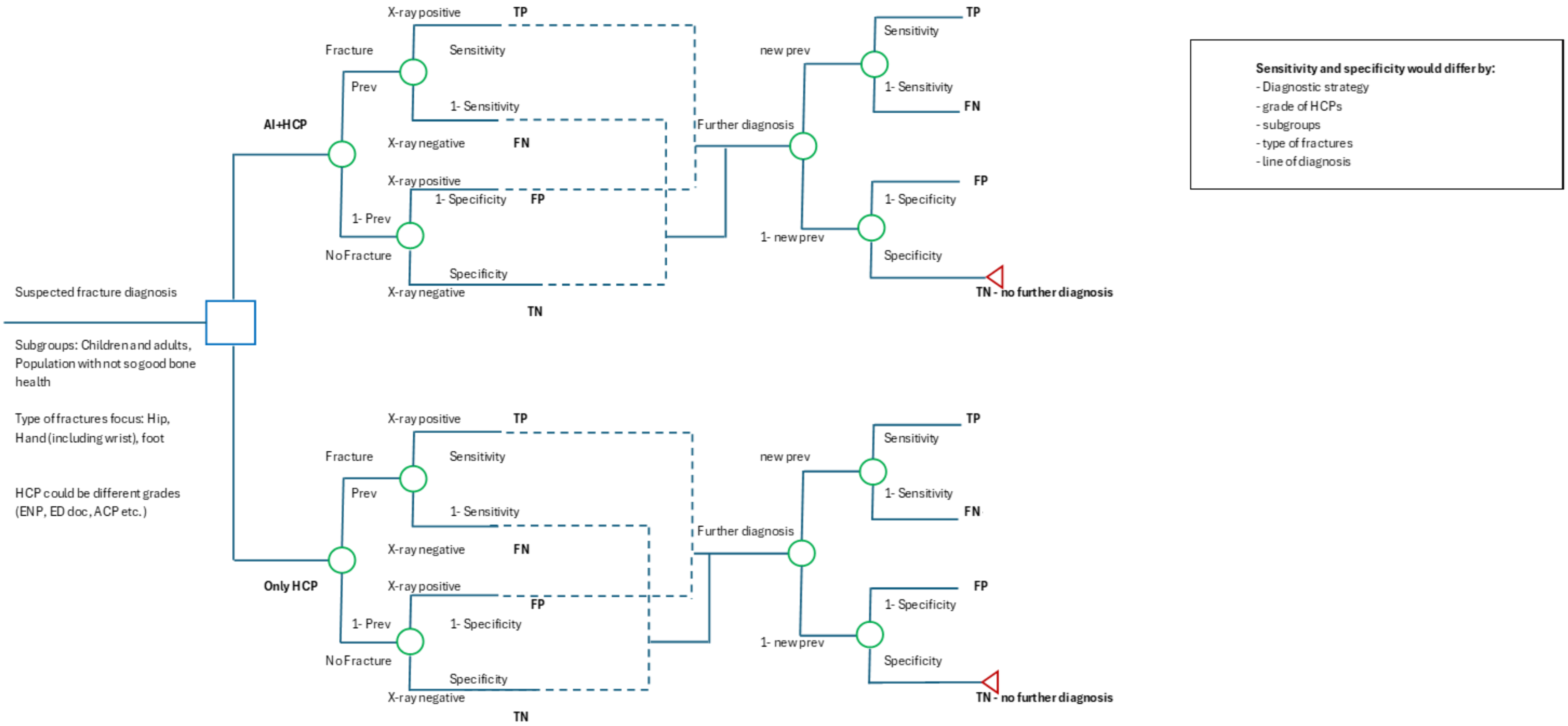
24 21 or 22 or 23 1068

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

Appendix B: Proposed model structure

Figure 1 presents an early iteration of the potential model concept. However, please note that this is currently pending clinical validation and may alter during the assessment in response to the evidence base. This is a general conceptual sketch of an early iteration of the proposed model, including the possibility for subsequent confirmatory review in the event of a positive, negative or in all cases following initial diagnosis. Longer term costs and QALYs will be attached to the terminal nodes which will vary by fracture type and population (e.g. frail vs non-frail).

Figure 1. An early iteration of the potential model concept



Abbreviations: Prev, Prevalence; AI, Artificial Intelligence; HCP, Healthcare professional, TP, True positive; FP, False positive; TN, True Negative; FN, False Negative; ENP, Emergency nurse practitioner; ACP, Advanced clinical practitioner; ED, Emergency department; doc, doctor.

Note: dashed lines show how intermediate diagnostic outcomes feed into the decision tree for further diagnosis. The final model will consider additional outcomes (e.g. costs, QALYs) resulting from intermediate diagnostic outcomes