NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE

Early Value Assessment consultation:

Supporting documentation – Committee papers

The enclosed documents were considered by the NICE Diagnostics advisory committee (DAC) when making their draft recommendations:

1. **Front sheet**
2. **Final Scope**
3. **Assessment report** – an independent report produced by an external assessment group (EAG) who have reviewed and critiqued the available evidence.
4. **Assessment report overview** – an overview produced by the NICE technical lead which highlights the key issues and uncertainties in the company's submission and assessment report.
5. **Factual Accuracy check comments table**
6. **Professional organisation submission**

Please use the above links and bookmarks included in this PDF file to navigate to each of the above documents.

NICE National Institute for Health and Care Excellence

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE

HealthTech Programme

GID-HTE10059 Artificial intelligence technologies to aid the opportunistic detection of vertebral fragility fractures

Final scope

1 Introduction

The prioritisation board identified artificial intelligence technologies to help detect vertebral fragility fractures on radiographic imaging as suitable for early value assessment (EVA) by the HealthTech Programme based on a topic intelligence briefing.

2 Technologies

This section describes the properties of the technologies based on information provided to NICE by manufacturers and experts, and publicly available information. NICE has not carried out an independent evaluation of these descriptions.

2.1 Purpose of the technologies

Vertebral fragility fractures (VFFs) are diagnosed on radiographic images involving the spine, including X-ray, CT, MRI and dual-energy X-ray absorptiometry (DXA) scans. However, up to 70% of VFFs are missed or underdiagnosed (Royal Osteoporosis Society, 2021). The Royal College of Radiologists' Radiological guidance for the recognition and reporting of osteoporotic VFFs notes several reasons why VFFs are missed:

- the spine is not routinely reviewed during reporting by the radiologist or radiographer

- a lack of awareness of the importance of early diagnosis of VFFs and use of ambiguous and inconsistent terms during reporting
- immature or underdeveloped departmental radiology information systems (RIS) and alert processes
- onward referral systems (in particular, fracture liaison services [FLSs]) may not be well developed locally.

Artificial intelligence (AI) technologies offer potential for improving the detection of VFFs, especially opportunistically (on images involving the spine taken for reasons other than VFF detection). AI technologies could be used to assist radiologists and radiographers when they interpret images. This could improve detection rates, leading to more people receiving care when necessary. The technologies may also reduce the time taken to interpret images, facilitating timelier referral for further evaluation.

Missed VFFs lead to complications such as a curved spine (causing the person to lean forward), height loss, immobility, pain, as well as loss of function. This can impact the person's quality of life and ability to perform daily tasks. VFFs are also associated with an increased risk of death and are a significant predictor of future fractures (Gonnelli et al. 2013). Over 55% of people with a hip fracture have evidence of previous VFFs (Royal College of Radiologists, 2021). The costs associated with VFFs and hip fractures are significant – the health and social care costs in the first year of post-hip fracture are over £33,000 per person (Royal College of Radiologists, 2021). 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.

2.2 Product properties

Several companies offer software with AI derived algorithms for analysing radiographic images to opportunistically detect VFFs. They use radiographs in DICOM (digital imaging and communications in medicine) format which are stored on the hospital's PACS (picture archiving and communications system). These technologies are designed to assist healthcare professionals in the interpretation of radiographic images to help diagnose VFFs. NICE notes that the ionising radiation (medical exposure) regulations (IRMER) state that clinical evaluation of X-rays requires a trained person. Therefore, AI technologies cannot currently be used autonomously without human interpretation. All of the identified technologies have algorithms that are fixed, but can be updated periodically. All of the technologies utilise either X-ray images or CT scans. Some of the technologies include additional features. Some of the companies provide their software directly on their own platforms, whereas others are available via multivendor platforms (for example, the Blackford Platform).

2.2.1 Annalise Enterprise (CXR) and Annalise Container (CXR) (Annalise.AI)

Annalise Enterprise (CXR) is a class IIb CE marked AI medical device intended to assist clinicians with the interpretation of X-ray images (both lateral and anterior-posterior) of people ≥16 years of age. The technology uses AI algorithms to notify of suspected findings, including VFFs. The company states that it is used in 48 trusts in the NHS. Suspected findings are displayed immediately alongside associated localisation information to the clinician as they view the study in the PACS viewer. Annalise Enterprise (CXR) is used in 48 trusts in the NHS. Annalise Container (CXR) has the same capabilities as Annalise Enterprise (CXR) and uses the same AI model, but it is a cloud-based technology that is hosted by AI marketplace platforms, such as Sectra Amplifier and Blackford.

2.2.2 BoneView (Gleamer)

BoneView is a class IIa CE marked software designed to assist clinicians in the interpretation of X-ray images. BoneView uses AI (deep learning) to detect anomalies in the appendicular skeleton, ribs and thoracic-lumbar spine. The company states that the software can detect VFFs, as well as other fractures, dislocations, effusions and bone lesions. Healthcare professionals view the results as DICOM secondary captures with bounding boxes around any abnormalities; a results summary sheet is also provided. The technology is currently available in the NHS.

2.2.3 BriefCase-Triage (Aidoc Medical)

BriefCase-Triage is a class I UKCA and class IIa CE marked software which uses an AI algorithm to analyse abdominal and chest CT images on adults ≥18 years of age. This technology is intended to detect suspected VFFs of the thoracic and/or lumbar spine (both acute and chronic). Suspected findings are notified to the clinician on annotated images. The technology includes additional features, such as image prioritisation options. BriefCase-Triage is currently being piloted in one centre in the NHS.

2.2.4 CINA-VCF Quantix (Avicenna.AI)

CINA-VCF Quantix is an AI algorithm that detects and labels VFFs on CT images of adults ≥50 years of age. The technology uses the Genant classification system. Suspected findings are displayed on annotated images along with a notification to the clinician and can be reviewed immediately. The technology is available as a cloud-based solution on the Sectra Amplifier and Blackford platforms. The company is in the process of obtaining CE certification and expects it to be certified as a Class IIb medical device. CINA-VCF Quantix has not been used in the NHS.

2.2.5 HealthVCF and HealthOST (Nanox AI)

HealthVCF is a cloud-based software which uses AI algorithms to detect VFFs on chest and abdominal pelvic CT scans of adults ≥50 years of age involving the T1-L5 portion of the spine. The technology uses the Genant classification system. Suspected findings are highlighted on a standalone desktop application in parallel with the standard of care. HealthVCF is CE certified and is currently used in several NHS trusts. HealthOST is a newer version of the algorithm, which includes additional features, such as analysis of low bone mineral density, labelling of vertebrae and measurement of the mean Hounsfield Units. CE certification of HealthOST is expected in 2025. The technology has not been used in the NHS.

2.2.6 IB Lab FLAMINGO (IB Lab)

IB Lab FLAMINGO is a software that uses AI algorithms to detect VFFs in CT scans of adults ≥50 years of age involving the thoracic and/or lumbar spine. The technology uses the Genant classification system. Suspected findings are displayed within a report which includes annotated images and other information, such as the range of vertebrae examined and labelling of the vertebrae, which can be reviewed immediately. The technology allows for image prioritisation based on risk. IB Lab FLAMINGO is CE certified as a class IIa medical device. It has not been used in the NHS.

2.2.7 TechCare Spine (Milvue)

TechCare Spine is a software which uses AI algorithms to detect VFFs in lateral X-ray images involving the lumbar and/or thoracic spine. The technology uses the Genant classification system. Suspected findings are displayed on annotated images along with a notification to the clinician and can be reviewed within 1-2 minutes. It is available as a cloud-based solution. CE certification of TechCare Spine is expected in June 2025, with anticipated classification as a class IIa medical device. The technology has not been used in the NHS.

3 Target conditions

A VFF is a break in the spine that occurs when bones are weaker than normal. It is defined as a reduction in vertebral height or vertebral deformity as a result of structural failure (when there is a height reduction or 20% or more, or an endplate deformation) after a fall from standing height or less. But they can also occur spontaneously as a result of day-to-day activities involving very little trauma or stress. VFFs are the most common type of fragility fractures caused by osteoporosis (a result of bone weakness) which reduces bone density and strength. They are often described as vertebral compression fractures. Other informal synonyms include wedge fractures, wedge deformity, collapse, compression and loss of height. Osteoporotic VFFs are common in the elderly and particularly in postmenopausal women, but they can also be associated with other conditions or factors, such as chronic or long-term corticosteroid/glucocorticoid usage or malignancy in the vertebrae. Other risk factors include history of falls, family history of hip fracture, low BMI (less than 18.5kg/m^2), smoking, alcohol intake of more than 3.5 units per day and secondary causes of osteoporosis such as rheumatoid arthritis, inflammatory bowel disease or malabsorption (CG146, 2017).

Many people with VFFs have no specific clinical signs, but they may experience sudden acute pain with local tenderness and chronic back pain in thoracic or lumbar spine, which gets worse when sitting and leaning backwards or standing and leaning forward, curved spine (causing the person to lean forward), limited mobility, functional disability or loss of height (more than 2.5cm). They may also have difficulties in breathing, performing daily activities, gastrointestinal problems, sleep disturbances and a range of psychological symptoms including anxiety, low mood, depression and low self-esteem. Multiple VFFs result in progressive height loss and abnormal curvature of the spine (kyphosis/hyperkyphosis). Therefore, VFFs cause significant morbidity and have a significant impact on a person’s quality of life, while also being associated with an increase in mortality (Clynes et al. 2020).

3.1 Epidemiology

It is estimated that approximately 2.5 million people in England and Wales have osteoporosis. Approximately, one in two women and one in five men will sustain one or more fragility fractures in their lifetime.

The incidence of VFFs increases with age. Recent data shows an incidence rate of 7.1 per 10 000 person years in adults aged over 50 (Curtis et al. 2016). Women are more commonly affected at all ages. An incidence of 12% has been reported in women aged 50 to 79 years, increasing to 20% in women over 80 years old.

4 Current management and care pathway

4.1 Identification, assessment and diagnosis of VFFs

VFFs can be identified when a person presents to a healthcare setting with symptoms suggestive of a VFF (Clinical Guidance for the Effective Identification of Vertebral Fractures, 2017). They can also be identified on DXA scans performed for bone densitometry as part of osteoporosis or secondary fracture prevention pathway. In addition, VFFs can be identified when a person has an imaging investigation involving the spine taken for reasons other than VFF detection. This is known as opportunistic detection. This can involve imaging for any reason and can be unrelated to the spine (for example, an X-ray for a chest malignancy), or further imaging of the spine (for example, DXA following a diagnosis of low bone mineral density). VFFs are most likely to be underreported on imaging obtained for non-musculoskeletal indications. Therefore, increasing the opportunistic detection of VFFs would be of most value.

In 2018/19, over 43 million radiological examinations were performed across the NHS in England (GIRFT Radiology Report, 2020). The specific imaging approach depends on the care setting, pathway and condition. NICE's guidelines on non-complex (NG38) and complex fractures (NG37) make recommendations on when different imaging modalities should be considered. Recommendations on imaging are also made in NICE's guidelines on suspected cancer (NG12), early and locally advanced breast cancer (NG101), advanced breast cancer (CG81), lung cancer (NG122), pancreatic cancer (NG85), oesophago-gastric cancer (NG83), prostate cancer (NG131), cancer of the upper aerodigestive tract (NG36), diverticular disease (NG147) and chronic obstructive pulmonary disease (NG115). Imaging can also be done for orthopaedic conditions.

There are several methods for the diagnosis and grading of VFFs. The semi-quantitative (Genant) and quantitative morphometric methods are primarily used in research, but they may also be applied in clinical practice. However, experienced radiologists will generally undertake a less formal visual read of the radiographic image.

The Royal College of Radiologists has made recommendations about the radiological reporting of VFFs. These include using standard terminology, indicating that the bones have been reviewed, if a VFF is diagnosed it, looking for and commenting on the presence of additional VFFs, the levels of fractures and their severity and if there is evidence of canal/cord/cauda equina compromise and discussing the reporting of VFFs at a departmental radiology events and learning meeting (REALM).

4.2 Treatment and management of VFFs

For people with osteoporosis and a new VFF confirmed on imaging, the clinician will assess using a recognised fracture risk assessment tool such as QFracture, FRAX or NOGG, urgent DXA and consider osteoporosis medication. NICE's guideline on Osteoporosis: assessing the risk of fragility fracture (CG146) recommends using either FRAX (without a bone mineral density value if a DXA scan has not previously been undertaken) or QFracture, within their allowed age ranges, to estimate 10-year predicted absolute fracture risk when assessing risk of fracture. CG146 is being updated to include up-to-date recommendations on risk assessment, as well as recommendations on the treatment and prevention of fragility fractures.

If a person has a new fracture while on osteoporosis treatment, they are referred to a metabolic/rheumatology clinic or local fracture liaison service (FLS). FLSs are specialised services that co-ordinate and deliver secondary fracture prevention through systematic identification, investigation, treatment recommendation and monitoring. These services are normally located within the acute hospital, community and primary care settings.

Management of VFFs is multidisciplinary and aims to reduce pain, restore mobility and minimise the incidence of new fractures. Conventional treatment plans include pain management with a range of pharmacological treatments, physical therapy including bed rest and supporting the spine to reduce the risk of further VFFs (Guidance for the management of symptomatic vertebral fragility fractures, 2022). Most people become symptom free through these measures and surgery is rarely indicated. However, if necessary, cement augmentation (percutaneous balloon kyphoplasty with or without stenting or vertebroplasty) or spinal fusion may be considered (IPG12, 2003, IPG166, 2006, TA279, 2013).

The Royal Osteoporosis Society recommends that people at high risk of fragility fracture are initiated on an appropriate osteoporosis drug treatment within 16 weeks of fracture diagnosis (i.e. within 4 weeks of the assessment being completed). People who are recommended interventions to reduce risk of fracture will be reviewed by the FLS within 16 weeks of fracture and at 52 weeks.

4.3 Position of AI technologies in the care pathway

The AI technologies described in section 2.2 can be used for the opportunistic detection of VFFs on images involving the spine taken for reasons other than VFF detection.

5 Comparator

The comparator is standard care where the radiologist or radiographer interprets the radiograph without AI assistance, usually within 24 hours of the image being taken. Retrospective assessment of radiographic images represents a form of screening and is outside the scope of this assessment.

The reference standard or assessment of ground truth is based on the consultant radiologist or reporting radiographer’s (ideally with specialist training in musculoskeletal imaging) interpretation and report. Although considered the reference standard, fracture detection by a radiologist or reporting radiographer is not 100% accurate as fractures may still be missed.

6 Decision problem

Decision question	Does the use of software with artificial intelligence (AI) derived algorithms for analysing radiographic images as an aid for the opportunistic detection of vertebral fragility fractures (VFFs) have the potential to be clinically and cost-effective to the NHS?
Population	People who have had a radiographic image involving the spine taken for reasons other than VFF detection
Subgroups	Depending on the availability of evidence, the following subpopulations may be included: • People aged 50 or older • People with osteoporosis or those who are at risk of osteoporosis • People with osteogenesis imperfecta • People with cancer.
Intervention	AI used as a decision aid for radiographic image interpretation and fracture assessment prior to radiology review, using any of the following softwares: • Annalise Enterprise (CXR) (Annalise.AI) • Annalise Container (CXR) (Annalise.AI) • BoneView (Gleamer) • BriefCase-Triage (Aidoc Medical) • CINA-VCF Quantix (Avicenna.AI) • HealthVCF (Nanox AI) • HealthOST (Nanox AI) • IB Lab FLAMINGO (IB Lab) • TechCare Spine (Milvue).
Comparator(s)	Radiologist or radiographer interpretation of the radiographic image without AI assistance.
Healthcare setting	Secondary care and community diagnostic centres
Outcomes	The outcome measures to consider include: **Intermediate outcomes** • Measures of diagnostic accuracy to detect VFFs • Accuracy when used by different healthcare professionals (radiologists, radiographers and other healthcare professionals) • Failure rate or rate of inconclusive AI reports • Number of missed fractures • Rate of missed fracture-related further injury • Proportion of people that need further imaging • Intervention related adverse events • Healthcare professional user acceptability of AI tools

	• Changes to clinical management. **Patient-reported outcomes** • Health-related quality of life. **Costs and resource use** • Cost of the AI software • Staff costs • Training and implementation costs • Other downstream costs for diagnosis or treatment • Time to produce a radiography report • Time to diagnosis or time to definitive radiology report • Time to further referral or treatment • Number of treatments and extent of treatments • Number of hospital appointment/visits, including referrals to fracture clinics and orthopaedic assessment • Number of hospital admissions • Type of healthcare professional interpreting the radiograph.

6.1 Patient issues and considerations

People may have concerns about the use of or sharing of their data when AI technologies are involved, and the need for consent. They value being well-informed about the decision-making process, including AI involvement and whether the clinician agrees with the AI output. People may also have concerns about whether their radiographic image will still be interpreted by a qualified clinician when AI is in use.

People may experience additional stress and require additional time to discuss the results of their scan if they receive a diagnosis with a VFF which has been identified opportunistically, because this will not have been the primary reason for imaging. Many of those people are likely to have other conditions as well.

6.2 Implementation issues

IT and workforce

The AI technologies will need to integrate into existing hospital PACS systems to ensure there is no disruption or delays to the workflow and will need to ensure compliance with cyber risk, information governance and data protection law. Some of the AI technologies require access to the internet for installation and/or running. All require hardware to run on if they are not available as a cloud-based option. There may be differences between trusts in the proportion of radiographic images interpreted by radiologists compared with radiographers or other staff. Companies have suggested that radiographs can also be interpreted by other health professionals, such as nurses, after appropriate training. However, clinical experts advised that this may lead to overdiagnosis and overtreatment. Increased detection of VFFs may temporarily or permanently increase the workload of radiologists and other clinicians.

The GIRFT Radiology Report highlights that many departments currently struggle to meet what might be deemed the minimum IT requirements for a modern radiology service. In addition, it notes that when considering the impact on staff, it should be noted that although there has been an overall growth in the demand for radiology services, the fastest increase in demand is for more complex imaging modalities such as CT and MRI.

Clinical experts highlighted that in order to benefit from identifying more VFFs, effective mechanisms need to be in place to communicate findings, with a structure established to receive and act upon each radiology report. A 2019 audit found that in 95% of cases people could access an appropriate bone service. But, only 19% of radiology departments had a defined pathway for people with a vertebral fracture to FLS or osteoporosis service (Royal Osteoporosis Society, 2021). The audit also highlighted that offsite 'teleradiology' reporting services (outsourcing a proportion of their reporting of CT imaging) are commonplace in the NHS.

Clinical experts explained that implementation may be more difficult in smaller hospitals, which are likely to have less technical staff.

Diversity in the technologies included

Although the technologies included are all based on AI, there are differences in the image used by each technology. Also, some of the technologies offer additional functions, for example prioritisation of high-risk cases. The technologies might also differ by the populations indicated for use, such as restrictions to specific age groups.

Procurement

Procurement may differ between the technologies with companies offering various pricing options including annual subscriptions and pay per use. Smaller or rural centres that see less people and perform fewer X-rays/CT scans may not have a sufficient volume to justify the cost of an annual site licence. Any requirement for the use of specific multi-vendor platforms may limit which trusts can access specific technologies. Any additional, bespoke company software may also be a potential barrier to implementation and may increase the risk of vendor lock-in.

6.3 Equality issues and considerations

There are geographical inequalities related to radiology services and service capacity. AI technologies to aid the opportunistic detection of VFFs can improve accessibility and equity of access by providing diagnostic services in areas lacking specialised radiologists.

Some bone disorders (for example, scoliosis, ankylosing spondylitis, bridging osteophytes, Scheuermann’s disease and degenerative disc disease) might affect the performance of the technologies. Clinical experts explained that AI technologies may also be less effective when detecting VFFs on images of people who are very elderly.

The incidence of VFFs increases with age from around the age of 50 years. VFFs occur more commonly in women than men at all ages. Osteoporosis is also more common in people of lower socioeconomic status.

People with certain conditions that affect bone density like osteoporosis are more likely to have VFFs. Certain medications like glucocorticoids, which are taken for a number of long-term conditions, may cause a reduction in bone density leading to osteoporosis and subsequently VFFs.

Compared to other ethnic groups, white men and women are at an increased risk of fragility fractures. Also, some ethnic groups may be underrepresented in the population used to train AI to detect VFFs. This may result in the algorithm performing differently in ethnic groups in which it was not developed, trained or validated with. There is a lack of a race-specific reference standard for measurement of bone density, noting that variations exist in bone mineral density across various ethnic groups, which could be a contributory factor to the misdiagnosis of VFF. The effectiveness of DXA and CT also varies significantly across different racial and ethnic groups.

Age, sex, disability and ethnicity are protected characteristics under the Equality Act 2010.

6.4 Other issues for consideration

There are several national datasets, the most established of which is the Digital Imaging Dataset (DID), managed by NHS Digital. In addition, there is the National Imaging Data Collection (NIDC), information from which is visible on the Model Hospital website. These datasets may contain relevant real-world data for this assessment.

7 Authors

Ivan Maslyankov

Topic lead

Chidera Mark-Uchendu, Lakshmi Mandava

Topic analysts

Judith Shore

Topic adviser

Deonee Stanislaus

Project Manager

January 2025

Appendix A. Related guidance

- **Related Guidelines:**

  Osteoporosis: assessing the risk of fragility fracture (2017). Clinical guideline 146.

  Management of osteoporosis and the prevention of fragility fractures (2021). SIGN national clinical guideline 142.

- **Guidelines in development:**

  Osteoporosis: risk assessment, treatment, and fragility fracture prevention (update). Publication date to be confirmed.

- **Related Interventional Procedures Guidance:**

  Percutaneous vertebroplasty (2003). Interventional procedures guidance 12.

  Balloon kyphoplasty for vertebral compression fractures (2006). Interventional procedures guidance 166.

- **Related Technology Appraisals:**

  Percutaneous vertebroplasty and percutaneous balloon kyphoplasty for treating osteoporotic vertebral compression fractures (2013). Technology appraisal guidance 79.

- **Related Quality Standards:**

  Osteoporosis (2017). Quality standard 149.

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE

Early Value Assessment

HTE10059 Artificial intelligence technologies to aid the opportunistic detection of vertebral fragility fractures (VFF) on radiographic images

External Assessment Group report

Produced by: **Newcastle External Assessment Group**

Authors: **Kim Keltie**, Lead Healthcare Scientist, The Newcastle upon Tyne Hospitals NHS Foundation Trust (NuTH);

Rachel O'Leary, Clinical Scientist, NuTH;

Elliot Blacklock, EAG Administrator, NuTH;

Paula Leslie, Pre-registrant Clinical Scientist, NuTH;

Eibhlin O'Sullivan, Trainee Healthcare Scientist, NuTH;

Paola Cognigni, Lead Clinical Scientist, NuTH;

Emma Belilios, Project Co-ordinator and Business Manager, NuTH;

Oleta Williams, Research Assistant, Information Specialist, Newcastle University;

Fiona Beyer, Principal Research Associate, Newcastle University;

Najmeh Moradi, Research Associate, Health Economics, Newcastle University;

Hangjian Wu, Research Associate, Health Economics, Newcastle University;

Gurdeep S Sagoo, Health Economics Group Co-Lead, Newcastle University;

Luke Vale, Professor of Health Economics, London School of Hygiene and Tropical Medicine; NuTH;

Andrew Sims, EAG Director, NuTH

Correspondence to: Andrew Sims

Northern Medical Physics and Clinical Engineering, NMPCE
(Medical Physics, NCCC Level 2)
Freeman Hospital
Freeman Road,
High Heaton,
Newcastle upon Tyne
NE7 7DN
nuth.nmpce.hta@nhs.net

Date completed: 16/04/2025 (draft submitted 03/04/2025)

Contains confidential information: Yes

Number of attached appendices: 4

Purpose of the early value assessment report

The purpose of this external assessment report (EAR) by an external assessment group (EAG) for early value assessment is to review the evidence currently available for technologies within the decision problem and advise what further evidence should be collected to help inform future decisions on whether the technologies should be widely adopted in the NHS. NICE has commissioned this work and provided the template for the report. The report forms part of the papers considered by the Committee when it is making decisions about the early value assessment.

Declared interests of the authors

None.

Acknowledgements

Yuriy Arlachov, MSK Consultant Radiologist, Nottingham University Hospitals NHS Trust

Irene Baker; Lay Specialist Committee Member

Daniel David George Chappell, Clinical Projects Manager, Cambridge University Hospitals NHS Foundation Trust

Emma M Clark, Professor of Clinical Musculoskeletal Epidemiology, University of Bristol; Honorary Consultant Rheumatologist, North Bristol NHS Trust; Trustee of the Royal Osteoporosis Society

Lisa Field, Consultant Radiographer, The Mid Yorkshire NHS Trust

Lucy Ellis Funnell, Product Implementation Lead NHS England; Physiotherapist, Circle Health Group

Jennifer Head; Lay Specialist Committee Member

Tim Hoogenboom, Head of Research, Open Medical Ltd

Amanda Isaac, Musculoskeletal Diagnostic & Interventional Radiologist, Guy's and St Thomas' NHS Foundation Trust

Chuck Lam, Medical Student, University of Birmingham

Gerald Lip, Clinical Director at North East of Scotland Breast Screening Programme; Consultant Radiologist at NHS Grampian; President at British Society of Breast Radiology; Chair at AI Special Interest Group British Institute of Radiology

Robert Meertens, Medical Imaging Senior Lecturer, University of Exeter

Rob Milner, Consultant Radiographer, The Rotherham NHS Foundation Trust

Annalisa Occhipinti, Associate Professor of Research, Department of Computing & Games, Centre for Digital Innovation and School of Computing, Engineering and Digital Technologies, Teesside University (UK)

Morrison Osiezagha, Support Worker, Prime Life

Opinder Sahota, Consultant Geriatrician, QMC, Nottingham University Hospitals NHS Trust

Jordan Tsigarides, Senior Clinical Fellow in Rheumatology, Addenbrookes Hospital (Cambridge)

Dijana Vilic, Senior Clinical Scientist in Artificial Intelligence Guy's and St Thomas' NHS Foundation Trust

Shendrine Wijendra, Interventional radiographer, University Hospitals of Birmingham – Heartland hospital

David John Wilson, Consultant Interventional Musculoskeletal Radiologist, St Luke's Radiology Oxford Limited and Oxford University

With thanks to staff from the Newcastle upon Tyne Hospitals NHS Foundation Trust (NuTH) and Newcastle University, including:

- Michael Hutton, Head of Section (MR Physics), Clinical Scientist, NuTH;
- Edward Meinert, Professor of Digital Health and Clinical Artificial Intelligence, Newcastle University;
- Andrew Mkwashi, Senior Research Associate, Health Innovations (Medical Devices), NIHR Innovation Observatory, Newcastle University;
- Hayley Richardson, Clinical Coding Trainer, NuTH;

- Kevin Robson, Head of Section (Diagnostic Radiation Protection), NuTH;
- Alexander Self, Consultant Radiologist, NuTH;
- John Tuckett, Consultant Radiologist, NuTH.

Responsibility for report

The views expressed in this report are those of the authors and not those of NICE. Any errors are the responsibility of the authors.

Contents

Abbreviations

Term	Definition
A&E	Accident and Emergency
AI	Artificial Intelligence
AP	Anterior-posterior
AUROC	Area under the receiver operating characteristic curve
BMD	Bone mineral density
CADTH	Canadian Agency for Drugs and Technologies in Health
CCEMG	Campbell and Cochrane Economics Methods Group
CEA	Cost-effectiveness analysis
CI	Confidence interval
CXR	Chest radiography
DXA or DEXA	dual energy X-ray absorptiometry (bone density)
DICOM	Digital imaging and communications in medicine
DID	Diagnostic Imaging Dataset
DPIA	Data Protection Impact Assessment
DTAC	Digital Technologies Assessment Criteria
EAG	External Assessment Group
EPPI	Evidence for Policy and Practice Information Centre
EVA	Early value assessment
FFFAP	Falls and fragility fracture audit programme
FLS	Fracture Liaison Service
GBP	British pound sterling
HCP	Healthcare professional
HES	Hospital Episodes Statistics
HU	Hounsfield units
HRQoL	Health-related quality of life
ICB	Integrated Care Board
ICER	Incremental cost-effectiveness ratio
IFU	Instructions for use
L1	First lumbar vertebra
L4	Fourth lumbar vertebra
L5	Fifth lumbar vertebra
LAT	Lateral
LoA	Limits of agreement

Term	Definition
MHRA	Medicines & Healthcare products Regulatory Agency
MSK	Musculoskeletal
NPV	Negative predictive value
NR	Not reported
OVCF	Osteoporotic vertebral compression fracture
PA	Posterior-anterior
PACS	Picture archiving and communication system
PARD	Public access registration database
PET	Positron emission tomography
PPV	Positive predictive value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSS	Personal social services
QALY	Quality-adjusted life year
RCR	Royal College of Radiologists
RCT	Randomised controlled trial
RFI	Request for information
RIS	Radiology Information Systems
SD	Standard deviation
SLA	Service level agreement
SoC	Standard of care
T1	First thoracic vertebra
VCF	Vertebral compression fractures
VF	Vertebral fracture
VFF	Vertebral fragility fractures
VHL	Vertebral height loss
QoL	Quality of life

Executive summary

Background and objectives: Vertebral fractures considered within this assessment are those resulting in a compressed vertebrae; this could be due to congenital, developmental, high-energy trauma or minimal trauma. They may be caused by osteoporosis (due to bone weakness), primary tumour or associated with malignancy. Vertebral fractures are a strong predictor of future fractures, and are associated with increased morbidity, mortality and healthcare costs. Artificial intelligence (AI) algorithms have been developed to support clinicians to opportunistically detect vertebral fractures, that is when reviewing diagnostic images for purposes other than investigation of vertebral fractures. The AI technologies in isolation are not fulfilling the role of diagnostic devices but are systematically searching for and alerting clinicians to incidental evidence for vertebral fractures; all require human interpretation as per their instructions for use. The premise is that detecting vertebral fractures earlier with AI technologies would lead to prompt management and treatment and that acting on incidental findings would be on average beneficial. The long-term impact could be higher quality of life for patients and lower subsequent healthcare resource use (related to further fractures, and GP and hospital visits associated with missed diagnoses).

The purpose of this early value assessment is to identify evidence for 9 AI technologies from 7 manufacturers used in the opportunistic detection of vertebral fractures when compared with standard care. The aims are then to identify evidence gaps to help direct further research and data collection. An early economic model has been created to determine the potential value proposition for these technologies in the NHS and explore key drivers and uncertainties to inform future economic evaluations.

Clinical evidence: The EAG conducted a literature search and reviewed evidence submitted by the companies and Clinical Experts resulting in 22 relevant studies. For technologies applied to X-rays this included 5 Annalise CXR, 1 BoneView, 0 TechCare Spine, and for technologies applied to CT images this included 8 HealthVCF, 3 CINA-VCF, 4 IB Lab, 1 BriefCase-Triage). Three studies were conducted in a UK setting, 5 were available as abstracts only, 1 poster only, 2 provided as academic in confidence and 1 as commercial in confidence by the companies. The configuration details and version of software used in the studies were generally poorly reported. Information regarding AI algorithm training and validation was only available for some technologies.

Most of the evidence described retrospectively applying the AI technologies to previously acquired images and then comparing with a reference standard (clinical expert reviewing the image specifically looking for vertebral fractures); however, the expertise and role of the individuals included as the reference standard varied across studies. Eleven studies reported diagnostic accuracy against a comparator (which was typically comparing against the original radiology report). The evidence generally demonstrated that AI technologies detect more vertebral fractures but can also misidentify anatomical features as vertebral fractures. The EAG considered the waiting lists for dual energy X-ray absorptiometry (DEXA) (bone density) scans and general radiologist staff shortages, and therefore the impact of false positives when implementing AI technologies on healthcare resources, workforce and waiting times needs careful consideration.

Economic evidence: The EAG identified no available economic evaluations for the technologies included in this assessment, but did identify one cost effectiveness study on an unnamed technology. The EAG used this, alongside the identified clinical evidence, to develop a de novo early decision tree economic model in *rdecision* (R programming language) which focused on the diagnostic pathway over a time horizon of 1 year. The model showed that using AI technologies for opportunistic detection of vertebral fractures could represent value for money for the NHS at a willingness to pay threshold of £20,000 per quality-adjusted life year (QALY) for some scenarios (a generic AI costing £7.36 per scan result in an ICER of £22,085). Increasing the sensitivity of standard of care (SoC) to 60% (base case was 25%) led to it dominating the AI technology as the sensitivity of the AI technologies would not be less than this but they would cost more. The results were also sensitive to utility gain from detecting and treating a vertebral fracture. If the utility gain associated with detecting and treating the vertebral fracture was reduced to 0.0265 over 1 year (a more conservative estimate than the base case value of 0.265 which was obtained from a symptomatic patient cohort), the ICER was £220,846 demonstrating AI was not cost effective. Further univariate sensitivity analysis found the model to be insensitive (across the range of values explored) for a number of parameters including failure rate (that is failure of the AI to process or analyse the image), prevalence and time to review the AI output. Uncertainties remain around how the eligibility criteria for each technology (including patient age, image modality, anatomical region scanned, X-ray projection where applicable), would influence uptake and cost-effectiveness, diagnostic accuracy of the technologies and any differences in this across the population in scope or between imaging modalities, and the downstream impact on other NHS services, such as DEXA scans and Fracture Liaison Services (FLS), for management of vertebral fracture detected by AI technologies.

Key points for decision makers:

- The value proposition of incidental detection of vertebral fractures proposed is that earlier detection and initiation of treatment could reduce further fractures, improve patient quality of life and reduce healthcare resource use.
- Some technologies had little (BoneView, BriefCase-Triage) or no (TechCare Spine) published evidence identified. There is a general lack of evidence from a UK setting across all technologies.
- Several studies were identified that reported diagnostic accuracy against a reference standard (the EAG considered that a retrospective study design was suitable for this comparison). The instructions for use for each AI-technology are that there is always a need for human interpretation that is AI-technology plus clinician. Even where not explicitly stated the EAG has assumed that the technologies are used in a way consistent with their instructions for use. The Royal College of Radiologists guidance for the recognition and reporting of osteoporotic vertebral fractures stated that radiologists and reporting radiographers are well placed to diagnose

VFFs on any imaging modality that includes the spine. The Clinical Experts advised that diagnosis of vertebral fracture would be conducted by a radiologist or reporting radiographer as standard practice in the NHS (in line with RCR guidance). Therefore, the EAG felt that comparison to a reference standard was appropriate. The clinical evidence suggests that specificity of technologies is generally high, however the diagnostic accuracy varies by technology, image modality and anatomical region of image.

- Limited prospective evidence was identified therefore a lack of evidence related to changes in clinical management and health-related quality of life attributable to implementation of the AI remains.
- The consent process for patients' diagnostic images to be processed (by an AI technology) for reasons other than their direct care, including transparently reporting how their data will be used and by who, is regulated by the Information Commissioners Office.
- The EAG notes that there is evidence from other fields (cancer detection) that a strategy to act on incidental findings increases healthcare interactions (including additional exposures to ionising radiation) and may not always yield expected benefits.
- Transparent reporting of software version, patient and diagnostic image characteristics used in training and validation, as well as configuration settings are required for future studies as diagnostic accuracy may vary by version.
- The implementation of AI requires consideration where new versions require training and validation.
- AI technologies show promise for cost-effective opportunistic detection of vertebral fracture, but key uncertainties should be explored further including diagnostic accuracy, cost of implementation and downstream impact on NHS services for management.

1. Decision problem

The decision problem is described in the Final Scope and EAG comments are included in the protocol. The EAG made no further changes.

Terminology

Multiple Clinical Experts noted that the term 'vertebral fragility fracture (VFF)' denoted the presence of an underlying weakened bone state secondary to conditions such as osteoporosis or osteopenia (where bone density scans show lower bone density than average for age, but not low enough to be classed as osteoporosis) (NHS, 2025). Experts also noted that the term 'vertebral compression fracture (VCF)' denoted the process of the vertebral body collapsing or reducing in height, which may be due to osteoporosis or other pathologies such as infection, cancer, trauma.

The EAG noted that terminology varied across the published literature and included: vertebral fragility fracture, vertebral compression fracture, loss of vertebral height, wedging, depression of the superior endplate, vertebral collapse, crush fracture, biconcave fracture, insufficiency fracture, frailty fracture, low-energy fracture, low-trauma fracture, stress fracture, spontaneous fracture and osteoporotic fracture (in cases where the underlying cause is considered related to a diagnosis of osteoporosis). Due to the limited evidence base the EAG considered all evidence which reported VFFs or VCFs, retaining the terminology used within each publication. However, because the detection of fracture does not determine aetiology, and to align with the latest Royal College of Radiologists guidance, the EAG has referred to 'vertebral fractures' throughout this report where possible. The EAG has referred to 'VFF' or 'VCF' only when specifically stated in the company Instructions of Use and when referring to the outcomes as defined in the published evidence.

The EAG highlighted that the Genant classification system (grade 1, 2 or 3) is the most widely used technique for systematic vertebral fracture definition (The Royal College of Radiologists, 2021). It is a semi-quantitative method which involves the visual recognition of a loss of vertebral body height on a lateral projection combined with assessment of the vertebral endplates to diagnose a fracture and to exclude non-fracture vertebral deformity:

- Grade 1 (Mild) fracture 20-25% loss of vertebral body height.
- Grade 2 (Moderate) fracture 25-40% loss of vertebral body height.
- Grade 3 (Severe) fracture >40% loss of vertebral body height.

One Clinical Expert highlighted that the AI technologies will be able to calculate the degree of compression on a numeric scale.

The EAG notes that dual energy X-ray absorptiometry (bone density) is referred to as DXA or DEXA within the published literature, however for the remainder of this report the EAG has referred to this as DEXA for consistency. Mention of radiographer throughout this report should be interpreted as diagnostic radiographer.

2. Technologies

A brief summary of the nine artificial intelligence (AI) technologies from seven manufacturers included in the early value assessment is included in Table 1. The EAG notes that Gleamer did not provide a response to the standard request for information (RFI) document for the BoneView technology for this topic; therefore, information for that technology was obtained from the Scope and from information supplied as part of the early value assessment on AI to detect fractures on X-rays in urgent care (HTE20, January 2025). Four technologies from three manufacturers process X-ray images and five technologies from four manufacturers process CT images.

As of March 2025, five of the nine technologies had regulatory approval (four as class IIa and one as class IIb medical devices under either the EU Council Directive 93/42/EEC or EU Regulation 2017/745). One of these technologies was also certified as a class I medical device under the UK Medical Device Regulation 2002. The remaining technologies are in the process of obtaining regulatory approval. Five of the nine technologies were registered on the Public Access Registration Database (PARD).

Three companies (Aidoc Medical, Annalise.AI and Gleamer) stated they meet the Digital technology assessment criteria (DTAC) and one company (Milvue) advised that DTAC evaluation was in progress. Two companies (Avicenna and Nanox AI) advised they do not have DTAC with Avicenna stating the reason as the technology is not yet in use within the UK. One company (IB Lab) noted that DTAC has not been performed by UK bodies, however the technical file has been audited by their notified body and that the absence of DTAC does not mean the product is not compliant with those standards. These declarations were supplied as part of company RFI responses.

Two local AI experts advised the EAG of the following in relation to regulations and best practices concerning the use of AI products for medical or clinical use:

- AI technologies intended for medical or clinical use must disclose their training dataset, information workflow and validation approach as part of regulatory compliance.
- AI technologies will be required to demonstrate they meet the DTAC, and the supplier should expect to provide a completed Data Protection Impact Assessment (DPIA).
- Significant changes to the algorithm that could affect clinical impact, patient safety or change their regulatory classification require either a new submission or variation of the regulatory approval. All model changes require performance validation and clinical risk assessment.

For more detailed information see Appendix D4.

From information provided by companies and from company websites, the EAG notes that all technologies included in this assessment:

- Involve a fixed AI algorithm (that is, an artificial intelligence or deep learning algorithm which has been reviewed by a notified body and released as a commercial product). Any further updates to this “fixed” state would require a review by the notified body. Typically, fixed algorithms do not learn or adapt to data that it processes during commercial use.
- Require internet access.
- Are to aid the clinician in reporting, that is, they will not be used autonomously without human interpretation.
- Use radiographs in DICOM (digital imaging and communications in medicine) format. The DICOM protocol provides a standardised method for inter-operability between different imaging devices and information systems understood without compatibility issues. DICOM also contains information about the patient, the type of scan, the imaging device and image parameters. DICOM images are stored on picture archiving and communications systems (PACS) which is a high-speed network for the storage, recovery and display or radiological images such as ultrasound, X-ray, CT and MRI. Some technologies refer to secondary DICOM

captures which is a DICOM image created from a non-DICOM image format.

Each technology reports findings and images in a different manner as summarised below:

- Annalise Enterprise CXR reports results in a desktop viewer application which synchronises with PACS and displays findings and localisation of the fracture. AI outputs from Annalise Enterprise CXR are also available as triage notifications, or DICOM secondary capture images inserted into the site PACS, which displays findings and localisation. Annalise Container CXR (a version of Enterprise which is packaged to optimise it for running on selected third-party marketplaces or platforms) outputs notifications and secondary capture images. The company has confirmed that Annalise Container utilises the same CXR AI model as Annalise Enterprise, and that the evidence should be considered generalisable between technologies.
- BoneView reports results in DICOM secondary captures with boundary boxes around abnormalities with a summary sheet. The EAG assumes that this means that AI annotation can be overlayed on the image as well as a report table, avoiding the need for the AI tool to have its own user interface.
- TechCare Spine reports 2D measurements of vertebrae deformation and height loss alongside labelling of each vertebra. Report flags potential fractures based on Genant classifications.
- BriefCase-Triage displays the analysed images in a standalone application separate from the usual standard of care platform. The user is notified of suspected findings with a notification and with low quality preview image when the user hovers over the notification to aid in triage.
- CINA-VCF Quantix provides annotated images to show vertebral labelling, vertebral height loss ratio and mean Hounsfield Unit measurements. A notification is sent if the height loss ratio exceeds 20% or 25% threshold.

- IB Lab FLAMINGO labels thoracic and lumbar vertebrae and identifies presence or absence of vertebral fracture in a report that includes a summary, overlay image and manufacturer information. Labels of the fractured vertebrae are coloured orange in the report.
- HealthVCF highlights compression fractures in the thoracic and lumbar spine CT scans with a boundary box on the affected vertebral body. HealthOST assesses T1-L4 the percentage compression deformity of each compression fracture highlighting those with 25% or more compression in yellow. The EAG notes that one publication (Page et al., 2023), using the predecessor (Zebra Medical Vision) did not locate the fracture but instead gave a binary output ('fracture' or 'no fracture' detected); therefore differences in functionality may occur between versions.

Each technology has different functionality for configuring the sensitivity and specificity of the AI. Three companies reported that the AI configuration can be adjusted at both installation and during ongoing use (Aidoc Medical, Annalise.AI and Nanox AI), one company reported that their technology was configurable only at installation (Avicenna.AI), one company advised that their technology is not configurable at installation or a per image basis (IB Lab). Two companies did not report whether their technology is configurable or not (Gleamer, Milvue).

Each company was sent a series of questions by the EAG to confirm specifics surrounding their technologies in particular AI training data, how the technology filters image types for suitability, how the technology integrates into the pre-existing radiology workflow in addition to confirmation on costs. Five companies responded to questions (for summary of these responses see Appendix C1).

The instructions for use of each AI technology are that there is a need for human interpretation – that is all should be considered as AI plus clinician. The EAG has assumed that that all technologies are used in way that is consistent with their indications for use.

Table 1: Description of AI technologies

Device (Company) [Previous Name]	Type of image	Compatible imaging	Exclusions (obtained from indications or contraindications in RFI or IFU)	Deployment Method	How are finding displayed	Where are the findings displayed	Additional features (as claimed by company)	Used in the NHS
Annalise Enterprise CXR and Annalise Container CXR, (Annalise.AI)	X-ray	CXR, anterior-posterior (AP) or posterior-anterior (PA) and optionally lateral (LAT) orientations	Is not to be used on patients under the age of 16 years for CXR.	Local or Cloud Enterprise supports hybrid local-cloud deployment Container hosted on partner platform – local or cloud	Annotated images with optional notification	Annalise Enterprise: Synchronised Desktop app and/or DICOM Secondary Captures. Annalise Container: DICOM Secondary Captures	Triage / prioritisation notifications to PACS or RIS reporting worklist. Device can also detect 123 additional findings on CXR, including spine related findings such as diffuse spinal osteophytes, kyphosis, osteopaenia, scoliosis, spinal fixation, spinal arthritis, spine lesion, and technical factors which may indicate poor image quality.	Yes (48 Trusts)
BoneView, (Gleamer)	X-ray	Appendicular skeleton, ribs and thoracic-lumbar spine	Unsuitable for patients below 2 years old. [redacted]	Local or cloud	DICOM Secondary capture with bounding boxes	DICOM Secondary Captures	Can detect fractures, dislocations, effusions and bone lesions	Yes
TechCare Spine, (Milvue)	Lateral X-ray	Thoracic or lumbar spine lateral views	VFF detection will not function with Frontal X-rays or Cervical spine.	Cloud	Annotated images with a notification	DICOM Secondary Captures	[redacted]	No
BriefCase-Triage, (Aidoc Medical)	CT	Chest, Abdominal	All studies that are technically inadequate, including motion artifacts, severe metal artifacts, suboptimal bolus timing or an inadequate field of view. Is not to be used on patients under the age of 18 years old.	Hybrid Local and Cloud	Annotated images	Standalone Desktop App	Image prioritisation options, can be set up to analyse only a subset of patient classes, depending on the site's preference and clinical workflow	Yes (1 centre)
CINA-VCF Quantix [Previously CINA-VCF] (Avicenna.AI)	CT	Chest, Abdominal	Images need to include at least three consecutive vertebrae in the T1-L5 portion of the spine without cement or surgical hardware. Is not to be used on patients under the age of 50 years old	Local or Cloud	Annotated images with a notification	PACS Viewer	NR (the EAG notes that vertebral height loss and Hounsfield Units as a measure of vertebral bone mineral density are also reported as outcomes of this technology some published evidence).	No
HealthVCF and HealthOST (Nanox AI) [Previously Zebra Medical]	CT	Chest, Abdominal Pelvic showing T1-L5	Is not to be used on patients under the age of 50 years old. Should not be used to analyse scans that are not from computed tomography (CT) modality. Should not be used to analyse scans that don't contain at least 4 complete vertebrae between T1-L4. Should not be used to analyse CT scans which are not in a sagittal or axial orientation. Should not be used to analyse CT scans with slice thickness and slice increment above 5 mm (for sagittal scans) or 3mm (for axial scans); as well as studies with slice increment bigger than the slice thickness. Should not be used to analyse scans with less than 20 slices. Should not be used to analyse CT scans with inconsistent distances between slices. Should not be used to analyse CT scans with an inconsistent number of columns and rows within the slices. Should not be used to analyse CT Attenuation correction or PET series (a compatible series within the PET study will be analysed).	Cloud	Annotated images	Standalone Desktop App	Can be used both retrospectively on any data base or prospectively on all exams as they come into the PACS system [HealthOST: Analysis of low bone mineral density, labelling of vertebrae and measurement of the mean Hounsfield Units]	Yes [HealthOST: No]
IB Lab FLAMINGO (IB Lab) Powered by UCB's, BoneBot AI model	CT	Thoracic or lumbar spine	Use in adults aged 50 years and older. IB Lab FLAMINGO is not indicated for use on patients with spinal metalwork. The software should only be used with medical-grade quality CT scans that meet the specific image requirements outlined in the instructions for use. The software should not be used on images with slice thickness greater than 3 mm. The user must ensure that ImageOrientationPatient is set in the DICOM tag (0020,0037) and that PixelSpacing is set in the DICOM tag (0028,0030) or in DICOM tag (0018,1164). The CT scan must contain at least two visible vertebrae. The device should not be used without proper training.	Local	Annotated images	PACS Viewer. No additional software needed	Image prioritisation based on risk	No

Abbreviations: AI, Artificial Intelligence; AP, Anterior-posterior; CXR, Chest radiography; DICOM, Digital Imaging and Communications in Medicine; L4, fourth lumbar vertebra; L5, fifth lumbar vertebra; LAT, Lateral; NR, Not Reported; PA, posterior-anterior; PACS, Picture archiving and communication system; PET, Positron emission tomography; RIS, Radiology Information Systems; T1, first thoracic vertebra; VFF, Vertebral fragility fracture.

3. Clinical context

For a description of the care pathway including diagnosis and management of vertebral fractures please refer to the [Final Protocol].

The prevalence of vertebral fractures is currently uncertain due to most being asymptomatic in nature. Clinical Experts estimated that approximately 50% to 70% of vertebral fractures remain undiagnosed due to lack of symptoms or self-management before presenting to a healthcare professional. Two Clinical Experts advised that having Osteoporosis Lead clinicians at centres has contributed to the improved reporting of vertebral fractures.

Multiple Clinical Experts advised that the ability for vertebral fractures to be opportunistically detected on images captured for indications other than spinal fractures (for example lateral chest X-ray) depends on the imaging parameters and image quality. One Clinical Expert additionally advised that lateral chest X-rays are not standard in many imaging departments, due to most patients being referred for CT. Multiple Clinical Experts advised that transparent reporting of how the AI technology has been developed, trained and validated would support adoption in the NHS.

Multiple Clinical Experts highlighted that AI could trigger alerts and reduce the number of missed vertebral fractures. Concern was raised about the downstream impact of such identification, especially the over-identification or false positives. As stated in the NICE Final Scope, ionising radiation (medical exposure) regulations (IRMER) state that clinical evaluation of X-rays requires a trained person, therefore AI outputs require human oversight. Therefore, vertebral fractures identified by AI will need additional radiologist or reporting radiographer review, may require additional imaging (and therefore may have an increased radiation dose) and may increase pressures on DEXA scan facilities. Differentiating vertebral fractures from non-fracture deformities (such as cupid’s bow or limbus vertebra which are developmental variants, Schmorl’s nodes or Scheuermann’s disease, H-shaped vertebrae associated with sickle cell disease or Gaucher’s disease) can be challenging (Lenchik et al., 2004). Several Clinical Experts highlighted that there is uncertainty regarding the ability of AI systems to differentiate vertebral fractures from non-fracture deformities, which could lead to false positive identification.

The time taken between image acquisition and interpretation varies depending on Trust protocols, clinical priority (urgent or routine), and access to hot reporting (for example accidents and emergency [A&E] referrals are recommended to be reported within 24 hours). AI technology may identify additional patients with vertebral fracture and add this data to the local radiology information system (RIS). Processes or protocols will need to be developed by Trusts to ensure radiological actionable reports and alerts are issued for these patients to access appropriate onward assessment and care. Ideally radiology reports should include a clear statement that there is a newly diagnosed vertebral fracture (which would require clinician review of patient notes including previous diagnostic images where available, and documentation of whether treatment is already being taken for prior vertebral fractures) and the referrer must ensure arrangement of appropriate assessment for osteoporosis and fragility fracture and subsequent treatment which has consequences for downstream services.

3.1 Royal College of Radiologists guidelines

In 2019 the Royal College of Radiologists (RCR) led a UK-wide audit which confirmed a lack of compliance with audit standards (derived from Royal Osteoporosis Society guidance, 2021) relating to the reporting of incidental VFFs on CT and onward referral mechanisms. Following this, in 2021 the RCR issued their Radiological guidance for the recognition and reporting of osteoporotic vertebral fragility fractures (The Royal College of Radiologists, 2021). This guidance reported that *“radiologists and reporting radiographers are well placed to diagnose VFFs on any imaging modality that includes the spine (including X-ray, barium studies, computed tomography [CT] and magnetic resonance imaging [MRI])”* and that “*the diagnosis of VFFs can also be made opportunistically where the spine is not the focus of the study*”.

The key points from the RCR guidance included:

- Appointment of a radiology osteoporosis lead is desirable to support development, delivery and audit of policy and protocol in the identification and reporting of fragility fractures (including vertebral fragility fractures [VFFs]) and to act as part of a multidisciplinary team (within a local fracture liaison service [FLS] if available).

- Agree local policies for the opportunistic reporting of VFFs from imaging that includes the spine.

- Agree local policy for adopting a consistent approach to the identification and reporting of VFFs.

- Implement a policy of automatic sagittal spinal reformatting, display and storage on PACS for cross-sectional imaging studies that include the spine.

- Consider a policy for template reporting of cross-sectional imaging studies that capture the spine, to include bone integrity, presence of VFF, level and grade or severity.

- Implement a policy for standardised use of terminology for VFFs – using the term ‘vertebral fracture’.

- Implement routine audit processes around identification and reporting of VFFs.

- Agree local policy for onward alerting of referrers or referral to fracture prevention pathways.

- Agree service level agreements (SLA) with teleradiology contractors to adopt and adhere to VFF-reporting policies.

- Consider the use of standard phrases or short codes to create actionable reports or alerts to referrers. [One Clinical Expert highlighted that variation in coding practices is a particular problem when teleradiology reporting is used].

- Where artificial intelligence (AI) technology is implemented to actively screen cross-sectional imaging, agree processes and policy for radiological correlation and alert reporting.

- Discuss and agree the use of alerts, report content and automisation and clarify onward referral pathways. Clinical engagement is beneficial. Severe or multiple VFFs or cases with canal compromise will warrant urgent clinical evaluation.

In 2023 the RCR conducted a UK-wide re-audit to assess the impact of a series of RCR interventions (including publication of the initial audit, the above mentioned guidance, and webinar) on VFF awareness, reporting practice and patient outcomes (Howlett et al., 2023). The first component of re-audit included an online questionnaire that assessed organisational aspects relating to reporting provision, infrastructure and onward referral systems which were recommended in the RCR guidance. Key results of this component of the re-audit included:

- 46% (n=55) of responding departments had an alert policy in place that included VFFs (increased from 26% in 2019). Alert notifications to the referrer were facilitated via an embedded radiology information system or picture archiving communication system (RIS or PACS) and 49% of departments had incorporated this technology in 2022.
- Several respondents commented in free-text fields that teleradiologists (radiologists who review images for primary reports, second opinions, clinical review, who can be based in the same or different organisation, and in the same or different international location) are not able to routinely access alerts.
- Most departments did not review or audit terminology used by reporters.
- FLS notification occurred in 47% of departments (increased from 27% in 2019). Radiology access to FLS increased to 44% (from 37% in 2019) but access to an appropriate bone service decreased to 79% (from 95% in 2019).

- 11% of departments had appointed a radiology lead for osteoporosis and 38% of departments agreed a local policy for VFF reporting.
- 17% of departments reviewed the 2021 RCR VFF reporting guidance in their local governance meeting. 82% of departments had already or were very likely or likely to change practice as a result of the RCR guidance or 2022 audit.
- Only 1% of departments had implemented an artificial intelligence (AI) solution to screen cross-sectional imaging for VFFs, 13% were exploring the possibility.

The second component included an audit of consecutive CTs that included the thoracolumbar spine (CT chest, abdomen, pelvis). Of 194 eligible departments 129 (66%) submitted patient data from 7,316 patients (aged 70 years and older). Following auditor review, 21.7% of images (1586/7316) had a vertebral fracture detected. From the primary reports:

- 50.4% (799/1586) had the fracture mentioned in the original report (absolute increase of 5.2% from 2019).
- 67.5% (539/799) included the term “vertebral fracture” (absolute increase of 7.4% from 2019).
- 34.4% (275/799) commented on fracture severity (absolute increase of 8.3% from 2019), and the auditor agreed with the severity reported in 85.8% of cases.

The proportion of CT reports which included the terminology of “vertebral fracture” and reporting of severity of fracture increased further when addendums to the original reports were considered by the auditors.

3.2 NICE Clinical Guidelines

The use of AI technology is not currently mentioned in the NICE Clinical Guideline on Osteoporosis: assessing the risk of fragility fracture (CG146, updated 2017). However, the EAG notes that NICE guidance on osteoporosis: risk assessment, treatment and fragility fracture prevention (GID-NG10216) is currently in development.

3.3 Routinely collected data in the NHS

A number of routine sources collect data that may be relevant to the scope of this EVA and may provide clinical context. However, they all have limitations and may not fully address the decision problem.

3.3.1 Fracture Liaison Services (FLS)

One Clinical Expert (Appendix D1) highlighted the Falls and Fragility Fracture Audit Programme (FFFAP) from the FLS database which is a national clinical audit commissioned by the Healthcare Quality Improvement Partnership and managed by the Royal College of Physicians. While this database is not specific to opportunistic detection of vertebral fracture, its results from 2024 demonstrate:

- 29.5% of 74,767 submitted records reported spine fractures (Key Performance Indicator 3 on the database) nationally; an increase from 4.5% (of 43,578 records submitted) in 2016.
- variation in reporting of spine fracture between 9.2% (of 1630 records) in Northern Ireland, to 40.4% (of 11,548 records) from South West England.
- 44.2% of patients nationally received a DEXA scan within 90 days (Key Performance Indicator 5).

3.3.2 The Model Hospital

The Model Hospital (part of the Model Health System) reports on the proportion of people aged 50 or above with osteoporosis, and a fragility fracture in the primary care Quality Outcomes Framework.

[redacted]

3.3.3 Diagnostic Imaging Dataset (DID)

The Diagnostic Imaging Dataset (DID) collects data from local Radiology Information Systems (RIS) on imaging of NHS patients in England. Data are collated from RIS and uploaded into a database maintained by NHS England. In the 2023-2024 financial year, 7,681,150 CT scans were taken (36.8% in an outpatient setting, 32.6% in A&E, 22.3% inpatient, 6.0% GP direct access, 2.3% other). For the same period there were 22,576,650 X-rays (34.4% in A&E, 25.9% GP direct access, 21.4% in outpatients, 16.1% inpatients, 2.3% other). Counts by body site can be retrieved but are limited to body sites which may have been used to diagnose or rule out cancer (DID Annual Statistical Release). In the 2023/24 annual statistical report NHS England acknowledge that images are used for wider clinical uses, and that it is not possible to distinguish between the different uses of diagnostic imaging within the DID dataset. Relevant imaging includes 8,194,405 chest X-rays (36.3% of all X-rays), and 806,625 chest or abdominal CT (10.5% of all CT scans). Looking at the reported counts of imaging activity using groups of tests suitable for diagnosing cancer labelled by body site (DID, Table 4; retrieved on 03 April 2025) the median ([Q1:Q3]; min, max) scan volume across NHS hospitals in England for the month of April 2024 were 5,412 ([3,385:7,335]; 90 to 17,825) for chest X-rays, and 546 ([310:690]; 5 to 2,390) for chest and abdominal CTs per site, which scaled to a full 12 months would be approximately median 65,000 ([40,000:88,000]; 1,000 to 215,000) chest X-rays and 6,500 ([3,700:8,300]; 100 to 29,000) chest and abdominal CTs. The EAG notes that the number of scans eligible for AI technologies is likely higher than this (given imaging of other anatomical areas involving the spine); however difficult to determine the exact volume given the different eligibility criteria of each technology. No additional information is available on X-ray projection (posterior anterior, anterior posterior or lateral view). The dataset includes diagnostic imaging scans on patients of all ages, and cannot be stratified into the specific age groups that individual technologies may be indicated for use in.

The National Imaging Data Collection also collects data on imaging, but its focus is operational data such as number of machines in operation across the

NHS, staffing and resources, rather than a breakdown of patient activity as reported in the DID. This dataset highlights that within the 2023/24 financial year there were 2,919 X-ray, 723 CT and 146 DEXA scanners across England in the NHS. This highlights that any increased detection of vertebral fractures with AI technologies when applied to routine X-rays or CT scans, needs consideration if a proportion were to require a confirmatory DEXA scan, as this may increase NHS waiting lists further, and may result in a delay in the initiation of treatment for this group of people.

3.3.4 Hospital Episode Statistics (HES)

Hospital Admitted Patient Care Activity reports released by NHS Digital for 2023 to 2024 contain data relating to diagnoses and procedures that may be of relevance to this EVA. There were 14,460 finished consultant episodes with a primary diagnosis of osteoporosis with pathological fracture (ICD-10 code M80.0-9). The 4-character ICD-10 codes are not specific enough to indicate the site of the fracture, so this number includes non-vertebral fractures. Relevant ICD-10 codes for a primary diagnosis of vertebral fracture identified before a diagnosis of osteoporosis may include M48.4 for fatigue fracture of vertebra (n=810), M48.5 for collapsed vertebra not elsewhere classified (n=3,717), and M49.5 for collapsed vertebra in diseases classified elsewhere (n=590). These codes may include other types of vertebral fracture, such as stress fractures, and fractures caused by metastatic cancer. This dataset is an *admitted patient* care resource with procedure counts for CT and X-ray, this is likely to underestimate how many images are taken, when compared with the counts recorded in the DID.

Hospital Outpatient Activity reports for 2023 to 2024 include 1,218 primary diagnoses of osteoporosis with pathological fracture (ICD-10 code: M80.0-9), and 256 primary diagnoses of either fatigue fracture of vertebra (M48.4) or collapsed vertebra not elsewhere classified (M48.5). Imaging data is collected in this dataset, but it lacks the granularity needed to determine counts of CT scans or X-rays on specific body parts. For example, a CT scan of the abdomen and pelvis would use the OPCS procedure codes U21.2 (computed tomography not elsewhere classified), Y98.2 (radiology of two body areas), Z92.6 (abdomen), and O16.1 (pelvis), but this combination of codes is not reflected in annual reports available online. In 2023 to 2024, the Hospital Accident and Emergency Activity dataset recorded 51 attendances with a primary or secondary reason of fragility fracture (SNOMED CT code: 306171000000106, not specific to site), and no publicly available data relating to imaging was identified. The clinical coding team within the Newcastle upon Tyne Hospitals NHS Foundation Trust advised that they do not code imaging used in A&E and outpatient settings, which limits using this dataset for future research.

3.4 Equality issues

In addition to the equality considerations described in the equalities impact assessment and the Final Scope the EAG notes that the following contraindications of some technologies should be considered:

- The patient age as listed in the instructions for use varies across technologies.
- Some of the technologies advise that image sites must not contain cement, surgical hardware or spinal metalwork.
- AI technology may misidentify scans when large fields of view are needed, this could be more prevalent in patients with obesity. See section 5.2.2 for further clarification.

4. Clinical evidence

4.1 Search strategies and study selection

A search strategy was designed by the EAG information specialists using the original NICE scoping searches with additional candidate search terms identified from browsing database thesauri (for example, Medline MeSH and Embase EMTREE), and existing literature identified during the initial scoping searches. The searches aims were to find: 1) studies that explicitly named the AI technologies (or associated manufacturers) as listed in the Final Scope in the title or abstract of study records, and 2) studies about detection of vertebral fragility fractures that did not name one of the relevant technologies in the database record (but might in the full text). The final clinical effectiveness search strategy consisted of free-text search terms for each of the AI technologies and associated manufacturers combined with the Boolean operator OR to maximise sensitivity. The searches to find records explicitly naming AI technologies or manufacturers were run separately to allow for categorisation and tagging in Endnote, to enable prioritised screening (records mentioning ‘opportunistic’ terms were sifted first, then those including the technology name, followed by the rest). Economic studies were identified using the same core strategy used in the clinical effectiveness search, with an additional filter developed by the Canadian Agency for Drugs and Technologies in Health (CADTH), designed to identify economic evaluations and models, applied to the search strategy in MEDLINE and EMBASE (CADTH, 2016). An English language limit was applied to all results. Additional, targeted searches were run in Embase and Google Scholar to retrieve papers about long-term impact of treatment. See Appendix A1 for full search details.

4.2 Included and excluded studies

Clinical evidence records (N=1,120) were reviewed by title and abstract, of which 1057 were excluded by a single reviewer (KK). A random sample of 10% of the records (titles and abstracts) was checked by a second reviewer (PL). The remaining 63 records had full papers reviewed of which 18 were considered in scope by a single reviewer (KK; confirmed by a second reviewer PL); 1 of which reported on economic outcomes. Eight additional studies provided via other sources (by companies, Clinical Experts, scoping) were considered relevant to the decision. A total of 25 studies were considered relevant to the scope, of which the EAG excluded 1 duplicate and deprioritised 1 pre-print and 1 abstract (neither was conducted in a UK setting, and full publications or those with larger sample size were included for that specific technology); therefore a total of 22 clinical studies and 1 economic study were included, study selection described in Appendix A2.

A summary of the 22 studies included in the clinical evidence is presented in Table 2. The technology name and version were poorly reported across the included studies:

- Five used Annalise CXR (Enterprise explicitly mentioned in 2 studies, Annalise Enterprise CXR Trauma Triage in 1 study which the company confirmed was generalisable to the current version of the technology). The company also advised that Annalise Container is a version of Enterprise CXR which is packaged to optimise it for running on selected third-party marketplace / platforms, and therefore the company confirmed that the evidence should be considered generalisable between the two technologies.
- One used BoneView, which was applied to thoracic and lumbar spine X-rays.
- One used Aidoc Medical (assumed to be BriefCase-Triage although not explicitly reported).
- Three used CINA-VCF (none explicitly stating use of CINA-VCF Quantix which is currently not CE or UKCA marked).
- Eight used HealthVCF or an algorithm by Zebra Medical (which the company confirmed was generalisable to the current version of the technology; noting that Nanox AI acquired Zebra Medical Systems in 2021). No evidence was identified which explicitly stated use of HealthOST.
- Four used a technology by IB Lab including one using IB Lab FLAMINGO (commercial in confidence report), and the company confirmed that two used BoneBot (Nicolaes et al., 2024; Nicolaes et al., 2023); which the company confirmed was identical to the IB Lab FLAMINGO technology.
- No evidence was identified for TechCare Spine (by Milvue).

When extracting data from each included paper the EAG interpreted comparison of the AI to a “ground truth” to mean comparison to a clinical reference standard (for example a radiologist specialising in musculoskeletal imaging), and any description of the detection of vertebral fracture when compared with the original radiology report to mean comparator.

Studies did not always report whether the intervention was AI plus clinician. The indication for use for all the technologies make it clear that it there is always a need for human interpretation; that is, AI plus clinician. The discrepancy may be due to limited reporting in the identified evidence or that a study has consider the use of the AI technology outside its terms of use. It is unclear which of these may be the case and the EAG has assumed that all technologies used in a way consistent with their terms of use.

A total of 71 full papers were reviewed and subsequently excluded by the EAG, Appendix A5.

Table 2: Description of key studies (N=22) in the clinical evidence base

Technology (version)	Author (reference, year; page)	Study design (n=patients or scans); Country	Population characteristics	Comparator	Reference standard	Anatomical location and scan modality	Populations used to train AI (number and type) for details, where reported, see Appendix A4	Populations used to validate AI (number and type) for details, where reported, see Appendix A4	Outcomes
Annalise CXR	Talwar (RSNA, 2023; W5B-SPCH-2) *Abstract*	Retrospective diagnostic accuracy study (n=1,559 scans) Australia	NR (Age ≥18 years)	Initial radiology report	Radiologist	Chest X-ray	NR†	NR†	Diagnostic accuracy, Failure rate or inconclusive AI reports
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]†	[redacted]†	[redacted]
Annalise.AI	Frias (RSNA, 2023; W5B-SPCH-4) *Abstract*	Retrospective diagnostic accuracy study (n=3,760 patients) Country=NR	NR (Age >18 years)	Initial radiology report	Thoracic radiologist (1 of 2, blinded to Ai result)	Chest X-ray	NR†	NR†	Diagnostic accuracy
Annalise Enterprise CXR (v.1.2); modified version of the commercially available software used in study	Jones (BMJ Open, 2021; e052902)	Pilot (n=11 radiologists, 2972 cases) and post-study survey (n=10 radiologists) Australia	NR (Age ≥16 years)	NR	Three of seven diagnostic radiologists.	Chest X-ray (at least one frontal chest radiograph)	✓ †	✓ †	Healthcare professional user acceptability of AI tools
Annalise Enterprise CXR Triage Trauma (v 2.2.0)	Ghatak (J Am Coll Radiol, 2024; 220-229)	Retrospective diagnostic accuracy study (n=596 scans) US	Age, mean years: 67.0 Male: 41.2% Ethnicity: 6.2% Hispanic, 90.4% not Hispanic, 3.3% unavailable	Original clinical report	Two thoracic radiologists (arbitration by a third if required)	Chest X-ray (frontal and lateral)	✓†	✓†	Diagnostic accuracy, Failure rate or inconclusive AI reports
BoneView by Gleamer	Oppenheimer (Skeletal Radiol, 2024; 1563)	Retrospective diagnostic accuracy study (n=512 X-rays) Germany	Age, years: mean 67.5 (19 to 100) Male: 37.4% Ethnicity: NR	NR	Consensus of 2 radiologists with experience in MSK imaging	Thoracic and lumbar spine X-ray	NR	NR	Diagnostic accuracy, Failure rate or inconclusive AI reports
BriefCase-Triage, Aidoc Medical	Wiklund (J Bone Mineral Res, 2024; 1113-1119)	Retrospective diagnostic accuracy study (n=1,112) Sweden	With VCF Age, years: 76.2 Male: 47.1% Ethnicity: NR Without VCF	Original radiologist report	General radiologist reviewed CTs, all identified VCFs also reviewed by MSK radiologist	Abdominal CT	NR	NR	Diagnostic accuracy, Failure rate or inconclusive AI reports

Technology (version)	Author (reference, year; page)	Study design (n=patients or scans); Country	Population characteristics	Comparator	Reference standard	Anatomical location and scan modality	Populations used to train AI (number and type) for details, where reported, see Appendix A4	Populations used to validate AI (number and type) for details, where reported, see Appendix A4	Outcomes
			Age, years: 73.5 Male: 53.9% Ethnicity: NR						
[redacted]	Avicenna.AI [AiC] [redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
CINA-VCF Quantix (v0.60)	Guenoun (Clin Radiol, 2025; 106831)	Retrospective diagnostic accuracy study (n=100 patients) France	Age, mean years: 76.6 Male: 28% Ethnicity: NR	-	Two radiologists	Chest-abdominal-pelvis CT	✓	✓	Diagnostic accuracy
CINA-VCF (v1.0)	Dai (J Comput Assist Tomogr, 2025)	Retrospective diagnostic accuracy study applied to retrospective images (n=474 scans) US and France	Age, mean years: 72.1 Male: 47.7% Ethnicity: NR	Original clinical radiology report (available for subset only)	2 neuroradiologists (arbitration by a third)	Chest or abdomen, axial or sagittal acquisition CT	✓	Purpose of study	Diagnostic accuracy
HealthVCF (v5.1.1)	Bendtsen and Hitz (Calc Tissue Int, 2024; 468-479)	Prospective diagnostic accuracy study (n=10,012 scans; of which 1,000 used in validation, and 538 patients followed for 6 months) Denmark	NR (only reported for n=538 patients with 6 months follow-up)	Original radiology report	Specialised radiographers, and senior radiologist if inconclusive	Thorax and abdomen CT	NR	✓	Diagnostic accuracy, Change to clinical management, Time to run the software
HealthVCF	Pereira (Radiol Bras, 2024; e20230102)	Retrospective diagnostic accuracy study (n=964 scans) Country NR	Age, mean years: 70.2 Male: NR Ethnicity: NR	General radiologist (original radiology report)	Two specialists in MSK imaging	Chest and abdominal CT	NR	NR	Diagnostic accuracy, Failure rate or inconclusive AI reports
HealthVCF	Roux (Rheumatology, 2022; 3269-3278)	Retrospective diagnostic accuracy study (n=173,720 patients; subset of 500 for accuracy) France	Age, mean years: 73.2 years Male: NR Ethnicity: NR	-	Two experts (not further defined) for AI accuracy subset	Abdominal or lumbar CT	NR	NR	Diagnostic accuracy

Technology (version)	Author (reference, year; page)	Study design (n=patients or scans); Country	Population characteristics	Comparator	Reference standard	Anatomical location and scan modality	Populations used to train AI (number and type) for details, where reported, see Appendix A4	Populations used to validate AI (number and type) for details, where reported, see Appendix A4	Outcomes
HealthVCF	Connor (J Bone Min Res, 2024; 296-297) *Abstract*	Prospective diagnostic accuracy study(n=36,620); UK	NR	1 centre had synchronous augmented live reporting by radiologists, 3 centres had asynchronous local clinical confirmation	NR	Thoraco-lumbar CT	NR	NR	Diagnostic accuracy
HealthVCF	Chappell (2024) *Poster*	Retrospective diagnostic accuracy study (n=2,000 scans) UK (presumed from authorship and ADOPT study)	Age: NR Male: 50.3% Ethnicity: NR	Original radiology report	Clinician with local radiologist adjudication	CT involving spine (including abdomen and pelvis CT, pulmonary angiogram CT)	NR	NR	Diagnostic accuracy
Zebra Medical Vision*	Page (JBMR Plus, 2023; e10778)	Retrospective diagnostic accuracy study (n=1,200) US	Age, median years: 73 Male: 51% Ethnicity: NR	-	Two neuroradiologists	Chest, abdominal or pelvis CT	NR	Purpose of study	Diagnostic accuracy, Failure rate or inconclusive AI reports
Zebra*	Kolanu (Journal of Bone and Mineral Research, 2020; 2307-2312)	Retrospective diagnostic accuracy study (n=2,357) Australia	NR (Age >50 years)	Original radiology report	Radiologist	Abdomen, thorax CT	NR	NR	Diagnostic accuracy, Failure rate or inconclusive AI reports
Zebra*	Connacher (Osteoporosis Int, 2019; S428) *Abstract*	Prospective diagnostic accuracy study (n=4,623 scans) UK	NR	-	Two FLS nurses	Thoracic or lumbar spine CT	NR	NR	Diagnostic accuracy, Change to clinical management
[redacted]	IB Lab [CiC] [redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
BoneBot* confirmed by IB Lab	Nicolaes (J Bone Miner Res. 2023; 1856-1866)	Retrospective diagnostic accuracy; summary of development (n=666 CT scans) and external validation (n=2,000 CT scans)	<u>With VF (n=1,536)</u> Age, median years: 74 years Male: 46% Ethnicity: NR	-	Medical doctor triaged the scans (certain VF, potential VF, no VF). Secondly, blinded vertebral readings (of certain VF, potential VF	Abdominal, chest CT	✓	Purpose of the study	Diagnostic accuracy, Failure rate or inconclusive AI reports

Technology (version)	Author (reference, year; page)	Study design (n=patients or scans); Country	Population characteristics	Comparator	Reference standard	Anatomical location and scan modality	Populations used to train AI (number and type) for details, where reported, see Appendix A4	Populations used to validate AI (number and type) for details, where reported, see Appendix A4	Outcomes
		Denmark	Without VF (n=407) Age, median years: 68 years Male: 54% Ethnicity: NR		and 5% subset of scans categorised as no VF) reviewed by radiologist				
BoneBot* confirmed by IB Lab	Nicolaes (Osteoporosis International, 2024; 143-152)	Retrospective diagnostic accuracy study; external validation (5,195 CT scans) China	Age, median years: 62 Male: 44.8% Ethnicity: NR	-	Sub-specialist radiologists	Abdominal, chest, thoracolumbar spine CT	✓	Purpose of the study	Diagnostic accuracy, Failure rate or inconclusive AI reports, Time to run the software
IB Lab FLAMINGO	Spangeus (Arch Osteopor, 2025; 42)	Retrospective diagnostic accuracy study (n=101,246 scans)	Age, mean years: 84 Male: 58% Ethnicity: NR	-	2 radiologists	CT for thoracic pathologies, aortic assessment, spinal imaging, and abdominal pathologies	NR	Purpose of the study	Diagnostic accuracy

Key: *company confirmed evidence was generalisable to the technology listed in the scope, † at stakeholder consultation the company stated that the population used to train and validate the AI for this study was reported in Seah et al. 2021.
Abbreviations: AI, Artificial intelligence; AP, anterior-posterior; CXR, Chest radiography; FLS, Fracture liaison service; MSK, Musculoskeletal; NR, Not reported; PA, Posterior-anterior; VCF, Vertebral compression fracture; VF, Vertebral fracture.

5. Clinical evidence review

The EAG extracted study characteristics for the 22 included papers, Appendix A4.

5.1 Quality appraisal of studies

The 22 clinical evidence studies (2 provided AiC and 1 CiC by 3 companies) include 5 abstracts. Of these 21 were diagnostic accuracy studies; four studies explicitly stated that their purpose was related to external validation of the AI algorithm (IB Lab, 2023; Nicolaes et al., 2024; Nicolaes et al., 2023; Page et al., 2023), and one included a pilot and post-study survey of radiologists (Jones et al., 2021; which was also reported in the abstract by Karunasena et al., 2022). Three studies were prospective (Bendtsen and Hitz, 2024; Connacher et al., 2019; Connor et al., 2024); of which 2 were available in abstract only. The remainder applied the AI tool to diagnostic images retrospectively. One Clinical Expert noted that prospective studies may be at risk of participation bias, where involvement in a study aimed to opportunistically detect vertebral fractures may influence the reporting of vertebral fractures thus not reflective of current practice (Appendix D1).

Generally, the software name and version number were poorly reported. Five studies explicitly mention the number of patients or scans, prevalence of vertebral fractures and population demographics used to train the AI technology (Dai et al., 2025; Ghatak et al., 2024; Guenoun et al., 2025; Nicolaes et al., 2024; Nicolaes et al., 2023). Three studies explicitly mention the study characteristics for validation of the AI technology (Bendtsen and Hitz, 2024; Ghatak et al., 2024; Guenoun et al., 2025), and validation was the purpose of six studies (Dai et al., 2025; IB Lab, 2023; Nicolaes et al., 2024; Nicolaes et al., 2023; Page et al., 2023; Spangeus et al., 2025).

Sixteen studies reported AI technologies when applied to CT images:

- chest, abdominal and pelvis (Avicenna.AI, 2024; Guenoun et al., 2025; Page et al., 2023)

- chest and abdominal (Dai et al., 2025; Nicolaes et al., 2023; Pereira et al., 2024) or thorax and abdominal (Bendtsen and Hitz, 2024; Kolanu et al., 2020)
- chest, abdominal, or thoracic or lumbar spine (Nicolaes et al., 2024; Roux et al., 2022)
- all CT that included the thoracic or lumbar spine (Connacher et al., 2019)
- all CT that included the spine (including abdomen and pelvic CT, and pulmonary angiogram CT) (Chappell et al., 2024)
- all CT including the abdomen (Wiklund et al., 2024)
- thoracic, lumbar or whole spine (IB Lab, 2023)
- thoracic pathologies, aortic assessment, spinal imaging and abdominal pathologies (Spangeus et al., 2025)
- anatomical location undefined (Connor et al., 2024).

Six studies included reported AI technologies when applied to X-ray images:

- Five studies included chest X-rays: anterior-posterior or posterior-anterior (Annalise.AI, No date); anterior-posterior or posterior-anterior or lateral (Frias, 2023); lateral in combination with anterior-posterior or posterior-anterior (Ghatak et al., 2024); projection undefined (Talwar, 2023) or at least one frontal projection (Jones et al., 2021).
- One study included spine (thoracic or lumbar) X-ray (Oppenheimer et al., 2024).

Eight studies explicitly reported the use of deidentified images when using the technology (Annalise.AI, No date; Dai et al., 2025; Ghatak et al., 2024; Nicolaes et al., 2024; Nicolaes et al., 2023; Page et al., 2023; Talwar, 2023; Wiklund et al., 2024). Company clarifications regarding de-identification of images before AI processing are summarised in Appendix C1. [redacted]

[redacted] The EAG assumes from company responses (Appendix C1) that the AI technologies use information from DICOM tags to determine if eligible for processing, but that patient information (other than the diagnostic image itself) is not used when generating an output of the AI.

One poster and two abstracts were conducted in a UK setting (Chappell et al., 2024; Connacher et al., 2019; Connor et al., 2024). Sample size of included evidence ranged between 100 patients (Guenoun et al., 2025) and 173,720 patients (Roux et al., 2022). One study reported on ethnicity of the population where AI was applied (Ghatak et al., 2024). One Clinical Expert noted that patients undergoing long-term steroid treatment (longer than 3 months duration) should be a considered subgroup, however, the EAG notes that the duration of medication was not reported across the included evidence. One abstract provided academic in confidence reported [redacted] (Avicenna.AI, 2024). Thirteen studies were conducted in a population aged 50 years or older (Annalise.AI, No date; Bendtsen and Hitz, 2024; Dai et al., 2025; Guenoun et al., 2025; IB Lab, 2023; Kolanu et al., 2020; Nicolaes et al., 2024; Nicolaes et al., 2023; Page et al., 2023; Pereira et al., 2024; Roux et al., 2022; Spangeus et al., 2025; Wiklund et al., 2024).

The prospective study by Bendtsen and Hitz (2024) closest described the use of an AI technology and how it may be used in the NHS, including a real-time synchronous review of 10,012 thorax and abdomen CT images by HealthVCF prior to upload to RIS or PACS as well as a validation cohort of 1,000 images which were randomly selected and evaluated by both HealthVCF and one specialised radiographer (Bendtsen and Hitz, 2024). Two academic-in-confidence reports supplied by manufacturers reported [redacted]. The majority of studies compared the output of the AI with a reference standard; however the reference standard varied (described in the literature as: radiologist, diagnostic radiologist, thoracic radiologist, general radiologist, interventional radiologist, neuroradiologist, specialised radiographer, senior radiologist, expert, FLS nurse, sub-specialist radiologist) and this may introduce heterogeneity into the evidence base and may not reflect the reference standard in the UK NHS. Twelve studies described incidental vertebral fracture detection by an AI technology when compared with standard care; most commonly the original radiology report (Annalise.AI, No date; Avicenna.AI, 2024; Bendtsen and Hitz, 2024; Chappell et al., 2024; Connor et al., 2024; Dai et al., 2025; Frias, 2023; Ghatak et al., 2024; Kolanu et al., 2020; Pereira et al., 2024; Talwar, 2023; Wiklund et al., 2024). No evidence reported the use case of initial radiographer with AI compared with a reference standard, which is how it would be used in the NHS. One Clinical Expert noted that diagnostic accuracy should be considered “per vertebra” and “per patient”, as a single fracture may be missed, however the patient may still undergo correct treatment pathway if there are multiple fractures. While several studies reported subgroup analysis (for example by age, sex), none of the included evidence was statistically powered to determine differences in diagnostic performance by subgroup and as is common with sub-group analyses in clinical studies should be considered exploratory. The reference standard in 6 studies explicitly stated they were blinded to the AI result. A further study reported that radiologists reviewing discrepancies were blinded.

The majority of studies retrospectively processed images and determined the diagnostic accuracy of the AI technology when compared with a reference standard or standard of care. Only 1 prospective study reported changes to clinical management (Bendtsen and Hitz, 2024), and none reported on health-related quality of life (which was expected given the study designs included in the evidence base). One study reported the time taken to process an image using an AI technology (Nicolaes et al., 2024) and another study reported the time taken for a radiologist to review (look-up, analyse, potentially report) the output from an AI technology (Bendtsen and Hitz, 2024).

Results from the evidence base

5.1.1 Annalise Enterprise CXR or Annalise Container CXR

Five studies reported the use of an Annalise AI technology; this included 2 abstracts (Frias, 2023; Talwar, 2023).

Measures of diagnostic accuracy

Two studies (including one provided academic in confidence) reported on diagnostic accuracy when compared with a reference standard, Table 3.

The study by (Ghatak et al., 2024) reported overall diagnostic accuracy of the AI technology when applied to frontal and lateral chest X-rays when compared with a reference standard (two thoracic radiologists, arbitration by a third), and additional subgroup analysis by sex, age, race, ethnicity and radiograph machine manufacturer. The Clinical Experts noted that it would be rare in NHS practice to obtain a lateral projection for chest X-ray, therefore the generalisability of this specific study to the decision problem is uncertain. This study reported that in the majority of subgroup analyses the point estimates of sensitivity and specificity exceeded 80% with three exceptions:

- Black or African American race (n=30 total cases, 4 positive cases, sensitivity: 75.0%, specificity: 100%)
- Agfa manufacturer (n=52 total cases, 21 positive cases, sensitivity: 76.2%, specificity: 87.1%),
- GE Healthcare manufacturer (n=8 total cases, 4 positive cases, sensitivity: 75.0%, specificity: 100%).

However, the authors highlight that the study was powered to assess overall model performance and not for individual subgroups so these estimates will be subject to a degree of imprecision and point estimates may change should further data become available.

The academic in confidence report provided by Annalise.AI stated that [redacted]

[redacted]

Three additional studies commented on elements of diagnostic accuracy of the AI technology when compared with the original radiology report:

- The abstract by (Talwar, 2023) reported that the Annalise Enterprise CXR technology detected 169 abnormalities across 1,559 chest X-rays, of which 97 were confirmed by a radiologist and missing from the original radiology report; however, the EAG notes that this abstract looked at 60 clinical findings not all related to vertebral fractures. This study reported that spinal compression fractures were the third most commonly missed finding (that is missing from the original radiology report but detected by the Annalise Enterprise CXR technology) in 12% of cases. The top two most commonly missed findings were pulmonary nodules and pleural effusions, both missed in 16% of cases.

- The abstract by (Frias, 2023) reported that the algorithm by Annalise.AI correctly identified compression fracture false negatives (3.1%; 114/3,760) and false positives (0.2%; 8/3,760) in the original radiology report when discrepancies between the original report and the algorithm were reviewed by a thoracic radiologist reference standard.

- One study (Ghatak et al., 2024) also reported that among true positive cases (confirmed by reference standard) that 36.4% of patients (86/236) had a diagnosis of vertebral compression fracture identified by ICD-10 codes, and that 33.1% (78/236) were receiving disease-modifying medication for osteoporosis as identified by ICD-10 codes and medication records for the same study duration and 1 additional year follow-up.

Table 3: Measures of diagnostic accuracy (Annalise.AI technologies) compared with a reference standard

Study	Reference standard	Anatomical location and scan modality	Definition fracture	Prevalence	False positive	False negative	True positive	True negative	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy	Kappa	PPV	NPV	AUROC
Annalise.AI [AiC]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted] [redacted]	[redacted] [redacted]	[redacted] [redacted]	[redacted] [redacted]	[redacted] [redacted]	[redacted] [redacted]	[redacted]
Ghatak (J Am Coll Radiol, 2024; 220-229)	Agreement of 2 thoracic radiologists (arbitration of 3rd)	Chest X-ray (frontal and lateral)	Vertebral compression fracture (**per patient**)	45.7% (272/595)	35 Female: 17 Male: 18	29 Female: 15 Male: 14	243 Female: 163 Male: 80	288 Female: 155 Male: 133	0.893 (0.857 to 0.927) Female: 0.916 (0.871 to 0.955) Male: 0.851 (0.777 to 0.915)	0.892 (0.854 to 0.923) Female: 0.901 (0.855 to 0.942) Male: 0.881 (0.828 to 0.927)	-		-	-	0.955 (0.939 to 0.968) Female: 0.965 (0.947 to 0.980) Male: 0.939 (0.909 to 0.964)

Abbreviations: AI, Artificial intelligence; AP, Anterior-posterior; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; LAT, Lateral; NPV, negative predictive value; PA, Posterior-anterior; PPV, positive predictive value

Failure rate or inconclusive AI reports

Failure rate was reported in two studies and was negligible, Table 4.

Table 4: Summary of failure rate (Annalise AI technologies)

Study	Failure definition	Failure outcome
Ghatak (J Am Coll Radiol, 2024; 220-229)	Unsuccessful model inference in chest X-ray	0.2% (1/596)
Talwar (RSNA, 2023; W5B-SPCH-2) *Abstract*	Unable to process chest X-ray	0% (0/1,559)

Healthcare professional user acceptability

One study reported on healthcare professional acceptability (Jones et al., 2021) which summarised staff perception following a pilot conducted across a radiology network in Australia. During the almost 6-week pilot phase, 90% of users (9/10) reported subjective improved accuracy while using the Annalise CXR viewer (with modified user interface), with 90% (9/10) also reporting satisfaction with the accuracy of the AI model findings.

At the end of the pilot phase, 90% reported a more positive attitude towards the Annalise.AI CXR viewer and 90% reported a more positive attitude towards AI in general. No radiologist reported a negative attitude towards the Annalise CXR viewer or towards AI in general.

Health-related quality of life

Not reported; no RCTs identified.

5.1.2 BoneView

One diagnostic accuracy study applied BoneView to 512 thoracic and lumbar spine X-rays retrospectively and compared to a reference standard of consensus between 2 radiologists with experience in musculoskeletal imaging (Oppenheimer et al., 2024).

Measures of diagnostic accuracy

The diagnostic accuracy of BoneView was reported on a per-vertebra basis (Oppenheimer et al., 2024), Table 5; no results reporting per patient were identified. This study demonstrated generally higher sensitivity and specificity achieved in the lumbar spine when compared with the thoracic spine and slight increase in sensitivity when both lateral and AP projections were considered in a second view (when compared with considering lateral projections alone). Sensitivity and specificity were lower for grade 1 fractures, and similar between grade 2 and grade 3 fractures.

The study also reported subgroup analysis showing a reduction in sensitivity and specificity in fracture detection, across both lumbar and thoracic spine, when foreign materials (metal screws, cement kyphoplasty, vertebral replacement) were present. Sensitivity and specificity were also impacted by the presence of degenerative changes, however the number with degenerative disease was small leading to wide confidence intervals, such that the clinical significance of this finding is unknown.

An incidental finding reported by one study was that in a small number of cases in lateral spine radiographs BoneView mislabelled intervertebral spaces as a fracture (Oppenheimer et al., 2024). The authors suggest (in a separate article) that this incorrect marking may be due to degenerative changes mimicking a fracture line (Oppenheimer et al., 2023).

Table 5: Measures of diagnostic accuracy (BoneView) per vertebra basis compared with a reference standard

Study	Reference standard	Anatomical location and scan modality	Definition fracture	Prevalence	Lumbar: Sensitivity (95%CI)	Lumbar: Specificity (95%CI)	Thoracic: Sensitivity (95%CI)	Thoracic: Specificity (95%CI)
Oppenheimer (Skeletal Radiol, 2024; 1563-1571)	Two radiologists (one with MSK experience)	Thoracic and lumbar spine X-ray	Positive and doubt fractures were classed as fracture positive by the AI for further analysis (per vertebrae)	Lumbar fracture: 12.9% (323/2504 vertebrae) Thoracic fracture: 10.7% (172/1610 vertebrae)	Lateral: 63.2 (5.3)% Lateral and AP: 72.4 (4.9)%	Lateral: 96.7 (0.8)% Lateral and AP: 94.2 (1.0)%	Lateral: 51.2 (7.5)% Lateral and AP: 60.6 (7.5)%	Lateral: 98.3 (0.7)% Lateral and AP: 94.0 (1.3)%
Oppenheimer (Skeletal Radiol, 2024; 1563-1571)	One radiologist reported Genant classification	Thoracic and lumbar spine X-ray	Grade 1 Grade 2 Grade 3 (per vertebrae)	NR	52.5 (7.7)% 72.3 (8.7)% 75.8 (10.7)%	NR	42.4 (10.5)% 60.0 (14.3)% 60.0 (14.3)%	NR
Oppenheimer (Skeletal Radiol, 2024; 1563-1571)	One radiologist reported Genant classification	Thoracic and lumbar spine X-ray	Anterior (wedge), total: Grade 1 Grade 2 Grade 3 (per vertebrae)	NR	64.4 (6.4)% 53.4 (10.4)% 68.9 (10.6)% 75.9 (11.4)%	NR	51.9 (9.6)% 46.7 (14.6)% 57.1 (18.3)% 58.0 (17.4)%	NR
Oppenheimer (Skeletal Radiol, 2024; 1563-1571)	One radiologist reported Genant classification	Thoracic and lumbar spine X-ray	Middle (crush), total: Grade 1 Grade 2 Grade 3 (per vertebrae)	NR	60.8 (9.3)% 51.4 (11.5)% 81.8 (14.7)% 75.0 (30.0)%	NR	50.0 (11.9)% 40.0 (15.8)% 64.7 (22.7)% 60.0 (13.1)%	NR

Abbreviations: AI, Artificial intelligence; AP, Anterior-Posterior; CI, Confidence interval; MSK, Musculoskeletal; NR, Not reported

Failure rate or inconclusive AI reports

The study by Oppenheimer et al. 2024 reported that 1.4% (5/357) of lumbar spine and 5.1% (8/155) of thoracic spine X-rays were excluded as being of unsupported anatomical regions by incorrectly classifying thoracic spine as chest X-rays and incorrectly classifying lumbar spine as abdominal X-rays (Oppenheimer et al., 2024). The authors speculated that this was often the case with obese patients where the field of view for the radiograph may include a higher proportion of surrounding tissue.

Healthcare professional user acceptability

Not reported.

Changes to clinical management

Not reported.

Health-related quality of life

Not reported; no RCTs identified.

5.1.3 BriefCase-Triage

One diagnostic accuracy study applied an AI technology by Aidoc Medical (BriefCase- Triage was not specified) to 1,112 patients over the age of 60 years who had undergone an abdominal CT scan and compared with a reference standard (general radiologist, with only VCFs identified being reviewed by a musculoskeletal radiologist) and also compared with the original radiology report (Wiklund et al., 2024).

Measures of diagnostic accuracy

The study by Wiklund (et al. 2024) reported measures of diagnostic accuracy of moderate or severe (grade 2 or 3) vertebral compression fractures when compared with a reference standard (Wiklund et al., 2024), Table 6. The study noted in a post hoc subgroup analysis that there was no evidence of a difference in AI sensitivity or specificity when stratifying by sex or age (above and below population mean of 74 years), p>0.05 in both cases. However, the authors acknowledged that the study sample size prevented additional subgroup analyses by different scan settings, patient position or presence of intravenous contrast.

Table 6: Measures of diagnostic accuracy (Aidoc Medical) compared with a reference standard

Study	Reference standard	Anatomical location and scan modality	Technology (definition fracture)	Prevalence	Original radiologist reporting	Sensitivity	Specificity	PPV	NPV
Wiklund (J Bone Mineral Res, 2024; 1113-1119)	General radiologist with VCF detected also reviewed by MSK radiologist.	Abdominal CT	VCF grade 2 or 3; patient-level	VCF prevalence 16.8% (187/1112 patients), of which 62 were incidence VCFs and 49 had a previously unknown prevalent VCF.	29.7% (33/111) for all grades; 48% (30/62) for incident VCFs and 6% (3/49) for previously unknown prevalent VCFs 42.8% (27/63) for grade 2 or 3	85.2%	92.3%	57.8%	98.1%

Abbreviations: MSK, Musculoskeletal; NPV, Negative predictive value; PPV, Positive predictive value; VCF, Vertebral compression fracture

The study by Wiklund (Wiklund et al., 2024) also reported that:

- 48% (30/62) of incident VCFs (new fracture compared with previous images) were described in the original radiology report, and
- 6% (3/49) of previously unknown prevalent VCFs ("prevalent" fracture was present in a previous image and "unknown" were not reported before) were reported.

However, it is unclear how many of these were detected by the AI technology.

Failure rate or inconclusive AI reports

A total of 0.6% (7/1,112) patients were excluded from analysis; 5 due to failed upload to the AI technology and 2 due to failed analysis.

Healthcare professional user acceptability

Not reported

Changes to clinical management

As Wiklund et al. (2024) was a diagnostic accuracy study applied to retrospective images, this outcome was not reported (Wiklund et al., 2024). However, the authors noted that only 10% (3/30 patients) of patients with a reported incident or previously unknown prevalent VCF had started pharmacologic osteoporosis treatment within 1 year follow-up. Five additional patients started treatment between 15 and 26 months after baseline CT, of which 2 had suffered additional VCF before treatment was initiated.

Health-related quality of life

Not reported; no RCTs identified.

5.1.4 CINA-VCF Quantix

Three diagnostic accuracy studies reported the use of CINA-VCF (none explicitly mentioning Quantix) when applied to images retrospectively; this included one abstract provided as academic in confidence.

Measures of diagnostic accuracy

Three studies reported on elements of diagnostic accuracy when compared with a reference standard, Table 7. Sensitivity and specificity were greater than 90% in both studies where these measures were reported.

- The study by (Dai et al., 2025) reported that most VCF cases missed by the AI technology were mild or borderline VCFs (with vertebral height loss around 25%), and that other pathologies (Schmorl's nodes, scoliosis, artefact) may have influenced findings. The study reported on subgroup analysis by scanner manufacturer, slice thickness, presence of contrast, reconstruction kernel (also referred to as a filter or algorithm by some CT vendors), and patient age and sex, and sensitivity and specificity remained above 90% in all cases. The EAG notes that the study by (Dai et al., 2025) reported that there were discordances between two radiologists in 10.7% (51/474) cases; however Cohen's Kappa 0.75 (95%CI 0.69 to 0.81) showed strong correlation.

- The study by (Guenoun et al., 2025) reported that the majority of false positives and false negatives had a vertebral height loss close to the definition of grade 2 fracture, that 1 false negative was due to a compression not being visible on the plane passing through the midsagittal plane of the vertebral body, and 1 false positive was caused by natural deformation of L5 vertebra.

- The academic in confidence abstract provided by Avicenna.AI only reported on [redacted]

Two studies (including one abstract) also reported on detection of fractures when compared with the original radiology report (standard care):

- In the study by (Dai et al., 2025) the original radiology report was only available in a subset of cases, the authors noted that the original reports only included mention of VCF correctly in 36.7% (44/120) of cases which compared with the reference standard; therefore standard of care missed 76 cases – of which the AI algorithm was able to detect 88.2% (67/76).

- The abstract provided AiC by Avicenna.AI reported that [redacted]

Table 7: Measures of diagnostic accuracy (CINA-VCF) compared with a reference standard

Study	Reference standard	Anatomical location and scan modality	Definition fracture	Prevalence	False positive	False negative	True positive	True negative	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy	Kappa	PPV	NPV	AUROC
Dai (J Comput Assist Tomogr, 2025)	Two radiologists (arbitration by a third)	Chest or abdomen, axial or sagittal acquisition CT	Vertebral compression fracture (grade 2 or 3) per patient	35.0% (166/474 cases)	22	8	158	286	0.952 (0.907 to 0.979)	0.929 (0.894 to 0.965)	0.937 (0.911 to 0.957)	-	0.878 (0.821 to 0.922)	0.973 (0.947 to 0.986)	0.974 (0.962 to 0.986)
Guenoun (Clin Radiol, 2025; 106831)	Two radiologists	Chest-abdominal-pelvis CT	Vertebral compression fracture (grade 2 or 3) per patient	52.0% (52/100 patients)	4	4	48	44	0.923 (0.815 to 0.979)	0.917 (0.800 to 0.977)	0.920 (0.848 to 0.965)	-	-	-	-
Avicenna.AI [AiC] *Abstract*	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Abbreviations: AUROC, area under the receiver operating characteristic curve; CI, confidence interval; NPV, Negative predictive value; PPV, Positive predictive value;

Failure rate or inconclusive AI reports

The EAG notes that the study by (Guenoun et al., 2025) did not explicitly report failure, however did report that 1% (17/1,700) of vertebrae presented did not have vertebral height loss calculated due to cement or hardware and therefore were excluded from analysis. The EAG notes that these are explicit contraindications for that technology and therefore their exclusion from analysis appropriately ensures that the technology is being used within its approved instructions for use.

Healthcare professional user acceptability

Not reported.

Changes to clinical management

Not reported; all 3 studies using a CINA-VCF technology were retrospective studies.

Health-related quality of life

Not reported; no RCTs identified.

Other outcome measures

The study by (Guenoun et al., 2025) also reported on vertebral height loss measurements, and Hounsfield Units. As these were not explicitly listed as outcomes in the Final Scope, the EAG has reported these results within Appendix C2.

5.1.5 HealthVCF, HealthOST

Five studies reported the use of HealthVCF (Bendtsen and Hitz, 2024; Chappell et al., 2024; Connor et al., 2024; Pereira et al., 2024; Roux et al., 2022) and three reported the use of Zebra Medical Vision (Connacher et al., 2019; Kolanu et al., 2020; Page et al., 2023) noting that Nanox AI acquired Zebra Medical in 2021, and the company stated that this evidence was generalisable to HealthVCF.

Measures of diagnostic accuracy

Six studies reported on diagnostic accuracy of HealthVCF or Zebra algorithm when compared with reference standard, Table 8. One study explicitly stated that the software did not localise the fracture but alerted the radiologist to the potential presence of the fracture (binary outcome: fracture, no fracture) which the authors stated may limit its benefits in clinical use (Page et al., 2023). The EAG notes that this study used a predecessor technology (Zebra Medical Vision); and that functionality may differ with newer versions, so the EAG questions the generalisability of this evidence. One study reported on diagnostic accuracy in a subset of 1,000 patients (Bendtsen and Hitz, 2024), another on a subset of 500 patients where the prevalence was set to 50% (250 patients with VFF, and 250 without VFF) (Roux et al., 2022), and an abstract on a subset of 683 scans. The study by (Pereira et al., 2024) reported 73.8% sensitivity and 92.7% specificity of HealthVCF when compared with the reference standard (2 radiologists specialised in musculoskeletal imaging) across 899 total chest and abdominal CT scans. Subgroup analysis by type of scan demonstrated a decreased sensitivity in chest CT angiography (sensitivity: 66.6%, specificity: 100%), increased sensitivity in chest CT (sensitivity: 80.6%, specificity: 91.3%) and no change in abdominal CT (sensitivity: 73.9%, specificity: 95.0%).

The prospective study by Bendtsen and Hitz (2024) closest described the use of an AI technology and how it may be used in the NHS (Bendtsen and Hitz, 2024). This described the real-time synchronous review of 10,012 thorax and abdomen CT images by HealthVCF (configured for highest specificity) conducted in a single hospital in Denmark, where annotations of moderate or severe VCFs were applied by the AI before the images were returned to the Radiology Information System (RIS) which were reviewed by a radiologist or specialised radiographer and radiology report completed. The study reported that HealthVCF detected a VCF grade 2 or 3 in 1,543 of 10,012 thorax and abdomen CT scans. When the 1,543 HealthVCF positive scans were reviewed by radiologists, 177 scans were excluded (reasons not reported) and true positives were detected (by radiologists) in 630 of the remaining 1,366 scans (46% true positives, 54% false positives). The study also described a validation cohort of 1,000 images which were randomly selected and evaluated by both HealthVCF and one specialised radiographer. The authors noted that the AI was over-calling patients (65 true positives, 81 false positives) and missed 3 in 100 VCFs (30 false negatives). The authors noted that the tool could not be used in isolation (human interpretation required as per device regulation), that the low performance was due to lack of generalisability to a Danish population, however recommended that the AI tool be used to prioritise which CT scans should be analysed for potential VCF.

The abstract by (Connacher et al., 2019) reported that when applied to 4,623 scans in a UK setting the algorithm by Zebra identified 1,305 images as having a vertebral fracture; however, only partial results were reported. Of these, 633 (49%) required more detailed analysis (however no additional detail was provided to explain why), therefore the confirmation of fracture was only available in the 672 remaining scans (EAG noting that the number is incorrectly reported in the paper); 279 of which (42%) were confirmed by an FLS nurse to have included a vertebral fracture, and 393 (58%) did not include a vertebral fracture. The abstract also reported that only a subset of 2% (55/3,318) scans where a vertebral fracture was not detected by the AI were checked by an FLS nurse; therefore, the specificity cannot be identified from this study. The poster by (Chappell et al., 2024) reported data from 4 UK centres contributing to the ADOPT study where 1 site chose a ‘high specificity’ configuration of the HealthVCF technology, and 3 sites chose a ‘balanced’ configuration. The sensitivity and specificity were different between the high specificity site and those using the balanced configuration. The poster also reported that prevalence of vertebral fragility fractures varied between 9.7% for pulmonary angiogram CT and 42.7% for abdomen and pelvic CT; which is important when considering the baseline characteristics of patients where the AI technology is being used. They also reported that prevalence of VFFs varied between 5% and 17% across sites.

Five studies reported on detection of VCF by the AI technology when compared with standard care (4 compared with the original radiology report, and 1 compared with submissions to the Fracture Liaison Service database),

Table 9. The abstract by (Connor et al., 2024) reported that the AI technology detected between 42% more and 2539% more fractures than the number submitted to the Fracture Liaison Service database across 4 UK centres; however, this lacked detail. The study by (Kolanu et al., 2020) reported that across 183 true positive grade 2 or 3 VCFs (as determined by Zebra and confirmed by the reference standard) that 62% (111/183) were documented in the original radiology report, demonstrating that 72 additional fractures were detected by the AI technology (65% increase). When compared with standard care, the prospective study by (Bendtsen and Hitz, 2024) reported that 91% of the VCFs detected were described in the primary radiology report and 9% were added secondarily. However, the authors noted that more VCFs were described in the primary report towards the end of the study period (in the first half of the study 12.9% to 26.5% of findings were added after the primary report, and in the second half of the study this occurred in 0% to 4.2% of cases). The authors stated that this demonstrated a time factor or potentially increased awareness among radiologists during the study which may have influenced findings (that is the Hawthorne effect). The study by (Pereira et al., 2024) reported that there were 62 moderate to severe VCFs retrospectively identified by the radiologists specialised in musculoskeletal imaging that were not reported by the original reporting radiologist, of which the AI technology successfully detected 38 (61.2%). This restriction to detection of moderate and severe VCF, is consistent with the earlier study by (Page et al., 2023) which reported results when using Zebra Medical Vision (predecessor to the HealthVCF) and demonstrated that the sensitivity was increased from 66% when considering mild, moderate and severe VCF to 78% when considering moderate and severe VCF only.

The study by (Roux et al., 2022) states that the software does not measure vertebral heights, and that it was trained on moderate and severe fractures (mild fractures excluded) detecting fracture in L1 to L4 only. The study also reported diagnostic accuracy of the AI compared with 2 blinded experts in a subset where the prevalence was artificially set to 50% (including 250 patients with vertebral fracture and 250 without). Therefore, the EAG would advise caution when interpreting the results from this study with the others identified from HealthVCF.

One additional abstract by (Connor et al., 2024) reported the use of HealthVCF across 4 UK centres including a total of 36,620 CT scans (anatomical location undefined). However, the EAG noted that the AI technology settings varied across centres (1 centre chose high specificity mode, 3 centres chose balanced mode) as did the reference standard (1 centre had synchronous live reporting by radiologists, 3 centres had asynchronous local clinical confirmation and FLS management). While the proportion of scans including a moderate to severe thoraco-lumbar VFF as identified by the AI technology is reported (ranging between 12.4% and 40.6% of scans), it is unclear how the proportion of clinically reviewed scans were selected, or the overlap of outcomes. For the 1 centre choosing synchronous live reporting by a radiologist, a total of 410 scans (12.4%) were identified by HealthVCF as including a VFF, a total of 410 scans were also reviewed clinically (assumed to be the same 410 scans) of which 221 (53.9%) were considered true positives.

Table 8: Measures of diagnostic accuracy (HealthVCF and Zebra technologies) compared with a reference standard

Study	Reference standard	Anatomical location and scan modality	Definition fracture	Prevalence	False positive	False negative	True positive	True negative	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy	Kappa	PPV	NPV	Youden Index
Chappell (2024) *Poster*	Clinician with local radiologist adjudication	All CT involving spine	Vertebral fragility fracture (grade undefined); reported per scan (Per patient)	12.8% (255/2,000)	-	-	-	-	<u>Sites with high specificity configuration:</u> 48.3% <u>Sites with balanced configuration:</u> 79.0% [radiologist report: 51%]	<u>Sites with high specificity configuration:</u> 98.5% <u>Sites with balanced configuration:</u> 81.2% [radiologist report: 100%]	-	-	-	-	-
Kolanu (Journal of Bone and Mineral Research, 2020; 2307-2312)	1 radiologist	Abdomen or thorax CT	Moderate to severe VCF (grade 2 or 3) (Per patient)	17.8% (280/1,570)	99	97	183	1,191	0.65 (0.598 to 0.709)	0.92 (0.909 to 0.938)	0.88 (0.858 to 0.891)	-	0.65 (0.593 to 0.705)	0.93 (0.910 to 0.939)	-
Page (JBMR Plus, 2023; e10778)	2 neuroradiologists	Chest, abdominal or pelvis CT	Moderate to severe VCF (grade 2 or 3) (Per patient)	12.6% (137/1087)	124	30	107	826	0.78 (0.70 to 0.85)	0.87 (0.85 to 0.89)	-	-	-	-	-
Bendtsen and Hitz (Calc Tissue Int, 2024; 468-479	Radiographer	Thorax and abdomen CT	Moderate or severe VCF (grade 2 or 3) (Per patient)	9.5% (95/1000 patients; subset of larger study)	81	30	65	824	0.684 (0.581 to 0.776)	0.910 (0.890 to 0.928)	0.889	-	0.445 (0.363 to 0.530)	0.965 (0.950 to 0.976)	0.59
Pereira (Radiol Bras, 2024; e20230102)	2 radiologists specialised in musculoskeletal imaging	Chest and abdominal CT	Moderate or severe VCF (grade 2 or 3) (Per patient)	16.1% (145/899 scans)	55	38	107	699	0.738 (0.657 to 0.805)	0.927 (0.905 to 0.944)	0.896 (0.874 to 0.915)	-	0.660 (0.581 to 0.731)	0.948 (0.929 to 0.962)	-
Kolanu (Journal of Bone and Mineral Research, 2020; 2307-2312)	1 radiologist	Abdomen or thorax CT	Any VCF (Per patient)	23.9% (406/1,696)	-	-	-	-	0.54 (0.496 to 0.593)	0.92 (0.909 to 0.938)	0.83 (0.814 to 0.850)	-	0.69 (0.640 to 0.741)	0.87 (0.848 to 0.884)	-
Page (JBMR Plus, 2023; e10778)	2 neuroradiologists	Chest, abdominal or pelvis CT	Any VCF (Per patient)	20.9% (227/1087)	82	78	149	778	0.66 (0.59 to 0.72)	0.90 (0.88 to 0.92)	-	-	-	-	-
Roux (Rheumatology, 2022; 3269-3278)	2 experts (undefined)	Abdominal or lumbar CT	At least one moderate or severe vertebral fracture between L1 and L4. (Per patient)	50% (250/500; subset of larger study)	-	-	-	-	0.94 (0.89 to 0.98)*	0.65 (0.60 to 0.70)*	-	-	-	-	-

Key: *EAG suspects the sensitivity and specificity have been reported in the wrong order in the paper. Abbreviations: AUROC, Area under the receiver operating characteristic curve; CI, Confidence interval; L1, First lumbar vertebra; L4, Fourth lumbar vertebra; NPV, Negative predictive value; PPV, Positive predictive value; VCF, Vertebral compression fracture.

Table 9: Measures of diagnostic accuracy of AI technology (HealthVCF, Zebra) when compared with the original radiology report

Study	Study design	Comparison with original radiology report
Connor (J Bone Min Res, 2024; 296-297) *Abstract* ADOPT study	Diagnostic accuracy study prospective (n=36,620 total scans) Site A, n=3,298 Site B, n=20,239 Site C, n=3,654 Site D, n=9,429	Compared to submission to Fracture Liaison Service Database in 2022: Site A: 42% increase Site B: 2,539% increase Site C: 354% increase Site D: 447% increase
Chappell (2024) *Poster* ADOPT study (500 scans from 4 sites as shadow test before main study)	Diagnostic accuracy study applied to retrospective images (n=2,000 scans)	Additional 22.5 patients per 1,000 patient scans
Kolanu (Journal of Bone and Mineral Research, 2020; 2307-2312)	Diagnostic accuracy study applied to retrospective images (n=2,357 patients)	72 additional fractures detected by AI not included in original radiology report
Bendtsen and Hitz (Calc Tissue Int, 2024; 468-479)	Diagnostic accuracy study prospective (n=1,000 patients)	91% of fractures detected by AI were also in the original radiology report
Pereira (Radiol Bras, 2024; e20230102)	Diagnostic accuracy study applied to retrospective images (n=964 scans)	62 additional fractures detected by AI not included in original radiology report

Abbreviations: AI, Artificial Intelligence.

Failure rate or inconclusive AI reports

Failure rate between 5.6% and 9.4% was reported by three studies (Kolanu et al., 2020; Page et al., 2023; Pereira et al., 2024), Table 10.

Table 10: Summary of failure rate (HealthVCF and Zebra technologies)

Study	Failure definition	Failure outcome
Pereira (Radiol Bras, 2024; e20230102)	Unable to process	5.6% (54/964) - 3.8% (37/964) due to unavailable axial series, incomplete examinations - 1.8% (17/964) due to failure during analysis (spine could not be segmented due to distortion of the vertebral body, presence of metallic objects)
Page (JBMR Plus, 2023; e10778)	Unable to process	9.4% (113/1200)

Study	Failure definition	Failure outcome
(Kolanu et al., 2020)	Unanalysable	6.0% (108/1,804 eligible scans) - 28 non-axial series - 24 less than 20 images - 3 inconsistencies in imaged data (no further detail provided) - 1 could not segment vertebrae - 1 non-CT modality - 1 could not analysis - 50 no data (no further detail provided)

Healthcare professional user acceptability

Not reported.

Changes to clinical management

The study by Bendtsen and Hitz (2024) described the prospective application of the HealthVCF in a single hospital in Denmark. In the subset of 223 patients identified as having a vertebral fracture (without a prior diagnosis of osteoporosis at baseline) who were followed for 6 months, 37 (16.6%) were referred for a DEXA scan, 25 (11.2%) had a new diagnosis of osteoporosis in their electronic patient record, and 23 (10.3%) had started anti-osteoporosis treatment (where 14 patients overlapped all 3 of these groups) (Bendtsen and Hitz, 2024). The EAG notes that only true positives (flagged by HealthVCF and confirmed by radiologists) were followed, and therefore it is unclear whether the above proportions are higher than expected for the population via standard care.

The study by Bendtsen and Hitz (2024) also reported that of the 218 patients who had known osteoporosis at baseline, 154 were on medical treatment at baseline and that 16 had medication changes based on results (15 changing from oral bisphosphonates, 1 changing from intravenous bisphosphonates). Of the 64 patients that had a diagnosis of osteoporosis at baseline but were not on medical treatment, 41 were referred for DEXA scans, and 11 started on anti-osteoporosis treatment during the 6-month follow-up; but the overlap between these groups is unknown. The EAG notes that the proportions of the patients referred for DEXA scans were similar between those with prior osteoporosis diagnosis and those without (18.8%, 41/218 compared with 16.6% 37/223; EAG calculated p=0.63); as were the proportion receiving anti-osteoporosis treatment (5.0%, 11/218 compared with 10.3%, 23/223; EAG calculated p=0.06). Therefore, it is unclear of the downstream impact of earlier detection.

The abstract by (Connacher et al., 2019) reported clinical outcomes from a subset of 50 patients who were identified as having a vertebral fracture by the Zebra algorithm and had completed the FLS assessment in a UK setting. Of these 50 patients, 57% had no additional assessment after FLS, 25% had a GP referral, 6% bone clinic referral and 12% had died. However, it is unclear whether these 50 patients had the vertebral fracture confirmed by the FLS nurses (who were the reference standard in this particular study) and the timepoint at which the FLS assessment was completed was not reported. It is also unclear how the 50 patients were selected given that a total of 279 patients had a vertebral fracture detected by Zebra and confirmed by FLS nurse.

Health-related quality of life

Not reported; no RCTs identified.

Time to run the software

The study by Bendtsen and Hitz (2024) reported that radiographers spent on average 5 to 10 minutes analysing each image generated by the HealthVCF (including look-up, analysing and potential reporting) (Bendtsen and Hitz, 2024) which the EAG considered would be in addition to normal reporting times. For context one Clinical Expert stated that radiologists (reference standard) often take less than 2 minutes to report a chest X-ray.

Other outcome measures

The study by Roux et al. 2022 also reported on Hounsfield Units (relative quantitative measurement of radio density used in CT) and simulated DEXA scan T-scores (Roux et al., 2022). The EAG notes that these were not explicitly listed as outcomes in the Final Scope, but as these specific results were not compared with DEXA scan results, the accuracy and generalisability of these remains unclear.

5.1.6 IB Lab FLAMINGO

Four studies reported the validation of AI technologies by IB Lab; this included one internal report submitted commercial in confidence by the company.

Measures of diagnostic accuracy

Four studies reported on diagnostic accuracy of an IB Lab technology per patient or scan (Table 11) and per vertebrae (Table 12) when compared with a reference standard. The EAG notes that in 2 studies improvement in diagnostic accuracy was achieved when the definition of vertebral fracture was restricted to moderate and severe fractures (grade 2 and 3) when compared to definitions including mild VF (combining grade 1, 2 and 3) (Nicolaes et al., 2024; Nicolaes et al., 2023).

The study by Nicolaes et al. 2024 also reported that the presence of Schmorl's nodes were found in 17% (n=819) of all CT scans, however the specificity remained high (above 90%) across patient level and vertebrae level analysis (Nicolaes et al., 2024). When focusing on moderate or severe VFs (grade 2 or 3) the study also reported that 30% of patient-level false positives (87/285) were explained by errors involving the first and last visible vertebrae in the CT scan, and that 20% of false positives (56/285) contained Schmorl's nodes. The authors reported that the majority (97%) of false negative results were mild (grade 1) VFFs.

Two studies reported subgroup analysis by sex and age (Nicolaes et al., 2023; Spangeus et al., 2025). Both reported that sensitivity was higher in females (per patient reported in Nicolaes et al. 2023, and per scan reported in Spangeus et al. 2025). However, no significant difference (evidenced by overlapping 95% confidence intervals) in sensitivity by age, or specificity across sex or age subgroups was found. The EAG notes that in the commercial in confidence report (IB Lab, 2023) that [redacted]

[redacted]

The study by Spangeus et al. 2025 reported that vertebral fractures were identified by the IB Lab FLAMINGO AI technology in 111 out of 246 CT scans, and that only 49 (44%) were reported in the original radiology report (Spangeus et al., 2025). However, the EAG notes that patients had between 1 and 5 scans each, and therefore it is unclear how many patients had a missed VF diagnosis in standard of care.

Table 11: Measures of diagnostic accuracy (IB Lab) compared with a reference standard, reported per patient

Study	Reference standard	Anatomical location and scan modality	Technology (definition fracture)	Prevalence	False positive	False negative	True positive	True negative	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy	Kappa	PPV (95% CI)	NPV (95% CI)	AUROC
Nicolaes (Osteoporosis International, 2024; 143-152)	Sub-specialist radiologist	Abdominal, chest, thoracolumbar spine CT	Moderate or severe; Grade 2 or 3 (patient level)	13.1% (628/4,810)	-	-	-	-	0.944 (0.923 to 0.960)	0.932 (0.925 to 0.939)	0.933 (0.926 to 0.940)	0.749 (0.722 to 0.774)	0.675 (0.646 to 0.708)	0.991 (0.988 to 0.994)	0.938 (0.928 to 0.947)
Nicolaes (J Bone Miner Res. 2023; 1856-1866)	Initial triage by medical doctor then reference standard produced by experienced radiologists	Abdominal, chest CT	Moderate or severe; Grade 2 or 3 (patient level)	15.3% (297/1,943)	-	-	-	-	0.808 (0.762 to 0.851) Female: 0.868 (0.812 to 0.909) Male: 0.724 (0.635 to 0.795) Age 50-69 years: 0.833 (0.736 to 0.902) Age 70+ years: 0.797 (0.739 to 0.850)	0.945 (0.933 to 0.955) Female: 0.941 (0.923 to 0.957) Male: 0.948 (0.931 to 0.961) Age 50-69 years: 0.951 (0.936 to 0.964) Age 70+ years: 0.937 (0.918 to 0.952)	0.92 (0.91 to 0.93) Female: 0.93 (0.91 to 0.94) Male: 0.92 (0.90 to 0.94) Age 50-69 years: 0.94 (0.92 to 0.95) Age 70+ years: 0.91 (0.89 to 0.93)	0.72 (0.67 to 0.76) Female: 0.77 (0.72 to 0.82) Male: 064 (0.56 to 0.71) Age 50-69 years: 0.69 (0.60 to 0.76) Age 70+ years: 0.73 (0.67 to 0.78)	0.725 (0.675 to 0.770) Female: 0.774 (0.710 to 0.828) Male: 0.654 (0.568 to 0.734) Age 50-69 years: 0.636 (0.546 to 0.718) Age 70+ years: 0.775 (0.717 to 0.827)	0.965 (0.955 to 0.973) Female: 0.969 (0.954 to 0.979) Male: 0.962 (0.947 to 0.973) Age 50-69 years: 0.982 (0.971 to 0.991) Age 70+ years: 0.945 (0.926 to 0.989)	0.876 (0.852 to 0.898) Female: 0.905 (0.876 to 0.930) Male: 0.836 (0.790 to 0.869) Age 50-69 years: 0.892 (0.851 to 0.927) Age 70+ years: 0.867 (0.838 to 0.893)
IB Lab Internal Report [CiC]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Spangeus (Arch Osteopor, 2025; 42)	2 experienced radiologists.	CT for thoracic pathologies, aortic assessment, spinal imaging, and abdominal pathologies	Moderate or severe VFF; grade 2 or 3 (per scan)	45.1% (111/246 scans; not reported per patient)	-	-	-	-	0.86 (0.78 to 0.92) Female: 0.95 (0.89 to 1.00) Male: 0.75 (0.63 to 0.87) Age <85 years: 0.80 (0.70 to 0.90)	0.99 (0.96 to 1.00) Female: 0.98 (0.94 to 1.00) Male: 0.99 (0.96 to 1.00) Age <85 years: 0.99 (0.96 to 1.00)	0.93 (0.89 to 0.96) Female: 0.96 (0.92 to 0.99) Male: 0.89 (0.83 to 0.94) Age <85 years: 0.90 (0.85 to 0.95)	0.85 (0.78 to 0.91) Female: 0.93 (0.84 to 0.98) Male: 0.77 (0.64 to 0.87) Age <85 years: 0.80 (0.69 to 0.90)	0.98 (0.95 to 1.00) Female: 0.98 (0.95 to 1.00) Male: 0.97 (0.91 to 1.00) Age <85 years: 0.98 (0.93 to 1.00)	0.89 (0.84 to 0.94) Female: 0.95 (0.88 to 1.00) Male: 0.86 (0.78 to 0.93) Age <85 years: 0.86 (0.79 to 0.94)	0.92 (0.88 to 0.96) Female: 0.96 (0.93 to 1.00) Male: 0.87 (0.79 to 0.94) Age <85 years: 0.89 (0.83 to 0.95)

Study	Reference standard	Anatomical location and scan modality	Technology (definition fracture)	Prevalence	False positive	False negative	True positive	True negative	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy	Kappa	PPV (95% CI)	NPV (95% CI)	AUROC
									Age >85 years: 0.91 (0.83 to 0.98)	Age >85 years: 0.98 (0.95 to 1.00)	Age >85 years: 0.95 (0.91 to 0.98)	Age >85 years: 0.90 (0.81 to 0.97)	Age >85 years: 0.98 (0.94 to 1.00)	Age >85 years: 0.93 (0.86 to 0.99)	Age >85 years: 0.95 (0.90 to 0.98)
Nicolaes (Osteoporosis International, 2024; 143-152)	Sub-specialist radiologist	Abdominal, chest, thoracolumbar spine CT	Mild, moderate or severe; Grade 1, 2 or 3 (patient level)	33.7% (1,622/4,810)	-	-	-	-	0.626 (0.602 to 0.648)	0.935 (0.926 to 0.943)	0.831 (0.821 to 0.841)	0.598 (0.574 to 0.621)	0.831 (0.809 to 0.851)	0.831 (0.819 to 0.843)	0.781 (0.768 to 0.794)
Nicolaes (J Bone Miner Res. 2023; 1856-1866)	Initial triage by medical doctor then reference standard produced by experienced radiologists	Abdominal, chest CT	Mild, moderate or severe; Grade 1,2 or 3 (patient level)	20.9% (407/1943)	-	-	-	-	0.757 (0.714 to 0.798)	0.873 (0.855 to 0.889)	0.85 (0.83 to 0.87)	0.58 (0.54 to 0.63)	0.612 (0.569 to 0.653)	0.931 (0.918 to 0.943)	0.815 (0.790 to 0.837)

Abbreviations: AUROC, Area under the receiver operating characteristic curve; CI, Confidence interval; NPV, Negative predictive value; PPV, Positive predictive value.

Table 12: Measures of diagnostic accuracy (IB Lab) compared with a reference standard, reported per-vertebrae

Study	Reference standard	Anatomical location and scan modality	Technology (definition fracture)	Prevalence	False positive	False negative	True positive	True negative	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy	Kappa	PPV (95% CI)	NPV (95% CI)	AUROC
Nicolaes (Osteoporosis International, 2024; 143-152)	Sub-specialist radiologist	Abdominal, chest, thoracolumbar spine CT	Moderate or severe; Grade 2 or 3 (vertebrae level)	1.9% (899/48,584)	-	-	-	-	0.874 (0.853 to 0.896)	0.990 (0.989 to 0.991)	0.988 (0.987 to 0.989)	0.728 (0.706 to 0.746)	0.632 (0.605 to 0.659)	0.998 (0.997 to 0.998)	0.932 (0.921 to 0.943)
Nicolaes (J Bone Miner Res. 2023; 1856-1866)	Initial triage by medical doctor then reference standard produced by experienced radiologists	Abdominal, chest CT	Moderate or severe; Grade 2 or 3 (vertebrae level)	2.7% (663/24,930)	-	-	-	-	0.532 (0.494 to 0.569)	0.993 (0.992 to 0.994)	0.98 (0.98 to 0.98)	0.58 (0.55 to 0.62)	0.667 (0.625 to 0.705)	0.987 (0.986 to 0.989)	0.763 (0.745 to 0.783)
IB Lab Internal Report [CiC]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nicolaes (Osteoporosis International, 2024; 143-152)	Sub-specialist radiologist	Abdominal, chest, thoracolumbar spine CT	Mild, moderate or severe; Grade 1, 2 or 3 (vertebrae level)	5.4% (2,623/48,584)	-	-	-	-	0.576 (0.557 to 0.595)	0.990 (0.989 to 0.991)	0.967 (0.966 to 0.969)	0.639 (0.623 to 0.656)	0.762 (0.744 to 0.782)	0.976 (0.975 to 0.977)	0.783 (0.773 to 0.792)
Nicolaes (J Bone Miner Res. 2023; 1856-1866)	Initial triage by medical doctor then reference standard produced by experienced radiologists	Abdominal, chest CT	Mild, moderate or severe; Grade 1,2 or 3 (vertebrae level)	4.3% (1,066/24,930)	-	-	-	-	0.470 (0.438 to 0.499)	0.987 (0.986 to 0.988)	0.96 (0.96 to 0.97)	0.52 (0.49 to 0.55)	0.616 (0.582 to 0.648)	0.977 (0.974 to 0.979)	0.728 (0.714 to 0.744)
Spangeus (Arch Osteopor, 2025; 42)	2 experienced radiologists.	CT for thoracic pathologies, aortic assessment, spinal imaging, and abdominal pathologies	No or mild VF, moderate or severe VF	NR (out of 2,136 vertebrae)	-	-	-	-	0.71 (0.64 to 0.78)	0.99 (0.98 to 0.99)	0.96 (0.95 to 0.97)	0.74 (0.69 to 0.80)	0.82 (0.76 to 0.88)	0.97 (0.97 to 0.98)	0.85 (0.81 to 0.89)

Abbreviations: AUROC, area under the receiver operating characteristic curve; CI, confidence interval; NPV, Negative predictive value; PPV, Positive predictive value

Failure rate or inconclusive AI reports

All 3 studies reported on failure of the IB Lab technology, however definitions varied across studies (Table 13); 1 reporting that images were not readable, 1 reporting not readable due to missing CT slides, and 1 commercial in confidence [redacted]

Table 13: Failed or inconclusive reports (IB Lab)

Study	Definition of error	Error rate
Nicolaes (Osteoporosis International, 2024; 143-152)	Not readable by BoneBot (poor image quality or missing slices in CT series)	1.0% (52/5,195) [Noting that an additional 13 scans were not interpretable by technology or by radiologist due to poor quality image]
IB Lab (CiC, 2023)	[redacted]	[redacted]
Nicolaes (J Bone Miner Res. 2023; 1856-1866)	Not readable by BoneBot	2.9% (57/2,000)

Healthcare professional user acceptability

Not reported.

Changes to clinical management

As the 3 studies using an IB Lab technology were all retrospective studies, this outcome was not reported.

Health-related quality of life

Not reported; no RCTs identified.

Time to run the software

The study by Nicolaes et al. 2024 reported that software required on average 2 minutes run time on each scan using graphics processing unit acceleration (Nicolaes et al., 2024).

5.2 Adverse events and clinical risk

False positives and false negatives are reported in the 'diagnostic accuracy' subsections above, and technical failures (that is failure of the AI to process or analyse an image) are reported in the 'failure rate or inconclusive AI report' subsections above. There was a lack of detail reported in the published evidence, therefore the EAG was unable to determine whether there is a common characteristic of images which fail to be processed or analysed by the AI technologies.

The EAG searched [MHRA Field Safety Notices] from 01 January 2020 to 04 March 2025 and found no mention of any of the company or technology names listed in the Final Scope.

5.3 Clinical evidence summary and interpretation

The main aim of this assessment was to determine the evidence available and identify evidence gaps to support future evidence generation. In line with the EVA process and methods, the EAG conducted a rapid review and identified 22 studies within this assessment. This included abstracts and information provided academic or commercial in confidence, which may lack peer-review. The majority of evidence described retrospective application of the AI when applied to already acquired images and compared this with a reference standard. The EAG judged this approach to be appropriate when considering the true diagnostic accuracy of the technologies (results from AI require human interpretation as per technology instructions for use, and Clinical Expert advice that diagnosis of vertebral fracture would be made by a radiologist or reporting radiographer in the NHS). This evidence demonstrated that the AI technologies were able to detect additional moderate to severe vertebral fracture (as confirmed by a reference standard) which were not reported in the original radiology report, with generally high specificity across all technologies (where evidence was available). The sensitivity, specificity, failure rate and vertebral fracture prevalence will depend on the eligible population as defined in the AI technology Instructions for Use (IFU); therefore, each technology should be considered separately.

For the technologies that process X-ray images:

- **Annalise Enterprise CXR:** 5 studies were identified, 2 reported sensitivity (range: 89% to 100% per patient) and specificity (range: 89% to 95.4% per patient) against a reference standard, with low failure rate of being able to process the image (range: 0% to 0.2%). No evidence was specifically related to Annalise Container CXR; however, the EAG acknowledges that this is a version of Enterprise which is packaged to optimise it for running on selected third-party marketplaces / platforms. Annalise Container uses the same CXR AI model as Annalise Enterprise, and the company confirms that evidence should be considered generalisable between technologies.. One abstract reported the application of AI to AP, PA or LAT chest X-rays, one full publication reported front and lateral chest X-ray, one academic-in-confidence study [redacted]; the projection was not explicitly reported in the remaining two abstracts.

- **BoneView:** 1 study identified, with sensitivity of 63.2% and specificity of 96.7% per vertebrae (results per patient were not reported), where 1.4% of lumbar spine and 4.5% of thoracic spine images were unable to be processed due to incorrectly classifying the image.

- **TechCare Spine:** No evidence was identified for this technology.

For the technologies that process CT images:

- **BriefCase-Triage**: 1 study including a technology by Aidoc Medical (name of technology not specified) was identified, with sensitivity of 85.2% and specificity of 92.3% per patient, and low failure rate of 0.6%.

- **CINA-VCF**: 3 studies were identified (none explicitly stating CINA-VCF Quantix), 2 of which reported sensitivity (range: 92.3% to 95.2% per patient) and specificity (range: 91.7% to 92.9% per patient) against a reference standard. Failure rate was not explicitly reported across identified evidence.

- **HealthVCF, HealthOST**: 8 studies were identified (none explicitly stating HealthOST), 6 of which reported sensitivity (range 48.3% to 79% per patient) and specificity (range: 81.2% to 98.5% per patient) against a reference standard, with a failure rate between 5.6% and 9.4%.

- **IB Lab FLAMINGO**: 4 studies were identified (2 using BoneBot which the company confirmed was identical to IB Lab FLAMINGO), all reporting sensitivity (range: 80.8% to 94.4%) and specificity (93.2% and 99%), and failure rate (range: 1.0% to [redacted]%).

There was limited prospective evidence (3 studies), and limited evidence from a UK setting (2 abstracts and 1 poster). Patient consent was not explicitly sought in any of the included prospective studies to apply an AI technology to their diagnostic image for reasons other than their direct care or where clinically indicated. One study (Bendtsen and Hitz, 2024) explicitly stated that informed consent was not obtained from patients as it was approved as a quality assurance study with retrospective collection of data and intervention by the local data committee. The remaining two studies (Connor et al. 2024 and Connacher et al. 2019) did not explicitly state whether patient consent was sought however the EAG notes that both these studies are abstracts only. Most included studies reported that images were anonymized or de-identified prior to processing. Regulatory approvals for all AI technologies require human interpretation of the outputs. Therefore, the impact of implementing AI technologies within the pathway and capacity on radiologists and reporting radiographers needs to be considered. The EAG notes that the reference standard described in the included evidence varied. Several identified studies have highlighted that the AI technologies may produce false positives which can be caused by intervertebral osteochondrosis, Schmorl's nodes, or congenital abnormalities. Therefore, the direct impact of integrating AI technologies for the opportunistic detection of vertebral fractures prospectively remains unclear due to the impact on workflow, on downstream confirmatory review by a radiologist, and the need for further ionising radiation exposure and treatment.

Many studies lacked detailed reporting on the configuration and version of AI technology used, which may impact the diagnostic accuracy. One Clinical Expert raised that transparent description of what the AI is doing is required (for example do they consider adjacent vertebrae above and below, two or more adjacent vertebrae, how they handle first or last vertebrae which have none above or below, how do they deal with scoliosis or tilted spines); as this may impact interpretation of results. The EAG recommends that transparent reporting of the inputs, patient and diagnostic imaging characteristics used for the training and validation for each version of the AI technology should be available in the public domain. The EAG judged this to be important when considering the generalisability of results to the UK NHS and technology performance in a real-world setting given the impact algorithm development of variation in image quality, different scanner manufacturers, noise characteristics, and patient characteristics. The EAG notes that previous diagnoses of vertebral fractures and treatment for previous fractures was not reported across included evidence, therefore sensitivity and specificity may be overestimated. An implementation consideration may include the time and resources to train and validate each new version of the AI technology released. Users should be notified of updates and ensure that staff are adequately trained.

Although there is a wealth of data available from the various sources described above, none of these sources fully meet the decision problem to accurately describe the number or proportion of CT scans or X-rays resulting in an incidental identification of a vertebral fracture in an NHS setting and how these impact on patient outcomes, including quality of life, and subsequent healthcare use. None of the routinely available datasets currently document the use of AI technologies in the diagnosis of vertebral fractures. The Fracture Liaison Service database may be well placed to expand its data collection to better quantify the use of imaging and AI, and routes to diagnosis of vertebral fracture in the future.

6. Economic evidence

6.1 Existing economic evidence

Search strategies developed for identifying clinical effectiveness evidence and AI technologies with an additional filter developed by CADTH (CADTH, 2016) were used as the EAG economic search strategy to identify evidence on economic evaluations and models. Please see Section 4.1 and Appendix A1 for details of the literature searching.

A total of 269 records were identified through the economic search strategy, with an additional 1 record identified from the clinical evidence search. After removing 81 duplicate records, 185 unique records were screened based on their titles and abstracts by a single reviewer (HW). A random sample of 10% of the records was checked by a second reviewer (NM). From this screening, 6 records were selected for full-text retrieval and reviewed independently in duplicate (HW,NM). Of these, 4 studies were excluded as they were not relevant to vertebral fractures and 1 study was excluded because it was not an economic evaluation. No economic evaluations relevant to the decision problem for the technologies of interest were identified. The sift was then widened to include studies that could aid the development of a de novo model (not specific to the interventions listed in the scope) to help with model structure and parameterisation. From this sift, a further study from the economic search economic search (Dalal et al., 2022) and 1 study from the clinical evidence (Curl et al., 2024) were used inform the economic modelling. A summary of these 2 papers is in Table 14. A summary of full papers which were excluded are summarized in Appendix A5.

Please see the PRISMA flow diagram of the search and screen process in Appendix A3.

Table 14: Key studies selected for developing and parameterising 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
Curl (J Am Coll Radiol, 2024; 1489-1496) Cost-utility study US	Intervention: Automated opportunistic screening for osteoporotic vertebral compression fractures (OVCFs) using AI-based software on existing lateral chest and abdominal X-rays. Comparator: Standard of care.	Study population: Patients with available radiographs and no existing osteoporosis diagnosis. Base case scenario: a representative 71-year-old woman without known osteoporosis and lateral radiographs of the chest and abdomen available.	Primary outcome: Estimated average per-patient costs and average per-patient QALYs, to obtain a patient level ICER. Average per-patient total lifetime costs and total lifetime QALYs gained or lost within each strategy. Key findings: The screening strategy was found to be cost-effective, resulting in lower costs and increased QALYs compared with the usual care strategy from a limited societal perspective (although what this entails was unclear from the study). The study indicated they included direct healthcare costs (reimbursement amounts) and long-term care after repeat fractures.	This study was used to inform the model development, including a cost per patient for a generic AI technology (£7.36 reflecting 2024 price year, converted from $10, using 2022 price year, conversion and inflation applied using CCEMG - EPPI-Centre Cost Converter).

Study name, design and location	Intervention(s) and comparator	Participants and setting, length of follow-up	Relevant outcomes and key findings	EAG comments
Dalal (Aging Clin Exper Res, 2022; 1909-1918) Qualitative research (a structured expert elicitation exercise) UK	Intervention: An automated approach for the incidental identification of VFFs on existing medical images visualising the spine, using machine learning-based computer-aided diagnostic systems. Comparator: Not applicable	Seven experts from different parts of the UK. Hypothetical population: population of individuals aged 70 years who have been referred for a CT scan in the NHS.	Estimating mean values and the uncertainty associated with key parameters for a cost-effectiveness analysis (CEA) of a machine-learning-computer-aided diagnostic system including: - Probability of VFF being correctly reported by the radiologist - Probability of absence of VFF being correctly assessed by - the radiologist - Probability of being referred for management when a VFF is identified by the radiologist - Probability of having a DEXA scan after GP referral	This study was used to inform the treatment pathway and the economic model structure. Additionally, this study informed the sensitivity and specificity for CT scans in the standard of care arm.

Abbreviations: AI, Artificial intelligence; CEA, Cost-effectiveness analysis; DEXA, Bone density X-ray scan (dual energy X-ray absorptiometry); OVCF, Osteoporotic vertebral compression fracture; QALYs, Quality-adjusted life-years; VFF, Vertebral fragility fracture.

Relevant economic literature to inform the economic model

From the systematic search, two relevant studies were identified. One study was a cost-utility analysis (Curl et al., 2024) estimating cost-effectiveness over a lifetime time horizon. It compared automated opportunistic screening using AI technology, with standard of care, in patients with available lateral chest and abdominal X-rays. The AI technology used was not explicitly reported, but one author disclosed employment with Aidoc Medical. The initial diagnosis and treatment decision were modelled using a decision tree, with transition into a Markov model for longer term outcomes. The structure of the model informed the EAG in development of their own de novo model, although the EAG considered it more complex than needed to address the decision problem of this EVA. The EAG did, however, use the cost per AI scan of $10 (from 2022 price year, converted to £7.36 using CCEMG - EPPI-Centre Cost Converter to reflect 2024 price year) for their generic AI technology base case.

A second study (Dalal et al., 2022) summarised a structured expert elicitation exercise to estimate probabilities and their associated uncertainties for key parameters related to the diagnosis and management of VFFs in the current UK care pathway. Data was collected from seven experts in VFF diagnosis and management from across the UK, and their responses were pooled. The four parameters included in the study were: (1) the probability of a VFF being correctly reported by a radiologist, (2) the probability of the absence of VFF being correctly assessed by a radiologist, (3) the probability of referral for management when a VFF is identified, and (4) the probability of having a DEXA scan after GP referral. The EAG used the estimated sensitivity and specificity of a radiologist in the SoC arm of their model for comparison with all AI technologies.

In addition to these two studies, the EAG identified five additional sources (from scoping and evidence trawling) which included 1 abstract, 2 full publications and 2 NICE technology appraisals that were relevant to different parts of the decision problem. These sources contributed information on model assumptions, utility and costs but were not specific to the interventions included in this assessment. A conference abstract by (Dalal et al., 2021) described a de novo decision analytic model over a lifetime horizon, from an NHS England perspective. The model was a decision tree with a discrete event simulation for a cohort of 400,000 70-year-olds, comparing identification of VFFs from existing CT scans by AI-based algorithm, ASPIRE, versus radiologist, and referral to start bisphosphonate treatment. One-way, two-way, scenario, threshold, and probabilistic sensitivity analyses were used to quantify uncertainty. The AI algorithm identified 47,029 more VFFs than SoC, resulting in an additional cost of £8,681,804 (assuming 2014 price year, GBP) and an increase of 139 QALYs, that is, an incremental cost per QALY of £62,459. The ICER was £185 for each additional VFF identified. The authors identified that the key economic drivers were the specificity and unit cost of the AI technology, the radiologists' time averted by using the AI technology, and the costs associated with referral to an FLS. The EAG used these key drivers to inform their own sensitivity analysis for this EVA.

NICE TA464 (2017, updated 2019) used a discrete event simulation over a time horizon up to age 100 years, to report on the cost-effectiveness of bisphosphonates for treating osteoporosis. The model applied acute care costs over the year after a fracture, and chronic costs over the longer term of the model. Key costs for the acute period were derived from the studies by Gutierrez et al. (2012), and Borgstrom et al. (2006). Gutierrez et al. (2012) used a GP database to quantify healthcare use attributable to fractures, providing data on hospitalisations, accident and emergency visits, GP visits, prescribed medications, number of medications and referrals. The study by Borgstrom et al. 2006 quantified the cost of home help for the year following fracture (Borgstrom et al., 2006). These costs (inflated to 2024 price year) were used by the EAG to estimate the costs associated with, and management of. vertebral fractures after identification by either AI technologies or SoC. Because the EAG model covers only a 1-year time horizon, it was not necessary to consider the longer-term costs of residential care arising in the chronic period following a fracture.

Svedbom et al. (2018) as part of the ICUROS study (included in TA791) investigated the health-related quality of life (HRQoL) outcomes associated with low-energy fractures, including vertebral fractures, over an 18-month period. The study population were aged 50 years or more and had vertebral fractures confirmed by X-ray. The EQ-5D-3L was used to measure quality of life at pre-fracture and four time points, post fracture. Time points were within 2 weeks after fracture (including a retrospective evaluation of pre-fracture HRQoL) and at 4, 12, and 18 months after fracture. It found that vertebral fractures had a prolonged negative impact on HRQoL. Persistent pain, reduced mobility, and psychological distress were key contributors to the diminished HRQoL in vertebral fracture patients. The utility estimates and decrements in this study were utilised in the EAG economic model, in cases where there was a fracture, it was detected and the patient was referred for vertebral fracture management, see section 6.2.5.

Gutiérrez et al. (2012) investigated the incremental cost of fractures, including 1,471 vertebral fractures in a cohort of postmenopausal women aged 50 years or older in the UK. The cohort had a mean age of 72 years and comparisons were drawn between a fracture cohort and a matched non-fracture cohort (Gutierrez et al., 2012). The total average cost over 12 months in the vertebral fracture cohort was £2,180 (2007 price year, GBP), and in the non-fracture matched group was £1,028. Therefore, the vertebral fracture had an incremental cost of £1,152. Gutiérrez et al. (2012) study also provided the mean incremental difference for all of the fracture management cost components. Of the £1,152 incremental cost of a fracture, hospitalisation accounted for 54% of the cost, and medication for a further 29%. Seventy per cent (70%) of these annual costs were concentrated in the first 6 months following the fracture. Gutiérrez et al. (2012) also compared their cost estimates with other published UK-based studies. They reported that their estimate of a total cost of £2,180 (not only the vertebral fracture cost) was within the range of estimates from other research.

NICE TA791 (2022) reported cost-effectiveness of romosozumab compared with existing treatment for severe osteoporosis in postmenopausal women at high risk of fracture. The cost-effectiveness analysis was conducted from an NHS and Personal Social Services (PSS) perspective using a Markov microsimulation model with a lifetime time horizon. In TA791, the acute costs associated with vertebral fracture, during the first year after a fracture were taken directly from Gutiérrez et al. (2012), a UK based study described previously, and the main source of estimating costs in TA464. The costs of fractures in subsequent years were from TA464. All of the costs in the first and subsequent year were updated to a 2020 price year using consumer price indices from the Office for National Statistics (ONS) which resulted in £2,131 and £361, respectively. As described above, Gutiérrez et al. provided total and incremental fracture costs in the first year but the company analysis reported in TA791 used the total cost value rather than the incremental cost value. This results in first year management costs being overestimated as a substantial proportion of total costs would have been incurred even if a fracture had not occurred. The EAG notes that the incremental cost estimate provided by Gutiérrez et al is more appropriate in a base case analysis.

The probabilities of discharge to institutional care overall and by age group were sourced from Nanjayan et al. 2014 (Nanjayan et al., 2014) which was in line with TA464.

In TA791, the utility value associated with fracture was estimated using utility multipliers for fractures from the International Costs and Utilities Related to Osteoporotic Fractures Study (ICUROS) combined with UK general population values for the EQ-5D-3L (2014). The multiplier used in the TA791 is a confidential value, however, it is indicated that while it is different from the value reported in TA464 it is similar to the value from another source (ID901, 2018). The value from this latter source is equal to 0.68 for vertebral fracture in the year after fracture.

6.2 Early economic model

A de novo economic model was developed by the EAG to investigate the potential cost-effectiveness of opportunistic detection of vertebral fractures by AI-assisted technologies compared with current SoC in a hospital setting. The analysis was from an NHS and Personal Social Services (PSS) perspective, as per the NICE reference case (NICE PMG36, 2022; updated 2023), and used a one-year time horizon. This timeframe is believed to be sufficiently long enough to capture the immediate impact of opportunistically detecting vertebral fracture including those initially missed by SoC. The EAG did not construct a longer-term model structure that would consider any ongoing longer-term events relating to vertebral fractures or their management noting that NICE guidance on osteoporosis: risk assessment, treatment and fragility fracture prevention is currently under development, GID-NG10216). Implications of this are discussed in section 6.4 with a draft model structure to consider longer term impacts presented in section 8.3.

6.2.1 Model structure

The general structure allowed comparison of the original reporting radiographer *without* AI assistance (standard of care) with the original reporting radiographer *with* AI assistance, for each technology separately. This structure enabled the EAG to explore whether each AI technology (with different eligibility criteria) has the potential to be cost effective compared with current practice. As this is an early economic evaluation aiming to explore the potential value for money as part of a NICE EVA, the data inputs (such as sensitivity and specificity for each AI technology) and outputs (such as cost effectiveness estimates) should not be seen as definitive but rather illustrative. Instead, they were applied from the clinical evidence available to consider key cost and QALY drivers and key areas of uncertainty.

Figure 1 presents a decision tree model incorporating prevalence, sensitivity and specificity, and cost per diagnostic strategy (AI-assisted versus unassisted) based on the existing literature of similar topics and including feedback from Clinical Experts (see Appendix D2). The EAG model assumes that a patient has a radiographic image involving the spine taken for any cause other than vertebral fractures. It goes on to assume that this image can then be used for opportunistic detection of vertebral fractures.

For the SoC arm, the model first incorporates the probability of the patient having a vertebral fracture (VF present) or not (1-VF present). If the patient has a vertebral fracture, it can either be detected opportunistically or not by the reporting *radiographer* within 24 hours of the image being taken. If a vertebral fracture is opportunistically detected (test positive) then a *radiologist* specialising in musculoskeletal imaging will review the image to correctly confirm its presence. As many images are chest, abdomen or pelvic X-rays or CT scans, the EAG has assumed that a proportion of test positive cases may require an additional spine X-ray scan to confirm the presence of a vertebral fracture. The EAG notes that an additional X-ray may not be required in all cases (especially where the initial diagnostic image was a CT where resolution is higher), and that subsequent CT or MRI imaging may be required in some instances (see Clinical Expert responses to Question 6c in Appendix D1). The EAG notes that a second review may not be required. To address this variation across clinical scenarios, the EAG has varied the proportion requiring additional imaging in sensitivity analysis (increased and decreased).

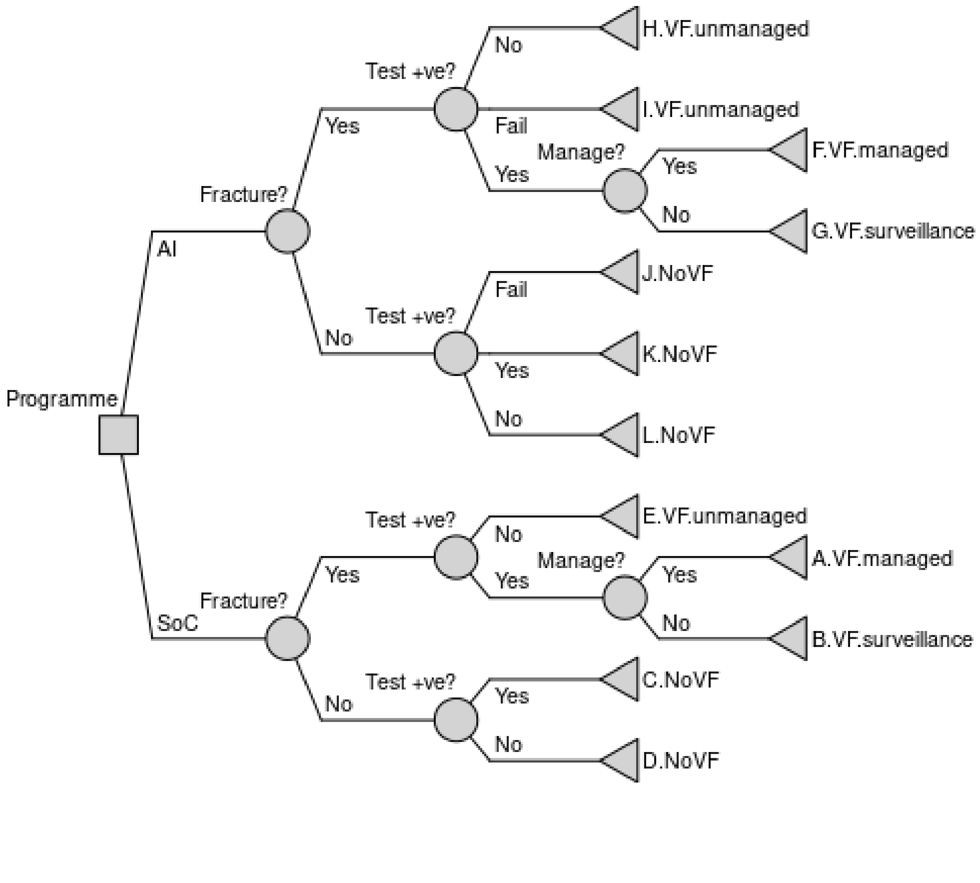
Where vertebral fracture is confirmed (true positive), a proportion will undergo additional investigations (for example DEXA scan to determine bone mineral density to help determine whether osteoporosis is the cause of the fracture) or initiate vertebral fracture management. If the vertebral fracture is not confirmed (false positive) no costs in addition to the second review by a radiologist are added. If a vertebral fracture is present but not detected (false negative, interpreted by the EAG as a missed opportunity of detecting vertebral fracture) no additional costs are incurred and there is no impact on QALYs. If a vertebral fracture is not present and not detected (true negative) no additional costs are incurred and there is no impact on QALYs. The AI-assisted arm follows a similar structure to the SoC arm, but includes additional costs, different sensitivity and specificity to the SoC, an extra branch with probability of the technology being unable to process the image and produce an output (test failure).

Costs and utilities were attached to appropriate chance or terminal nodes in the decision tree for both the SoC and AI-assisted strategies based on the existing studies in the literature (see Section 6.2.4 and 6.2.5 for details).

The EAG ran the economic model separately for each technology compared with SoC. While the EAG notes that there was limited information regarding

SoC (requiring several assumptions to run the model), the EAG felt modelling each technology separately was preferable considering that the eligibility criteria differed for each technology, and therefore the starting population characteristics including prevalence of vertebral fracture may differ by technology. An alternative model structure that starts with a broad population of which a proportion would be eligible for each technology could be specified and this would be of value for the definitive evaluation of these technologies, which may be possible at some point in the future. The current model would support future pairwise comparisons of a technology against SoC when data does become available.

Figure 1: General structure of the economic model (Abbreviations: AI, artificial intelligence; SoC, standard of care; VFF, vertebral fracture)



6.2.2 Model assumptions

EAG base case model assumptions included:

- The model assumes a CT or X-ray was used for the opportunistic detection of vertebral fracture for the SoC in the base case analysis. Sensitivity and specificity data are only available for the SoC arm using CT scan; and therefore, it was assumed that the sensitivity and specificity for CT were the same for X-rays (explored further in sensitivity analysis). The EAG notes that the sensitivity of single radiograph is likely lower than that of multi-slice CT. Therefore, the economic model may overestimate the number of additional vertebral fracture detected in the SoC X-ray arm, and therefore greater benefits when using AI to process X-ray images may be possible. Due to the uncertainty associated with the diagnostic accuracy of standard of care, the EAG has varied sensitivity of SoC in the opportunistic detection of vertebral fractures in the economic model univariate sensitivity analysis.

- To reduce complexity, and due to lack of data, the model assumes the opportunistic detection (sensitivity and specificity of AI technologies, current care) are applied per-patient once within the short time horizon. Individuals may receive more than one episode of imaging, and each episode may be used for opportunistic detection. This would incur the cost of the technology, potentially reduce the missed opportunities but also increase the false positive rate. The net impact on cost-effectiveness is uncertain.

- Given the subscription fee for many of the AI technologies is volume-based, the EAG calculated a cost-per-scan and assumed 65,000 chest X-ray images per year and 6,500 chest or abdominal CT images per year per centre (using values obtained from the DID, see section 3.3.3). Lower and higher throughput was explored within sensitivity analyses with changes to the cost-per-scan made where appropriate.

- Any image flagged as having a potential vertebral fracture (by the initial reporting *radiographer*, with or without AI assistance) may have a secondary review by a *radiologist* specialising in musculoskeletal imaging. The EAG notes that in some cases a secondary review may not be required. However, as the proportion of initial scans requiring review was unknown, the EAG did not include this within scenario analysis. The EAG notes the variation in secondary review of images could be modelled through changes in the proportion requiring additional spine X-ray, which has been considered in sensitivity analysis.

- As the initial diagnostic images include the spine but not for vertebral fracture detection (for example chest, abdomen, pelvic X-rays and CT scans) the EAG has assumed that a small proportion of test positive patients may require an additional spine X-ray to further investigate. Including this within the economic model will also enable the EAG to count the number of additional unnecessary scans as a consequence of false positives. The EAG has varied the proportion required additional spine X-ray within sensitivity analysis.

- The EAG assumes that only a proportion of patients who had the vertebral fracture confirmed by a radiologist would be referred for treatment, or further investigation as to the cause of the fracture (for example DEXA scan, referral to FLS) within the 1-year time horizon. This is due to multiple reasons, including minor fracture or resource constraint (due to limited access to DEXA scan facilities) or because treatment has already been initiated. The impact of this on cost-effectiveness is unknown as no evidence was identified on the natural history of vertebral fracture or the proportion requiring treatment and the impact of treatment on the natural history of vertebral fractures.

- For false negatives (that is, disease positive but not detected by initial reporting radiographer or AI technology and therefore no additional review and no additional scan), the EAG has assumed no further management costs or utility decrements within the 1-year time horizon. It is currently unclear the implication of these because the EAG has not identified data on the natural history of vertebral fracture, the proportion requiring treatment and how treatment changes the natural history. Nevertheless, the EAG estimated the number of missed opportunities for managing vertebral fracture, and there are fewer in the intervention arm. Therefore, we would expect that had the EAG been able to capture these impacts they would reduce the ICER.

- Only the patients with vertebral fracture detected and managed will experience a quality-of-life gain over the one-year time horizon modelled. No other utility gains are considered. Using this approach, differences in QALYs solely relate to differences in treatment rates of vertebral fracture for AI technologies over SoC.

6.2.3 Clinical parameters

Clinical parameters applied to all arms are summarised in Table 15, clinical parameters which varied by technology are summarised in Table 16.

The prevalence is defined as the proportion of people who have a radiographic image involving the spine taken for reasons other than vertebral fracture detection who have an undetected vertebral fracture prior to imaging. The EAG acknowledges that there is no consensus on the definition of prevalence used in the vertebral fracture literature, and that the wide range of values (from 7.3% to 52% in the clinical evidence described in Section 5.2) implies considerable uncertainty. Therefore, the EAG used the mid-point of this range (29.7%) in the base case and tested the extreme values in sensitivity analyses.

Sensitivity and specificity for the SoC for CT images were sourced from Dalal et al. (2022), in which expert opinions regarding sensitivity and specificity of vertebral fracture detection and rate of referral for vertebral fracture management (if positive) were elicited from 7 experts through organised elicitation tasks. This study was identified from the clinical search and was considered as an appropriate source for diagnostic accuracy reflecting the current clinical practice of opportunistic vertebral fracture detection in the UK. However, diagnostic accuracy of standard of care in opportunistic detection of vertebral fractures was varied in sensitivity analysis due to the uncertainty associated. Sensitivity and specificity for the AI technologies were directly extracted from the clinical evidence base described in Section 5.2. Where multiple values of sensitivity and specificity data were available for a specific technology: (a) studies reporting diagnostic accuracy per patient, those defining vertebral fracture as grade 2 or 3, those comparing with a radiologist specialising in musculoskeletal imaging (considered the reference standard, see Appendix D) were used as the base case value; (b) the other values of sensitivity and specificity were used in sensitivity analyses. The EAG notes that the sensitivity and specificity data were not available from the company for the AI technology TechCare Spine, therefore this technology was not included in economic modelling. This technology can only be applied to lateral thoracic or lumbar spine X-ray images, and therefore it is unclear what the number of images that would be eligible for this technology is.

Where possible, the EAG selected the failure rate from the same study which provided the sensitivity and specificity values used in base case. If this was not possible then the study with the largest sample size reporting failure was used. Other values for failure rate for the AI technologies were applied in the sensitivity analysis.

The EAG also included a “generic AI” as a scenario using data from the study by Curl et al. 2024 (AI sensitivity: 0.598, AI specificity: 0.999, cost per scan $10 converted to £7.36), and an estimated failure rate of 1% to enable transparent reporting of results for decision making purposes.

The EAG assume that not all patients who tested positive were referred for vertebral fracture management, and the probability of receiving vertebral fracture care conditional on positive test was also sourced from the expert elicitation (n=7 experts) reported by Dalal et al. (2022). Due to the uncertainty associated with this parameter, the proportion referred for vertebral fracture management was increased within sensitivity analysis.

Table 15: Main clinical parameters applied to all branches

Variable	Value	Source	EAG comment
Prevalence	0.297	From clinical evidence base (studies using AI included in this study)	Mid-point of prevalence reported in the clinical evidence (between 7.3% and 52%)
Sensitivity for SoC	0.253	Dalal et al. (2022)	Data is for CT, but due to lack of data applied to X-ray as well (explored in sensitivity analysis).
Specificity for SoC	0.891	Dalal et al. (2022)	Data is for CT, but due to lack of data applied to X-ray as well (explored in sensitivity analysis).
Probability of requiring an additional spine X-ray following test-positive	0.10	EAG assumption	Assumption (explored in sensitivity analysis).
Probability of referred for vertebral fracture management conditional on positive test	0.146	Dalal et al. (2022)	Data is for CT (expert elicitation of 7 experts), but due to lack of data, applied to X-ray as well (explored in sensitivity analysis). Dalal et reported: *"The low probability of being referred for management seems to vary and depends on different factors. It may be explained by the perceived lack of importance of VFFs and osteoporosis, particularly when compared to other diseases."* The EAG increased this proportion in sensitivity analysis.

Abbreviations: AI, Artificial intelligence; SoC, Standard of Care; VFF, Vertebral fragility fracture.

Table 16: Clinical parameters which vary by technology

Variable (image modality)	Sensitivity	Specificity	Source, EAG comment	Failure rate	Source, EAG comment
Generic AI base case (X-ray)	0.598	0.999	(Curl et al., 2024) The EAG notes that the sensitivity is lower, and specificity higher than the base case for all named technologies included in this assessment but is included to transparently demonstrate key drivers and uncertainties in economic model.	1.0%	EAG assumption
Annalise (X-ray)	0.893	0.892	(Ghatak et al., 2024); which reported per patient based on 595 chest X-ray (frontal and lateral) see clinical evidence in Section 5.1.1. Clinical Experts have stated that it would be rare in NHS practice to obtain a frontal and lateral X-ray (see Appendix D3), therefore generalisability of this evidence is unclear.	0.2%	(Ghatak et al., 2024); based on 596 chest X-ray (frontal and lateral) see clinical evidence in Section 5.1.1.
BoneView (X-ray)	0.724	0.942	(Oppenheimer et al., 2024); on thoracic and lumbar spine X-rays including positive and doubt as fracture positive by the AI, when applied to lateral and AP projections, reported per-vertebrae; see clinical evidence in Section 5.1.2. Per patient results not available.	2.5% (13/512)	(Oppenheimer et al., 2024); when combining 1.4% (5/357) failure in lumbar spine and 5.1% (8/155) failure in thoracic spine X-rays; see clinical evidence in Section 5.1.2.
TechCare Spine (Lateral X-ray)	NR	NR	No evidence identified, not included in economic model.	NR	No evidence identified, not included in economic model.
BriefCase-Triage (CT)	0.852	0.923	(Wiklund et al., 2024); when applied to abdominal CT, considering grade 2 and grade 3 only, reported per patient; see clinical evidence in Section 5.1.3.	0.6% (7/1112)	(Wiklund et al., 2024); combining failure to upload to AI and failed analysis; see clinical evidence in Section 5.1.3.
CINA-VCF Quantix (CT)	0.952	0.929	(Dai et al., 2025); when applied to 474 chest, abdomen, axial or sagittal acquisition CT, reporting VCF grade 2 or 3, per patient; see clinical evidence in Section 5.1.4.	1%	No data source explicitly reported failure rate but the EAG has assumed 1% failure rate in base case for this technology noting that 1% (17/1700) did not have the vertebral height loss calculated due to cement or hardware being present. This is not directly AI failure to process the image, however is used as a surrogate in the absence of better evidence.
HealthVCF and HealthOST (CT)	0.738	0.927	(Pereira et al., 2024); based on 899 chest and abdominal scans, detecting moderate or severe VCF (grade 2 or 3); reported per scan, see clinical evidence in Section 5.1.5.	5.6% (54/964)	(Pereira et al., 2024); based on 899 chest and abdominal scans where reasons of failure defined. Higher failure rates (6.0%, 9.4%) will be explored in sensitivity analysis (Kolanu et al., 2020; Page et al., 2023).
IB Lab FLAMINGO (CT)	0.944	0.932	(Nicolaes et al., 2024) based on 4,810 abdominal, chest and thoracolumbar spine CT, moderate or severe (grade 2 or grade 3), reported per patient; see clinical evidence in Section 5.1.6.	1.0% (52/5,195)	(Nicolaes et al., 2024); based on 4,810 abdominal, chest and thoracolumbar spine CT. Higher failure rates (2.9%, 4.8%) will be explored in sensitivity analysis (IB Lab, 2023; Nicolaes et al., 2023).

Abbreviations: AI, Artificial intelligence; AP, Anterior-posterior; SD, standard deviation; VCF, Vertebral compression fracture.

6.2.4 Resource use and cost parameters

Since the time horizon in this economic modelling is one year, the cost and QALY payoffs were used without applying discounting, as per NICE recommendation. All costs were inflated to 2024 by using web-based conversion tool Campbell and Cochrane Economics Methods Group - Evidence for Policy and Practice Information Centre Cost Converter (CCEMG - EPPI-Centre Cost Converter). This included converting reported costs in Euros (EUR) or US dollars ($) to GBP.

6.2.4.1 Initial reporting radiographer review

The initial reporting radiographer or radiologist review of the diagnostic image is considered to be the same for standard of care and intervention (with AI) arms. Therefore, for simplicity this has not been included in the economic model.

6.2.4.2 AI costs

A summary of costs associated with the AI technologies included in the economic model (including a generic AI base case) are described in Table 17. For the branches implementing AI technology, the AI cost was decomposed to product subscription, implementation, integration, training and maintenance costs. Assumptions were made to facilitate the calculation of the AI costs:

- Product subscription: for volume-based price provided by the technology companies, the EAG assumed 65,000 scans per site per year for technologies which process X-ray images (for chest and abdomen), and 6,500 scans per site per year for technologies which process CT scans in the base case. These assumptions are based on the median value of chest and abdominal X-ray or CT per organization from DID statistics (2023-2024). Additional values representing the 25% quartile, 75% quartile and maximum number of scans (using data from DID) were explored in sensitivity analyses.

- Only one company (IB Lab) stated the need for a dedicated server with minimum requirements defined. The company stated that the hardware costs would be incurred by the NHS and would be additional to the software costs. The EAG notes that the cost of server was only applied to one technology, but costs were small and therefore had little impact on the cost per scan.

- Training: 5 out of 7 companies stated that training is included as part of product subscription service, and 1 company provided a total annual estimate of training cost.
- The EAG has assumed that cost of AI is applied as a cost per scan to all diagnostic images (that is that the hospital cannot specify which images the AI is applied to). Therefore, even patients with known vertebral fracture or already receiving treatment for vertebral fracture will incur the cost per scan, but with no additional benefit.
- The EAG included a “generic AI” arm in the economic model using the AI cost per scan in 2022 from Curl et al. 2024 (AI cost per scan: $10 converted to £7.36).

For context, the EAG notes that the calculated costs per scan are between [redacted] across the technologies across image modalities, which is the equivalent of between [redacted] minutes of a radiologist specialist in musculoskeletal imaging per scan (assuming £147.91 per hour; (Jones et al., 2024) when inflated to 2024 costs). One Clinical Expert provided additional context and advised that they could read 60 DEXA examinations within an hour.

Table 17: Key cost parameters for the AI (no discounting applied)

Parameter	Generic AI	Annalise (X-ray)	BoneView (X-ray)	TechCare Spine (X-ray)	BriefCase-Triage (CT)	CINA-VCF Quantix (CT)	HealthVCF and HealthOST (Nanox AI) (CT)	IB Lab FLAMINGO (CT)
Product subscription (per year)	-	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Implementation (one-off)	-	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Integration (one-off)	-	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Training	-	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Maintenance (per year)	-	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Total per scan*: (no. of scans)	£7.36 as reported by (Curl et al., 2024); converted from $10 (2022 cost)	[redacted]	£1.00 (Notional cost)	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Key: *including implementation, integration, training and maintenance costs
Abbreviations: NR, not reported; PACS, Picture archiving and communication system, RIS, Radiology Information Systems.

6.2.4.3 Cost of confirmatory review by radiologist

The staff costs associated with the review and interpretation of the AI report are described in Table 18. For all cases where a vertebral fracture is flagged (either by AI or by SoC arm), the EAG assumed that the image will be re-reviewed by a radiologist specialising in musculoskeletal imaging, taking approximately 10 additional minutes (informed by Clinical Expert opinion; Appendix D2), with alternative durations (1 and 15 minutes) being considered in sensitivity analysis. Thus, all test positive cases incurred an additional £24.65 per scan.

Table 18: Cost of review and interpretation of AI report

Parameter	Unit Cost	Source	EAG comment
Radiologists (Consultant level, Cost per working hour)	£147.91	(Jones et al., 2024)	Hospital-based consultant – medical, including qualification: £143 inflated to 2024 costs.
Additional time needed review and interpretation AI outputs	10 minutes	Expert views (see Appendix D2)	Clinician's estimate ranges from 1 to 15 minutes. 10 minutes is used in the base case
Total cost of addition review (per scan)	**£24.65**	**-**	**-**

6.2.4.4 Cost of additional spine X-ray

In addition to the staff time associated with radiologist reviewing any test positive scans, the EAG assumed that 10% of all positive patients may require an additional spine X-ray (given that the initial image may have been a chest, abdominal or pelvic X-ray or CT) for the radiologist to check the presence of the fracture, Table 19.

Table 19: Cost of additional spine X-ray

Parameter	Unit Cost	Source	EAG comment
X-ray	£4.90	NHS reference costs 2023-2024	Non-Consultant Led – Currency code WF01D (Service code 811 – interventional radiology) £49; assuming that 10% will require an additional confirmatory spinal X-ray.

6.2.4.5 Total vertebral fracture management cost over 1 year

The total vertebral fracture management costs over 1 year as described in TA464 (updated in 2017) were £4,173, comprising 92.4% hospitalisation, 2.0% A&E, 1.1% GP, 3.5% referral, 0.4% prescribing, and 0.6% home care costs (assumed relevant to an NHS and PSS perspective). The EAG inflated this to £5,302 to reflect 2024 costs. The EAG considered that given the short time horizon of the economic model, a difference in the requirement of long-term care (whether home help or residential care) would not be expected between SoC and AI-assisted and therefore was omitted. The EAG has assumed that this cost will include referral for DEXA and FLS. To explore the impact of this uncertainty, the EAG increased management costs in sensitivity analysis.

6.2.4.6 True positive (fracture detected)

Each true positive incurred the cost of the second review (by radiologist specialising in musculoskeletal imaging) and a proportion required a confirmatory spine X-ray. An estimated 14.6% also receive management of their vertebral fracture (as estimated by Dalal et al. 2022). Therefore, each true positive case would incur a total of £5,331.55 with vertebral fracture management and £29.55 without vertebral fracture management.

Table 20: Additional cost for correct diagnosis (True Positive, TP)

Parameter	Unit Cost	Source	EAG comment
Additional review of scan	£24.65	See Table 18	10 minutes of radiologist specialising in musculoskeletal imaging.

Parameter	Unit Cost	Source	EAG comment
X-ray	£4.90	NHS reference costs 2023-2024	Non-Consultant Led – Currency code WF01D (Service code 811 – interventional radiology) £49; assuming that 10% will require an additional confirmatory spinal X-ray.
Management costs	£5,302	TA464 (2012, updated 2017), inflated to 2024	See section 6.2.4.5.
Total costs per true positive (with decision to manage)	**£5,331.55**	**As above**	**Combined additional review, proportion undergoing additional scan and management costs**
Total costs per true positive (with no management)	**£29.55**	**As above**	**Combined additional review, proportion undergoing additional scan (no management costs)**

6.2.4.7 False negatives (missed opportunity for vertebral fracture diagnosis)

The EAG assumed that if a fracture is not detected by initial reader or AI, then there is no second review by a radiologist and no management costs are incurred. Thus, no additional costs would be incurred for missed opportunities to identify a vertebral fracture.

6.2.4.8 True negative

Patients correctly diagnosed as not having a vertebral fracture (that is fracture not present and not detected by either initial reader or AI) were assumed to incur no additional cost.

6.2.4.9 False positive

The EAG assumed that all false positive cases (that is fracture identified by SoC or AI technology) would require a second review by a radiologist specialised in musculoskeletal imaging and therefore would be detected as a false positive. The EAG assumed that 10% of patients may require an additional spine X-ray (given that the initial image may have been a chest, abdominal or pelvic X-ray or CT; with this proportion being varied in sensitivity analysis), but that no further additional management costs would be incurred. Therefore, each false positive case would incur an additional £29.55Table 21.

Table 21: Additional cost for incorrect diagnosis (False Positive)

Parameter	Unit Cost	Source	EAG comment
Additional review	£24.65	See Table 18	10 minutes of a radiologist specialising in musculoskeletal imaging
X-ray	£4.90	NHS reference costs 2023-2024	Non-consultant Led – Currency code WF01D (Service code 811 – interventional radiology) £49; assuming that 10% will require an additional confirmatory spinal X-ray.
Total cost per false positive	**£29.55**	-	-

6.2.5 Measurement and valuation of health effects

Following the NICE reference case, a cost-utility was chosen to analyse cost effectiveness of opportunistic vertebral fracture detection using AI technologies. Due to the AI technologies being used in a general population referred for X-ray or CT scans which involve the spine but not specifically for vertebral fracture detection, and each of the AI technologies being used in different populations (due to different eligibility criteria) the EAG did not model absolute values of utility for a starting population, or absolute decrements associated with any events. Instead, the EAG modelled a utility gain for those where a vertebral fracture was present and opportunistically detected (true positive) and the patient was then referred for treatment. The utility gain was assumed to be 0 in all other patients (no change from baseline).

The utility gain applied was derived from the study by Svedbom et al. (2018), considered in both TA464 and TA791, which investigated the quality of life (QoL) outcomes associated with vertebral fractures (n=559), at different timepoints over an 18-month period using EQ-5D-3L (Svedbom et al., 2018). At enrolment patients with vertebral fracture had a utility of 0.17 (SD 0.43) which increased to 0.70 (SD 0.29) at 12 months. This would be the equivalent of a utility gain of 0.53 (difference between 12 months and baseline). The EAG acknowledges that this is likely to be a substantial overestimate, given that the population posed within the decision problem of this early value assessment are asymptomatic, and that not all patients diagnosed with fracture may undergo treatment within the 1-year time horizon. Therefore, the EAG assumed the utility gain was 50% smaller in the base case (0.265) with even smaller utility gains (0.1325, 0.053, 0.0265) considered within sensitivity analyses which may reflect the utility gain expected in an asymptomatic population.

6.2.6 Model validation

The economic model was built in *rdecision* package in R, which includes 1,300 self-tests which verify its computational methods. A detailed breakdown of the generic AI economic model base case results is provided in Appendix B1.

6.2.7 Presentation of results

The EAG explored a range of sensitivity and scenario analysis to determine the key drivers of the economic model:

- Diagnostic accuracy for standard care in review of X-rays: due to large uncertainty on the sensitivity and specificity of standard care the EAG explored a range of values.
- Diagnostic accuracy for standard of care in review of CTs: sensitivity 50% (Howlett et al., 2023), or 51% and specificity 100% (Chappell et al., 2024).
- Diagnostic accuracy for AI technologies processing X-rays:
    - Annalise.AI: [redacted]
    - BoneView: sensitivity: 63.2% (SD 5.3%) and specificity 96.7% (SD 0.8%) per patient when using lateral images only (Oppenheimer et al., 2024).
    - BriefCase: Only 1 study available there no change to sensitivity or specificity of AI technology.
- Diagnostic accuracy for AI technologies processing CTs:
    - CINA VCF: sensitivity: 92.3% (81.5% to 97.9%) and specificity 91.7% (80.0% to 97.7%) per patient (Guenoun et al., 2025).
    - HealthVCF: sensitivity: 65% (59.8% to 70.9%) and specificity: 92% (90.9% to 93.8%) (Kolanu et al., 2020).
    - IB Lab FLAMINGO: sensitivity 80.8% (76.2% to 85.1%) and specificity 94.5% (93.3% to 95.5%) (Nicolaes et al., 2023).

- Prevalence: decreased to 7.3% and increased to 52.0% using extremes reported in the clinical evidence.
- Failure rate:
    - Annalise.AI: decreased to 0% (0/1,559) (Talwar, 2023).
    - HealthVCF: increased to 6.0% (108/1804) and 9.4% (113/1200) based on other available published evidence (Kolanu et al., 2020; Page et al., 2023).
    - IB Lab FLAMINGO: increased to 2.9% (57/2,000) and 4.8% (143/2,797) based on other available published evidence (IB Lab, 2023; Nicolaes et al., 2023)
- Annual volume:
    - X-rays: 40,000, 88,000, 215,000 per site per year
    - CT scans: 3,700, 8,300, 29,000 per site per year.
- Time to read and interpret the AI report: decreased to 1 minute and increased to 15 minutes (spanning the range provided by Clinical Experts).
- Proportion of positive cases (TP, FP) requiring confirmatory spinal X-ray: varied to 5%, 20% and 50%.
- Management costs: Increased by 20% (would also cover increased proportion of DEXA scans) to £6,362.
- Utility gain for management of vertebral fracture: reduced from 0.265 in the base case to 0.1325 (50% reduction), 0.053 (80% reduction) and 0.0265 (90% reduction) because patients were identified for management via incidental findings (asymptomatic) and may have a smaller utility gain than those presenting with symptoms of a vertebral fracture.

In addition to the above, for the generic AI arm the EAG also varied the cost per scan to determine the threshold at which the AI technology would be cost-effective at the £20,000 willingness to pay. Due to considerable uncertainty surrounding some parameter values including on both point estimates and distributions of key parameters (prevalence, proportion requiring treatment, proportion requiring referral for subsequent spinal X-ray, utility gains), the EAG focused on how variation in parameter values would alter results rather than trying to characterise the joint uncertainty in costs and QALYs using probabilistic sensitivity analysis (PSA).

In addition to cost effectiveness results, the EAG included the following outcomes as model outputs for each paired comparison to provide additional insights: additional vertebral fracture detected with AI compared with SoC per 1,000 scans, extra reviews required with AI compared with SoC per 1000 scans, total missed opportunities with SoC per 1,000 scans, and total missed opportunities with AI per 1,000 scans.

6.3 Results from the economic modelling

6.3.1 Generic AI

The results from the generic AI (where AI costs were £7.36 per scan) are summarised in Table 22. In the base case the AI arm was £86.03 more expensive per patient (AI: £148.70, SoC: £62.65) and results in 0.003896 additional QALYs (AI: 0.006803, SoC: 0.002907) giving an ICER of £22,085, which is close to the willingness to pay threshold of £20,000 even when longer-term benefits are not considered. The EAG notes that for each 1000 people scanned, 100.7 additional people with vertebral fracture would be detected with the AI, 14.7 of whom would be managed, with fewer missed opportunities for management (AI: 118.2, SoC: 221.9). An additional 24.76 scans would need to be reviewed by a radiologist; however, the EAG notes that the sensitivity is lower and specificity higher than all named technologies and this limits the usefulness of these results to decision-making.

The economic model was sensitive to changes in sensitivity and specificity. With the same diagnostic accuracy as standard care (25.3% sensitivity, 89.1% specificity) the AI technology was dominated by standard of care (consequence of increased cost with AI, and slightly lower QALY due to a small proportion where the AI is unable to process the image and therefore unable to have diagnosis confirmed or referral for vertebral fracture management). For a 5% increase in sensitivity, the AI technology had an ICER of £33,816. The AI technology was dominated when the sensitivity of standard care was increased to 60% (SoC sensitivity in the base case 25.3%;

sensitivity and specificity of AI the same as the base case). The economic model was also sensitive to change in utility gain associated with vertebral fracture management. If this was reduced to 0.1325 over 12 months (from 0.265 in the base case) it resulted in an ICER of £44,169, which increased to £110,423 if the utility gain was reduced to 0.053 over 12 months (90% reduction), or £220,846 when the utility gain was reduced to 0.0265 over 12 months (reduced by 95%), the latter perhaps a more realistic scenario when applied to an asymptomatic population.

The economic model appeared relatively insensitive to changes in failure rate, prevalence, time to review AI output, and the proportion of images requiring additional X-ray. As well as changes affecting costs. A combined (extreme case) scenario (AI failure rate 10%, prevalence 52%, management costs £6,362, time to report 15 minutes, utility gain for management 0.0265, proportion requiring X-ray increased to 20%) resulted in an ICER of £260,778. However, the EAG would highlight that the modelling was restricted to a short 1-year time horizon and that sensitivity analysis was conducted to demonstrate impact of changes. Further evidence generation is required to obtain values of parameters with greater certainty.

Table 22: Deterministic results – generic AI base case (X-ray images)

Sensitivity analysis	SoC, Total costs £	AI, Total costs £	Difference, £	SoC, QALY	AI, QALY	Difference, QALY	ICER	Additional VFs detected with AI compared with SoC per 1,000 scans	Extra reviews required with AI compared with SoC per 1000 scans	Missed opportunities with SoC per 1,000 scans	Missed opportunities with AI per 1,000 scans
Base case	62.65	148.70	86.03	0.002907	0.006803	0.003896	22,085	100.7	24.76	221.9	118.2
SoC: sensitivity: 40%	97.74	148.70	50.95	0.004596	0.006803	0.002206	23,090	57.03	-18.9	178.2	118.2
SoC: sensitivity: 50%	121.60	148.70	27.08	0.005745	0.006803	0.001057	25,610	27.33	-48.6	148.5	118.2
SoC: sensitivity: 60%	145.50	148.70	3.21	0.006895	0.006803	-0.000092	Dominated	-2.37	-78.3	118.8	118.2
SoC: sensitivity 50%, specificity 100%	119.30	148.70	29.34	0.005745	0.006803	0.001057	27,751	27.33	28.03	148.5	118.2
AI: sensitivity: 100%, specificity: 95.4%	62.65	244.60	182.00	0.002907	0.01138	0.008469	21,485	218.9	174.3	221.9	0
AI: sensitivity: 60%, specificity: 80%	62.65	153.30	90.60	0.002907	0.006826	0.003918	23,122	101.3	163.8	221.9	117.6
AI: sensitivity: 25.3%, specificity: 89.1%	62.65	69.38	6.73	0.002907	0.002878	-0.0000297	Dominated	-0.7514	-1.518	221.9	219.6
AI: sensitivity: 30.3%, specificity: 89.1%	62.65	81.67	19.02	0.002907	0.00347	0.0005625	33,816	14.54	13.77	221.9	204.4
AI: sensitivity: 30.5%, specificity: 94.1%	62.65	80.64	17.99	0.002907	0.00347	0.0005625	31,988	14.54	-21.03	221.9	204.4
Prevalence: 7.3%	17.83	42.12	24.29	0.0007146	0.001672	0.0009575	25,368	24.75	-75.38	54.53	29.05
Prevalence: 52.0%	107.30	254.80	147.50	0.00509	0.01191	0.006821	21,626	176.3	124.4	388.4	206.9
Failure rate (AI): 0%	62.65	150.10	87.46	0.002907	0.006872	0.003964	22,062	102.5	26.54	221.9	119.4
Failure rate (AI): 2%	62.65	147.30	84.61	0.002907	0.006734	0.003827	22,108	98.91	22.97	221.9	117
Failure rate (AI): 10%	62.65	135.80	73.19	0.002907	0.006184	0.003277	22,332	84.7	8.71	221.9	107.5
Cost per scan (AI): £20	62.65	161.30	98.67	0.002907	0.006803	0.003896	25,329	100.7	24.76	221.9	118.2
Cost per scan (AI): £75	62.65	216.30	153.70	0.002907	0.006803	0.003896	39,448	100.7	24.76	221.9	118.2
Cost per scan (AI): £100	62.65	241.30	178.70	0.002907	0.006803	0.003896	45,865	100.7	24.76	221.9	118.2
Management costs: £6362	74.28	175.90	101.60	0.002907	0.006803	0.003896	26,086	100.7	24.76	221.9	118.2
Management costs: £7953	91.73	216.70	125.00	0.002907	0.006803	0.003896	32,088	100.7	24.76	221.9	118.2
Time to review AI report output: 1 minute	59.28	144.80	85.48	0.002907	0.006803	0.003896	21,944	100.7	24.76	221.9	118.2
Time to review AI report output: 15 minutes	64.52	150.90	86.34	0.002907	0.006803	0.003896	22,163	100.7	24.76	221.9	118.2
Proportion requiring additional spinal X-ray: 5%	62.28	148.30	85.97	0.002907	0.006803	0.003896	22,069	100.7	24.76	221.9	118.2
Proportion requiring additional spinal X-ray: 20%	63.39	149.60	86.16	0.002907	0.006803	0.003896	22,116	100.7	24.76	221.9	118.2
Proportion requiring additional spinal X-ray: 50%	65.63	152.10	86.52	0.002907	0.006803	0.003896	22,209	100.7	24.76	221.9	118.2
Proportion requiring additional spinal X-ray: 100%	69.34	156.50	87.13	0.002907	0.006803	0.003896	22,365	100.7	24.76	221.9	118.2
Proportion referred for management: 25%	104.10	245.60	141.60	0.004978	0.01165	0.006671	21,221	100.7	24.76	221.9	118.2
Proportion referred for management: 50%	203.70	478.70	275.00	0.009956	0.0233	0.01334	20,614	100.7	24.76	221.9	118.2
Utility gain for management: 0.1325	62.65	148.70	86.03	0.001454	0.003401	0.001948	44,169	100.7	24.76	221.9	118.2
Utility gain for management: 0.053	62.65	148.70	86.03	0.0005814	0.001361	0.0007791	110,423	100.7	24.76	221.9	118.2
Utility gain for management: 0.0265	62.65	148.70	86.03	0.0002907	0.0006803	0.0003896	220,846	100.7	24.76	221.9	118.2
Combined scenario (AI failure rate 10%, prevalence 52%, management costs £6362, time to report 15 minutes, utility gain 0.0265, proportion requiring X-ray 20%)	130.80	280.40	149.60	0.000509	0.001083	0.0005738	260,778	148.3	96.42	388.4	188.1

Abbreviations: AI, Artificial Intelligence; ICER, Incremental cost-effectiveness ratio, QALY; Quality adjusted life year; SoC, Standard of care; VF, Vertebral fracture.

6.3.2 Annalise.AI

The results for of Annalise.AI are summarised in Table 23. In the base case analysis, the use of Annalise.AI arm was £[redacted] more expensive than standard of care (AI: £[redacted], SoC: £[redacted]) and resulted in [redacted] additional QALYs (AI: [redacted], SoC:[redacted]) giving an ICER of £[redacted] close to the willingness to pay threshold of £20,000. The EAG notes that for each 1000 people scanned, [redacted] additional people with vertebral fracture would be detected with the AI, and fewer missed opportunities (AI: [redacted], SoC: [redacted]), but also that [redacted] additional scans would need to be reviewed by a radiologist. Univariate changes in utility gain associated with vertebral fracture management had the largest impact on the sensitivity analysis.

Table 23: Deterministic results - Annalise.AI (X-ray images)

Sensitivity analysis	SoC, Total costs, £	Annalise.AI, Total costs, £	Difference, £	SoC, QALY	Annalise.AI, QALY	Difference, QALY	ICER	Additional VFs detected with AI compared with SoC per 1,000 scans	Extra reviews required with AI compared with SoC per 1000 scans	Missed opportunities with SoC per 1,000 scans	Missed opportunities with AI per 1,000 scans
Base case	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 15.3%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 40%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 60%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity 50%, specificity 100%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annalise.AI: sensitivity: [redacted], specificity: [redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Prevalence: 7.3%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Prevalence: 52.0%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Failure rate (Annalise.AI): 0%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume X-ray: 40,000	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume X-ray: 88,000	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume X-ray: 215,000	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Management costs: £6,362	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Time to review AI report output: 1 minute	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Time to review AI report output: 15 minutes	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 5%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 20%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion referred for management: 25%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion referred for management: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Utility gain for management: 0.1325	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Utility gain for management: 0.053	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Abbreviations: AI, Artificial intelligence; ICER, Incremental cost-effectiveness ratio, QALY; Quality adjusted life year; SoC, Standard of care; VF, Vertebral fracture.

6.3.3 BoneView

The results for of BoneView.AI are summarised in Table 24. In the base case analysis, the BoneView.AI arm was £108.40 more expensive (AI: £171.00, SoC: £62.65) and results in 0.005221 additional QALYs (AI: 0.008128, SoC: 0.002907) giving an ICER of £20,755. The EAG notes that for each 1000 people scanned, 134.9 additional people with vertebral fractures would be detected with AI, with fewer missed opportunities (AI: 80.09, SoC: 221.9), but also that 98.15 additional scans would need to be reviewed by a radiologist. Univariate changes in utility gain associated with vertebral fracture management had the largest impact on the sensitivity analysis. The EAG notes that changes to volume made no impact in this analysis, as was modelled as a fixed cost per scan.

The EAG notes that the sensitivity and specificity applied were derived from lateral and anterior-posterior projections combined (which may not be generalisable to the population of the decision problem) and were reported per-vertebrae (not per patient). Therefore, these results should be interpreted with caution.

Table 24: Deterministic results - BoneView (X-ray images)

Sensitivity analysis	SoC, Total costs £	BoneView, Total costs £	Difference, £	SoC, QALY	BoneView, QALY	Difference, QALY	ICER	Additional VFs detected with AI compared with SoC per 1,000 scans	Extra reviews required with AI compared with SoC per 1000 scans	Missed opportunities with SoC per 1,000 scans	Missed opportunities with AI per 1,000 scans
Base case	62.65	171.00	108.40	0.002907	0.008128	0.005221	20,755	134.9	98.15	221.9	80.09
SoC: sensitivity: 15.3%	38.78	171.00	132.20	0.001758	0.008128	0.00637	20,758	164.6	127.9	251.6	80.09
SoC: sensitivity: 40%	97.74	171.00	73.27	0.004596	0.008128	0.003532	20,747	91.28	54.49	178.2	80.09
SoC: sensitivity: 50%	121.60	171.00	49.40	0.005745	0.008128	0.002383	20,735	61.58	24.79	148.5	80.09
SoC: sensitivity: 60%	145.50	171.00	25.53	0.006895	0.008128	0.001234	20,701	31.88	-4.908	118.8	80.09
SoC: sensitivity 50%, specificity 100%	119.30	171.00	51.67	0.005745	0.008128	0.002383	21,685	61.58	101.4	148.5	80.09
BoneView: sensitivity 63.2%, specificity 96.7%	62.65	149.00	86.40	0.002907	0.007095	0.004188	20,629	108.2	54.28	221.9	106.8
Prevalence: 7.3%	17.83	44.05	26.22	0.0007146	0.001998	0.001283	20,433	33.17	-15.35	54.53	19.68
Prevalence: 52.0%	107.30	297.40	190.10	0.00509	0.01423	0.009141	20,800	236.3	211.1	388.4	140.2
Annual volume X-ray: 40,000	62.65	171.00	108.40	0.002907	0.008128	0.005221	20,755	134.9	98.15	221.9	80.09
Annual volume X-ray: 88,000	62.65	171.00	108.40	0.002907	0.008128	0.005221	20,755	134.9	98.15	221.9	80.09
Annual volume X-ray: 215,000	62.65	171.00	108.40	0.002907	0.008128	0.005221	20,755	134.9	98.15	221.9	80.09
Management costs: £6362	74.28	203.50	129.20	0.002907	0.008128	0.005221	24,756	134.9	98.15	221.9	80.09
Time to review AI report output: 1 minute	59.28	165.50	106.20	0.002907	0.008128	0.005221	20,338	134.9	98.15	221.9	80.09
Time to review AI report output: 15 minutes	64.52	174.10	109.60	0.002907	0.008128	0.005221	20,986	134.9	98.15	221.9	80.09
Proportion requiring additional spinal X-ray: 5%	62.28	170.40	108.10	0.002907	0.008128	0.005221	20,709	134.9	98.15	221.9	80.09
Proportion requiring additional spinal X-ray: 20%	63.39	172.20	108.80	0.002907	0.008128	0.005221	20,847	134.9	98.15	221.9	80.09
Proportion requiring additional spinal X-ray: 50%	65.63	175.90	110.30	0.002907	0.008128	0.005221	21,123	134.9	98.15	221.9	80.09
Proportion referred for management: 25%	104.10	286.80	182.80	0.004978	0.01392	0.00894	20,444	134.9	98.15	221.9	80.09
Proportion referred for management: 50%	203.70	565.30	361.60	0.009956	0.02784	0.01788	20,226	134.9	98.15	221.9	80.09
Utility gain for management: 0.1325	62.65	171.00	108.40	0.001454	0.004064	0.00261	41,509	134.9	98.15	221.9	80.09
Utility gain for management: 0.053	62.65	171.00	108.40	0.0005814	0.001626	0.001044	103,773	134.9	98.15	221.9	80.09

Abbreviations: AI, Artificial intelligence; ICER, Incremental cost-effectiveness ratio; QALY; Quality adjusted life year; SoC, Standard of care; VF, Vertebral fracture.

6.3.3 BriefCase-Triage

The results for BriefCase-Triage.AI are summarised in Table 25. In the base case analysis, the BriefCase-Triage.AI arm was £[redacted] more expensive (AI: £[redacted], SoC: £[redacted]) and results in [redacted] additional QALYs (AI: [redacted], SoC: [redacted]) giving an ICER of £[redacted]. The EAG notes that for each 1000 people scanned, [redacted] additional people with vertebral fracture would be detected with AI, with fewer missed opportunities (AI: [redacted], SoC: [redacted]) but also that [redacted] additional scans would need to be reviewed by a radiologist. Univariate changes in utility gain associated with vertebral fracture management had the largest impact on the sensitivity analysis.

Table 25: Economic analysis of BriefCase-Triage (CT images)

Sensitivity analysis	SoC, Total costs £	BriefCase, Total costs £	Difference, £	SoC, QALY	BriefCase, QALY	Difference, QALY	ICER	Additional VFs detected with AI compared with SoC per 1,000 scans	Extra reviews required with AI compared with SoC per 1000 scans	Missed opportunities with SoC per 1,000 scans	Missed opportunities with AI per 1,000 scans
Base case	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 15.3%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 40%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 60%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity 50%, specificity 100%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Prevalence: 7.3%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Prevalence: 52.0%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume CT: 3,700	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume CT: 8,300	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume CT: 29,000	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Management costs: £6362	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Time to review AI report output: 1 minute	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Time to review AI report output: 15 minutes	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 5%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 20%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion referred for management: 25%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion referred for management: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Utility gain for management: 0.1325	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Utility gain for management: 0.053	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Abbreviations: AI, Artificial intelligence; ICER, Incremental cost-effectiveness ratio; QALY; Quality adjusted life year; SoC, Standard of care; VF, Vertebral fracture.

6.3.4 CINA-VCF

The results for CINA-VCF.AI are summarised in Table 26. In the base case analysis, the CINA-VCF.AI arm was £[redacted] more expensive (AI: £[redacted], SoC: £[redacted] and results in [redacted] additional QALYs (AI: [redacted], SoC: [redacted]) giving an ICER of £[redacted]. The EAG notes that for each 1000 people scanned, [redacted] additional people with vertebral fracture would be detected with the AI, with fewer missed opportunities (AI: [redacted], SoC: [redacted]) but also that [redacted] additional scans would need to be reviewed by a radiologist. Univariate changes in utility gain associated with vertebral fracture management had the largest impact on the sensitivity analysis.

Table 26: Economic analysis of CINA-VCF (CT images)

Sensitivity analysis	SoC, Total costs £	CINA-VCF, Total costs £	Difference, £	SoC, QALY	CINA-VCF, QALY	Difference, QALY	ICER	Additional VFs detected with AI compared with SoC per 1,000 scans	Extra reviews required with AI compared with SoC per 1000 scans	Missed opportunities with SoC per 1,000 scans	Missed opportunities with AI per 1,000 scans
Base case	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 15.3%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 40%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 60%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity 50%, specificity 100%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
CINA VCF: sensitivity: 92.3%, specificity: 91.7%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Prevalence: 7.3%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Prevalence: 52.0%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume CT: 3,700	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume CT: 8,300	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume CT: 29,000	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Management costs: £6362	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Time to review AI report output: 1 minute	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Time to review AI report output: 15 minutes	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 5%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 20%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion referred for management: 25%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion referred for management: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Utility gain for management: 0.1325	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Utility gain for management: 0.053	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Abbreviations: AI, Artificial intelligence; ICER, Incremental cost-effectiveness ratio; QALY; Quality adjusted life year; SoC, Standard of care; VF, Vertebral fracture.

6.3.5 HealthVCF

The results for HealthVCF are summarised in Table 27. In the base case analysis, the HealthVCF.AI arm was £[redacted] more expensive (AI: £[redacted], SoC: £[redacted]) and results in [redacted] additional QALYs (AI:[redacted], SoC:[redacted]) giving an ICER of £[redacted]. The EAG notes that for each 1000 people scanned, [redacted] additional people with vertebral fracture would be detected with the AI, with fewer missed opportunities (AI: [redacted], SoC: [redacted]) but also that [redacted] additional scans would need to be reviewed by a radiologist. Univariate changes in utility gain associated with vertebral fracture management had the largest impact on the sensitivity analysis.

Table 27: Economic analysis of HealthVCF (CT images)

Sensitivity analysis	SoC, Total costs £	HealthVCF, Total costs £	Difference, £	SoC, QALY	HealthVCF, QALY	Difference, QALY	ICER	Additional VFs detected with AI compared with SoC per 1,000 scans	Extra reviews required with AI compared with SoC per 1000 scans	Missed opportunities with SoC per 1,000 scans	Missed opportunities with AI per 1,000 scans
Base case	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 15.3%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 40%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 60%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity 50%, specificity 100%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
HealthVCF: sensitivity: 65%, specificity: 92%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Prevalence: 7.3%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Prevalence: 52.0%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Failure rate (HealthVCF): 6.0%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Failure rate (HealthVCF): 9.4%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume CT: 3,700	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume CT: 8,300	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume CT: 29,000	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Management costs: £6362	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Time to review AI report output: 1 minute	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Time to review AI report output: 15 minutes	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 5%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 20%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion referred for management: 25%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion referred for management: 50%,	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Utility gain for management: 0.1325	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Utility gain for management: 0.053	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Abbreviations: AI, Artificial intelligence; ICER, Incremental cost-effectiveness ratio; QALY; Quality adjusted life year; SoC, Standard of care; VF, Vertebral fracture.

6.3.6 IB Lab FLAMINGO

The results for IB Lab FLAMINGO are summarised in Table 28. In the base case analysis, the IB Lab FLAMINGO.AI arm was £[redacted] more expensive (AI: £[redacted], SoC: £[redacted]) and results in [redacted] additional QALYs (AI: [redacted], SoC: [redacted]) giving an ICER of £[redacted]. The EAG notes that for each 1000 people scanned, [redacted] additional people with vertebral fracture would be detected with the AI, with fewer missed opportunities (AI: [redacted], SoC: [redacted]) but also that [redacted] additional scans would need to be reviewed by a radiologist. Univariate changes in utility gain associated with vertebral fracture management had the largest impact on the sensitivity analysis.

Table 28: Economic analysis of IB Lab FLAMINGO (CT images)

Sensitivity analysis	SoC, Total costs £	IB Lab FLAMINGO, Total costs £	Difference, £	SoC, QALY	IB Lab Flamingo, QALY	Difference, QALY	ICER	Additional VFs detected with AI compared with SoC per 1,000 scans	Extra reviews required with AI compared with SoC per 1000 scans	Missed opportunities with SoC per 1,000 scans	Missed opportunities with AI per 1,000 scans
Base case	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 15.3%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 40%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity: 60%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
SoC: sensitivity 50%, specificity 100%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
IB Lab FLAMINGO: sensitivity: 80.8%, specificity: 94.5%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Prevalence: 7.3%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Prevalence: 52.0%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Failure rate (IB Lab FLAMINGO): 2.9%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Failure rate (IB Lab FLAMINGO): 4.8%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume CT: 3,700	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume CT: 8,300	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Annual volume CT: 29,000	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Management costs: £6362	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Time to review AI report output: 1 minute	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Time to review AI report output: 15 minutes	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 5%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 20%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion requiring additional spinal X-ray: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion referred for management: 25%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Proportion referred for management: 50%	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Utility gain for management: 0.1325	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Utility gain for management: 0.053	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Abbreviations: AI, Artificial intelligence; ICER, Incremental cost-effectiveness ratio; QALY; Quality adjusted life year; SoC, Standard of care; VF, Vertebral fracture.

6.4 Summary and interpretation of the economic evidence

The early de novo economic model developed for this EVA suggests that AI technologies have the potential to represent value for money in the NHS at a £20,000 per QALY willingness to pay threshold. However, additional evidence is required due to a general lack of information regarding prevalence, the proportion requiring treatment and diagnostic accuracy of opportunistic detection of vertebral fracture in standard care in the UK NHS. In addition to this, the EAG notes that the patient population referred for diagnostic imaging for other (non-vertebral fracture detection) clinical reasons may differ by image modalities, therefore diagnostic accuracy by image modality, anatomical location and potentially by indication for scan, may impact estimates of diagnostic accuracy. Some manufacturers have stated in the instructions for use that the software can be configured to clinical specification. These configuration settings may also impact results.

It is also unclear how introduction of AI technologies into the AI workflow works in practice. For example, it is unclear how patients with a prior diagnosis of vertebral fracture who are already undergoing treatment for fracture would be identified (this information is not routinely available in DICOM tags) such that processing the image through AI would not be required. This has a consequence in the economic case where the cost per scan would be applied but there could be no additional benefits as treatments are already being given. However, as the prevalence of asymptomatic vertebral fractures overall and for those currently on and not currently on treatment is unknown, then the impact on overall cost (and QALYs) is unknown.

The economic modelling was restricted to a one-year time horizon focusing on the diagnostic pathway, but the longer-term benefits remain unknown. One study highlighted that more vertebral fractures were described in the initial radiology reports over time when AI was implemented, which may demonstrate that reporting radiographers may be changing their behaviour as they are aware of observations being taken. However, increased reporting of vertebral fractures could also be a consequence of RCR guidelines which have also made specific recommendations in the incidental reporting of vertebral fractures. The impact of false negatives is minimal in a one-year time horizon, as Clinical Experts advised that many patients may not be re-present or be re-referred within 12 months, however reductions in quality of life and additional healthcare costs are plausible in those patients. A further area of uncertainty is the utility gain associated with the identification asymptomatic vertebral fracture. In the base case analysis very large gains in QALY over 12 months were assumed (0.265 QALYs over 12 months), reducing this by 50% (0.1325 QALYs over 12 months) increased the ICER above £40,000, and reducing by 95% (0.0265 QALYs over 12 months, which may be more realistic in an asymptomatic population) increased the ICER above £200,000. Ideally prospective data collection of utilities with a matched cohort not receiving treatment for opportunistically detected vertebral fracture is needed.

Furthermore, the workflow (considering eligibility criteria of each technology), associated vertebral fracture prevalence, diagnostic accuracy (which may vary across the diverse population in scope, and across imaging modalities) and downstream impact of vertebral fracture detection on other NHS services (such as the number of patients referred for DEXA scans and FLS and the time taken to attend these), when implementing each of the AI technologies in a UK NHS setting are all currently uncertain. Additional data collection should aim to reduce these uncertainties.

7. Integration into the NHS

Despite the call for early vertebral fracture detection, several systemic barriers exist:

- People diagnosed with osteoporosis may be referred to a FLS but there is limited access; only 51% of the NHS have access to an FLS (Royal Osteoporosis Society, 2021).
- Access to diagnostic services varies by region. The All-Party Parliamentary Group (APPG, 2021) on Osteoporosis and Bone Health reported a shortage of DEXA bone density scanners. In the latest NHS England figures (September 2024) 56,366 patients were waiting for a DEXA scan: 18.5% had been waiting more than the targeted six weeks (NHS England). This is an improvement on 33.6% in September 2023. The EAG notes that there were only 146 DEXA scanners available in England, therefore the downstream impact of increased vertebral fracture detection when implementing AI technologies needs careful consideration to ensure that there is not a negative impact on waiting times and time to treatment.
- All AI technologies require access to the internet, and the information governance may vary across NHS organisations. At the time of writing this report only 3 of the technologies included in this assessment have completed DTAC.
- Integration of AI technologies within the NHS poses a number of challenges which has been considered in the narrative review by Shelmerdine et al. (2024), which draws upon a case study from South West London.

In November 2024, the Royal College of Radiologists (RCR) developed guidance on AI deployment fundamentals for medical imaging, which recommended a number of key questions to consider when identifying available AI tools:

- Has the AI tool in question already been deployed in the same way with comparable healthcare organisations? A consultation with other trusts already using the tool could be considered.
- Have you considered the recommendations made by the manufacturer?
- What is the status of integration with the PACS and RIS?
- How was the algorithm trained – has it been trained on representative patients and pathologies?
- Is the performance acceptable – does it do the required task well?
- Under what circumstances should the AI tool not be applied and what are its limitations?
- Are the results generalisable – are the same results likely in your proposed population as those that have been tested?
- How will you avoid AI outputs adding bias to the clinical decision-making process when comparing AI to current diagnostic tools?
- What pre-deployment testing is likely to be required?
- What are the likely downstream effects of implementation? Consider the expected changes in existing pathways or services, whether services will be able to cope with those changes and the impact this may have on implementation.

8. Evidence gap analysis

8.1 Ongoing studies

The EAG pragmatically searched clinicaltrials.gov (on 21 March 2025) for the technology names listed in the Final Scope: no ongoing studies relevant to the decision problem were identified.

Several companies provided information regarding ongoing studies:

- Aidoc Medical have provided 2 studies, both are retrospective single centre studies. The locations and number of participants for these studies are not reported.
- Annalise.AI have provided [redacted].
- Avicenna have provided 3 studies: 1 RCT conducted in a single centre in France, 1 RCT multicentre (n=4 hospitals in France), and 1 RCT in Austria. The latter explicitly being conducted in patients aged 50 years and over.
- Gleamer did not complete a request for information for this topic. The EAG was supplied with the request for information documentation from HTE20 *Artificial intelligence software to help detect fractures in the emergency department (provisional title)* which stated that Gleamer has 30 ongoing studies of which there were 5 studies of interest to that topic.
- IB Lab have [redacted]
- Milvue provided no ongoing studies in the request for information.

- Nanox AI have provided one ongoing study (ADOPT; **A**I-enabled **D**etection of **O**steo**P**orosis for **T**reatment) that is a UK based multicentre (3 university hospitals and 1 teaching hospital). The aim is to compare outcomes from scans taken in 2017 with those taken in 2022 across a range of metrics including the AI-FLS pathway impact on patient care, and the clinical and cost effectiveness of the AI-FLS pathway. This study is not registered on a trial registry. One poster from ADOPT was already included in the evidence by the EAG (Chappell et al., 2024).

8.2 Evidence gap analysis

The EAG has summarised the evidence gaps across the technologies included in this early value assessment against the outcomes listed in the Final Scope, Table 29. Due to the lack of detail regarding methodology, it is unclear whether the ongoing studies (as summarised by the companies) will address these evidence gaps.

Table 29: Evidence gap analysis by AI technology (Key: GREEN evidence available, AMBER partial evidence available, RED no evidence available)

Outcomes	Annalise Enterprise (CXR) (Annalise.AI)	Annalise Container (CXR) (Annalise.AI)	BoneView (Gleamer)	BriefCase-Triage (Aidoc Medical)	CINA-VCF Quantix (Avicenna.AI)	HealthVCF (Nanox AI) (previously Zebra medical)	HealthOST (Nanox AI) (previously Zebra medical)	IB Lab FLAMINGO (Powered by UCB's, BoneBot AI model)	TechCare Spine (Milvue)
Diagnostic accuracy in detection of vertebral fracture	GREEN	RED	AMBER (per vertebrae only)	GREEN	GREEN	GREEN	RED	GREEN	RED
Accuracy by HCP profession	RED	RED	RED	RED	RED	RED	RED	RED	RED
Failure rate or rate of inconclusive AI reports	GREEN	RED	AMBER	GREEN	RED	GREEN	RED	GREEN	RED
Number of missed fractures	GREEN	RED	RED	RED	GREEN	GREEN	RED	AMBER (1 CiC study only)	RED
Rate of missed fracture-related further injury	RED	RED	RED	RED	RED	RED	RED	RED	RED
Proportion of people that need further imaging	RED	RED	RED	RED	RED	RED	RED	RED	RED
Intervention related adverse events	GREEN (no events identified)	GREEN (no events identified)	GREEN (no events identified)	GREEN (no events identified)	GREEN (no events identified)	GREEN (no events identified)	GREEN (no events identified)	GREEN (no events identified)	GREEN (no events identified)
HCP user acceptability of AI tools	GREEN	RED	RED	RED	RED	RED	RED	RED	RED
Changes to clinical management	RED	RED	RED	RED	RED	GREEN	RED	RED	RED
Health-related QoL	RED	RED	RED	RED	RED	RED	RED	RED	RED
Cost of the AI software	GREEN	GREEN	RED	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN
Staff costs	RED	RED	RED	RED	RED	RED	RED	RED	RED
Training and implementation	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN
Other downstream costs for diagnosis or treatment	RED	RED	RED	RED	RED	RED	RED	RED	RED
Time to produce a radiography report	RED	RED	RED	RED	RED	AMBER (time to analyse results from software)	RED	AMBER (time to run software)	RED
Time to diagnosis or time to definitive radiology report	RED	RED	RED	RED	RED	RED	RED	RED	RED
Time to further referral or treatment	RED	RED	RED	RED	RED	RED	RED	RED	RED
Number of treatments and extent of treatments	RED	RED	RED	RED	RED	RED	RED	RED	RED
Number of hospital appointment/visits	RED	RED	RED	RED	RED	RED	RED	RED	RED
Number of hospital admissions	RED	RED	RED	RED	RED	RED	RED	RED	RED
Type of healthcare professional interpreting the radiograph	RED	RED	RED	RED	RED	RED	RED	RED	RED

Abbreviations: AI, Artificial intelligence; HCP, Healthcare professional; QoL, quality of life.

Comparing the evidence gaps and referring to the decision problem, the EAG identified a number of key evidence gaps.

Population gaps:

- Indication for diagnostic imaging or confirmation that imaging was for reasons other than vertebral fracture detection was not reported in all cases.
- There is a lack of information on the patient demographics used in all studies. There is minimal evidence on the subgroups identified in the Scope ([redacted], none explicitly in people with osteoporosis or people with osteogenesis imperfecta which were subgroups listed in the Final Scope).
- One Clinical Expert noted that long term medication can impact bone density (such as steroid use); however long-term medication use was not routinely reported across the evidence.

Intervention gaps:

- Published evidence is not available for all technologies (no evidence for Annalise Container CXR (Annalise.AI), HealthOST (Nanox AI) or TechSpine (Milvue)). Some technologies have only a single study (BoneView by Gleamer, unnamed technology by Aidoc Medical).
- There is a general lack of transparent reporting of the software name, version number, and configuration settings which may influence results.

Comparator gaps:

- Comparator varied across studies; it is unclear how this aligns with standard practice in the NHS.
- The current standard of care is incidental detection by the initial reader (typically within 24 hours) and 11 studies reported AI detection compared with this; however, none were conducted in a UK setting.
- The sensitivity and specificity of original reporting radiographers (and other healthcare professionals) in incidental detection of vertebral fracture using X-ray images (for purposes other than vertebral fracture detection) remains unknown.

Outcome gaps

- No evidence was identified for any AI technology across a number of outcomes as defined in the Scope (accuracy by healthcare professional, rate of missed fracture-related further injury, proportion of people that need further imaging, health related quality of life, staff costs to implement the technology, other downstream costs for diagnosis or treatment, intervention related across a number of outcomes, time to diagnosis or time to definitive radiology report, time to further referral or treatment, number of treatments and extent of treatments, number of hospital appointments, visits or admissions, type of healthcare professional interpreting the radiograph).
- Prospective studies assessing diagnostic accuracy of standard of care (readers contributing to original radiology report) in incidental detection of vertebral fracture are at risk of bias related to changes in behaviour due to involvement in AI evaluation. Some studies reported review by a reference standard only for images which were flagged by the AI technology as potentially including a fracture, which may lead to biases in reported sensitivity and specificity values.
- Diagnostic accuracy (sensitivity and specificity) per-patient missing for BoneView.
- Definition of vertebral fragility or compression fracture varied across studies; however, the definition should closely match the population that the AI was trained and validated on.
- Long-term impact of incidental detection of vertebral fractures unknown. Acting on incidental findings does not always yield the expected benefits and can lead to net harm (as considered by (Davenport, 2023) in cancer detection).

Other considerations:

- Lack of evidence from the UK (2 abstract and 1 poster).
- Lack of evidence regarding clinic workflow (burden of false positives requiring review by specialist, influence on referrals and waiting times for further imaging including DEXA scans).
- No evidence to suggest that diagnosis (or earlier diagnosis of vertebral fracture) leads to faster diagnosis or treatment.

- Implementation and adoption across the NHS could be improved by more transparent reporting of the technology (such as version used) including detail of the datasets used during training and validation of the technology. May need to consider the implementation of the time and resources to train and validate each new version of the AI technology released.
- Consideration of patient consent regarding the use of AI applied to diagnostic images opportunistically out with their direct care or where clinically indicated, including transparently reporting how their data will be used and by who; this is regulated by the Information Commissioners Office.
- There is no direct head-to-head comparison of AI-technologies. Difference in patient suitability between technologies limit scope for this (noting decision problem states standard of care without AI as the relevant comparator), but do not prevent such comparisons. Additional work may include which optimal mix of technologies (and decision rules of how they are used) would be most appropriate for a given hospital throughput of diagnostic imaging.

8.3 Key areas for evidence generation

Considering the quality and quantity of evidence identified, the EAG considered 8 specific evidence generation recommendations, Table 30.

Across all recommendations including training and validation datasets there should be consistent detail on:

- patient demographics (including age, sex, ethnicity and medication use),
- detail on the technology (software name, version, and configuration settings),
- image modality (including anatomical location, projection when considering X-rays, manufacturer of CT or X-ray machine, and for CT slice thickness, use of contrast agents, kilovoltage peak which may confound results).

Studies should be done in the NHS setting with images captured from a wide range of scanners, settings and non-vertebral fracture clinical indications. Quality assessment of future evidence should consider the QUADAS-AI tool, and CONSORT-AI reporting guideline considered for clinical trials using an AI technology.

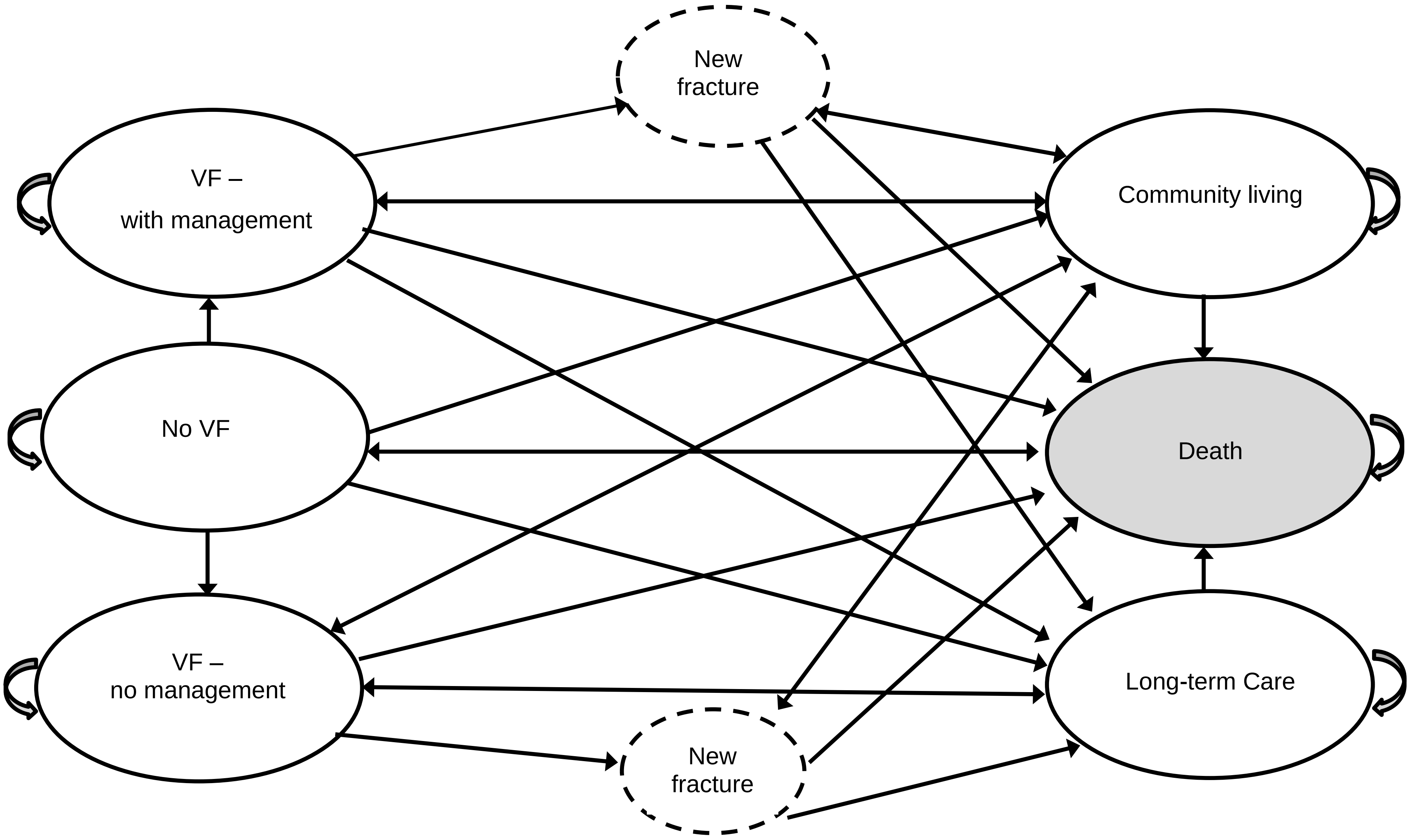
Future economic modelling, could include the decision tree, developed as part of this assessment, into a Markov model (Figure 2). This could involve temporary states to monitor the number of additional vertebral or other (for example hip fractures) to demonstrate potential benefits of treatment of vertebral fractures. With all states having allowed transition to an absorbing death state. In Figure 2 the EAG has separated out whether vertebral fracture is present or not and is treated from where someone lives. In reality, states may combine these two features for example, a person with a treated vertebral fracture may be in one of two states: treated vertebral fracture, community; treated vertebral fracture, long-term care, with the option of moving between these states over time.

Table 30: Evidence generation recommendations

#	Research question	Technologies	Recommended study design	Outcomes
1.	How many images are eligible to be processed by each AI technology?	All AI technologies	Prospective cohort	Throughput in NHS, cost per scan, failure rate (and reasons for failure), prevalence of vertebral fracture.
2.	What is the current sensitivity and specificity of radiology reports reporting opportunistic detection of vertebral fragility/compression fractures on diagnostic images (for reasons other than vertebral fracture detection) when compared with a reference standard of radiologists or reporting radiographer?	Standard of care (X-ray, CT)	Diagnostic accuracy (applied to retrospective images)	Diagnostic accuracy (per patient), prevalence of vertebral fracture.
3.	What is the management pathway of patients suspected of vertebral fracture opportunistically detected on X-ray or CT imaging?	Standard of care, and all AI technologies	Retrospective cohort (service evaluation)	No. of patients referred for spine X-ray or DEXA scan. No. of patients receiving medication for osteoporosis.
4.	Diagnostic accuracy of AI technologies in reviewing X-ray and CT images (not for vertebral fracture detection)? How many images are successfully processed by the AI technology?	All AI technologies	Diagnostic accuracy (applied to retrospective images)	Diagnostic accuracy (per patient), additional vertebral fracture detected, false positives.
5.	What staff roles and levels of experience/qualification were used in the initial reporting of eligible scans (and opportunistic detection of vertebral fracture)	Standard of care (X-ray, CT)	Retrospective cohort (service evaluation)	Staff costs
6.	What is the utility gain from detecting and treating vertebral fracture detected incidentally	All AI technologies	Prospective cohort with matched controls	Utilities, QALYs
7.	Healthcare acceptability of AI tool	All AI technologies	Survey	Ease of use, time taken to process and report image with AI assistance.
8.	Given AI technologies vary in terms of the patient group they are used for, what is the optimal mix of AI technologies for detection of vertebral fractures in the NHS	All AI technologies	Retrospective cohort comparing performance of all AI technologies and modelling cost-effectiveness of single and combination AI technologies	Costs, QALYs and cost per QALY

Abbreviations: AI, artificial intelligence; DEXA, dual energy X-ray absorptiometry (bone density); QALY, Quality-adjusted life year.

Figure 2: Proposed Markov model (following decision tree) for future economic evaluations [adapted from Curl et al. 2024]



Abbreviations: VF, Vertebral Fracture.

9. Conclusions

The EAG has identified preliminary evidence that AI technologies with high specificity may have benefit in the opportunistic detection of vertebral fragility fractures. The results of de novo economic modelling show that the individual technologies considered in this early value assessment could be similarly cost-effective under certain circumstances. However, additional evidence should be collected to confirm parameter estimates as some plausible combinations of parameter estimates may result in the AI technologies not being cost-effective. The workflow and impact on the NHS workforce should be closely monitored to ensure the effects of false positives do not further exacerbate referrals and waiting times for additional imaging. Such consequences may unnecessarily increase patient exposure to ionising radiation, further delay treatment (considering limited DEXA scanner availability), and add work pressures to radiologists and reporting radiographers (where there is a recognised capacity problem).

10. References

Annalise.AI. (No date). [AiC] Report.

Avicenna.AI. (2024). [AiC] Abstract Submission.

Bendtsen, M. G., and Hitz, M. F. (2024). Opportunistic Identification of Vertebral Compression Fractures on CT Scans of the Chest and Abdomen, Using an AI Algorithm, in a Real-Life Setting. *Calcified Tissue International*, *114*(5), 468-479. https://dx.doi.org/10.1007/s00223-024-01196-2

Borgstrom, F., Zethraeus, N., Johnell, O., Lidgren, L., Ponzer, S., Svensson, O., Abdon, P., Ornstein, E., Lunsjo, K., Thorngren, K. G., Sernbo, I., Rehnberg, C., & Jonsson, B. (2006). Costs and quality of life associated with osteoporosis-related fractures in Sweden. *Osteoporos Int*, *17*(5), 637-650. https://doi.org/10.1007/s00198-005-0015-8

Chappell, D., Brown, J., Sanders, V., Boylan, J., Turton, J., Threlkeld, J., Sampson, M., Santos, T., Guy, A., Curtis, E., Chisholm, C., Gerety, E., Eckert, R., Poole, K., Stone, M., Harvey, N., & Javaid, M. (2024). *Poster*

Connacher, S., Eckert, R., Mansour, R., & Javaid, M. K. (2019). Using artificial intelligence technology to improve case finding for vertebral fractures in the fracture liaison service (FLS) setting. *Osteoporosis International*, *30*(SUPPL 2), S428. https://dx.doi.org/10.1007/s00198-019-04993-w (World Congress on Osteoporosis, Osteoarthritis and Musculoskeletal Diseases, WCO-IOF-ESCEO 2019. Paris France.)

Connor, J., Threlkeld, J., Chisholm, C., Chappell, D., Sanders, V., Vindlacheruvu, M., Boylan, J., Turton, J., Santos, T., Sampson, M., Guy, A., Eckert, R., Curtis, E., Stone, M. D., Cooper, F., Kimmel, Y., Poole, K. E. S., Sahota, O., & Javaid, M. K. (2024). Artificial Intelligence Enabled Vertebral Fracture Identification for Fracture Liaison Services: opportunities and challenges- The ADOPT Study [Abstract]. *Journal of Bone and Mineral Research*, *39*(Supplement 1), 296-297.

Curl, P. K., Jacob, A., Bresnahan, B., Cross, N. M., & Jarvik, J. G. (2024). Cost-Effectiveness of Artificial Intelligence-Based Opportunistic Compression Fracture Screening of Existing Radiographs. *Journal of the American College of Radiology*, *21*(9), 1489-1496. https://dx.doi.org/10.1016/j.jacr.2023.11.029

Dai, M., Tiu, B. C., Schlossman, J., Ayobi, A., Castineira, C., Kiewsky, J., Avare, C., Chaibi, Y., Chang, P., Chow, D., & Soun, J. E. (2025). Validation of a Deep Learning Tool for Detection of Incidental Vertebral Compression Fractures. *Journal of Computer Assisted Tomography*, 10.1097/RCT.0000000000001726. https:/dx.doi.org/10.1097/RCT.0000000000001726

Dalal, G., Bromiley, P. A., Kariki, E. P., Luetchens, S., Cootes, T. F., & Payne, K. (2022). Understanding current UK practice for the incidental identification of vertebral fragility fractures from CT scans: an expert elicitation study. *Aging Clin Exp Res*, *34*(8), 1909-1918. https://doi.org/10.1007/s40520-022-02124-w

Dalal, G., Youn, J. H., Kariki, E., Bromiley, P. A., Luetchens, S., Cootes, T. F., & Payne, K. (2021). PMD5 Designing a Complex Intervention Using Early Model-Based Cost-Effectiveness Analysis: A Case Study of an Artificial Intelligence-based Algorithm to Identify Vertebral Fragility Fractures. *Value in Health*, *24*(Supplement 1), S123-S124. https:/dx.doi.org/10.1016/j.jval.2021.04.603 (HEOR: Evolving for Tomorrow's Challenges. Virtual, Online.)

Davenport, M. S. (2023). Incidental Findings and Low-Value Care. *AJR Am J Roentgenol*, *221*(1), 117-123. https://doi.org/10.2214/AJR.22.28926

Frias, E. G. (2023). Consecutive cohort analysis with natural language processing (NLP) and multi-finding AI algorithm for chest radiographs: multidimensional opportunity for QA/QC [Abstract] RSNA 2023, Chicago.

Ghatak, A., Hillis, J. M., Mercaldo, S. F., Newbury-Chaet, I., Chin, J. K., Digumarthy, S. R., Rodriguez, K., Muse, V. V., Andriole, K. P., Dreyer, K. J., Kalra, M. K., & Bizzo, B. C. (2024). The Potential Clinical Utility of an Artificial Intelligence Model for Identification of Vertebral Compression Fractures in

Chest Radiographs. *Journal of the American College of Radiology*, 220-229. https:/dx.doi.org/10.1016/j.jacr.2024.08.026

Guenoun, D., Quemeneur, M. S., Ayobi, A., Castineira, C., Quenet, S., Kiewsky, J., Mahfoud, M., Avare, C., Chaibi, Y., & Champsaur, P. (2025). Automated vertebral compression fracture detection and quantification on opportunistic CT scans: A performance evaluation. *Clinical Radiology*, 106831.

Gutierrez, L., Roskell, N., Castellsague, J., Beard, S., Rycroft, C., Abeysinghe, S., Shannon, P., Gitlin, M., & Robbins, S. (2012). Clinical burden and incremental cost of fractures in postmenopausal women in the United Kingdom. *Bone*, *51*(3), 324-331. https://doi.org/10.1016/j.bone.2012.05.020

Howlett, D. C., Drinkwater, K. J., Mahmood, N., Salman, L., Griffin, J., Javaid, M. K., Retnasingam, G., Marzoug, A., Greenhalgh, R., & Collaborating Author, G. (2023). Radiology reporting of incidental osteoporotic vertebral fragility fractures present on CT studies: results of UK national re-audit. *Clin Radiol*, *78*(12), e1041-e1047. https://doi.org/10.1016/j.crad.2023.09.004

IB Lab. (2023). [CiC] Report.

Jones, K. C., Weatherly, H., Birch, S., Casteli, A., Chalkley, M., Dargan, A., Forder, J. E., Gao, M., Hinde, S., Markham, S., Premji, S., Findlay, D., & Teo, H. (2024). *Unit Costs of Health and Social Care 2023 Manual. Technical report.*

Jones, C.M., Danaher, L., Milne, M.R., Tang, C., Seah, J., Oakden-Rayner, L., Johnson, A., Buchlak, Q.D., Esmaili, N..(2021) 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. 11*(12):e052902. doi: 10.1136/bmjopen-2021-052902

Kolanu, N., Silverstone, E. J., Ho, B. H., Pham, H., Hansen, A., Pauley, E., Quirk, A. R., Sweeney, S. C., Center, J. R., & Pocock, N. A. (2020). Clinical Utility of Computer-Aided Diagnosis of Vertebral Fractures From Computed

Tomography Images. *Journal of Bone and Mineral Research*, *35*(12), 2307-2312. https://dx.doi.org/10.1002/jbmr.4146

Lenchik, L., Rogers, L. F., Delmas, P. D., & Genant, H. K. (2004). Diagnosis of osteoporotic vertebral fractures: importance of recognition and description by radiologists. *AJR Am J Roentgenol*, *183*(4), 949-958. https://doi.org/10.2214/ajr.183.4.1830949

Nanjayan, S. K., John, J., Swamy, G., Mitsiou, K., Tambe, A., & Abuzakuk, T. (2014). Predictors of change in 'discharge destination' following treatment for fracture neck of femur. *Injury*, *45*(7), 1080-1084. https://doi.org/10.1016/j.injury.2014.02.005

Nicolaes, J., Liu, Y., Zhao, Y., Huang, P., Wang, L., Yu, A., Dunkel, J., Libanati, C., & Cheng, X. (2024). External validation of a convolutional neural network algorithm for opportunistically detecting vertebral fractures in routine CT scans. *Osteoporosis International*, *35*(1), 143-152. https://dx.doi.org/10.1007/s00198-023-06903-7

Nicolaes, J., Skjodt, M. K., Raeymaeckers, S., Smith, C. D., Abrahamsen, B., Fuerst, T., Debois, M., Vandermeulen, D., & Libanati, C. (2023). Towards improved identification of vertebral fractures in routine computed tomography (CT) scans: development and external validation of a machine learning algorithm. *J Bone Miner Res*, *38*(12), 1856-1866. https://doi.org/10.1002/jbmr.4916

Oppenheimer, J., Luken, S., Hamm, B., & Niehues, S. M. (2023). A Prospective Approach to Integration of AI Fracture Detection Software in Radiographs into Clinical Workflow. *Life (Basel)*, *13*(1). https://doi.org/10.3390/life13010223

Oppenheimer, J., Luken, S., S, G., Hamm, B., & Niehues, S. M. (2024). An overview of the performance of AI in fracture detection in lumbar and thoracic spine radiographs on a per vertebra basis. *Skeletal Radiology*, *53*(8), 1563-1571. https://dx.doi.org/10.1007/s00256-024-04626-2

Page, J. H., Moser, F. G., Maya, M. M., Prasad, R., & Pressman, B. D. (2023). Opportunistic CT Screening-Machine Learning Algorithm Identifies Majority of Vertebral Compression Fractures: A Cohort Study. *JBMR Plus*, *7*(8), e10778. https:/dx.doi.org/10.1002/jbm4.10778

Pereira, R. F. B., Helito, P. V. P., Leao, R. V., Bordalo Rodrigues, M., de Paula Correa, M. F., & Rodrigues, F. V. (2024). Accuracy of an artificial intelligence algorithm for detecting moderate-to-severe vertebral compression fractures on abdominal and thoracic computed tomography scans. *Radiologia Brasileira*, *57*, e20230102. https:/dx.doi.org/10.1590/0100-3984.2023.0102 (Acuracia de um algoritmo de inteligencia artificial na deteccao de fraturas compressivas moderadas a graves na tomografia computadorizada abdominal e toracica)

Roux, C., Rozes, A., Reizine, D., Hajage, D., Daniel, C., Maire, A., Breant, S., Taright, N., Gordon, R., Fechtenbaum, J., Kolta, S., Feydy, A., Briot, K., & Tubach, F. (2022). Fully automated opportunistic screening of vertebral fractures and osteoporosis on more than 150 000 routine computed tomography scans. *Rheumatology (United Kingdom)*, *61*(8), 3269-3278. https:/dx.doi.org/10.1093/rheumatology/keab878

Royal Osteoporosis Society. (2021). *State of the Nation report: Vertebral fracture identification in 2021*. https://strwebprdmedia.blob.core.windows.net/media/fcodhs1m/state-of-the-nation-report-vertebral-fracture-identification-in-2021.pdf

Shelmerdine, S.C., Togher, D., Rickaby, S., Dean, G. (2024). Artificial intelligence (AI) implementation within the National Health Service (NHS): the South West London AI Working Group experience. *Clin Radiol*. *79*(9):665-672. doi: 10.1016/j.crad.2024.05.018

Spangeus, A., Bjerner, T., Lindblom, M., Gotz, C., Hummer, A., Salzlechner, C., & Woisetschlager, M. (2025). Breaking the silence: AI's contribution to detecting vertebral fractures in opportunistic CT scans in the elderly-a validation study. *Arch Osteoporos*, *20*(1), 42. https://doi.org/10.1007/s11657-025-01524-5

Svedbom, A., Borgstom, F., Hernlund, E., Strom, O., Alekna, V., Bianchi, M. L., Clark, P., Curiel, M. D., Dimai, H. P., Jurisson, M., Kallikorm, R., Lember, M., Lesnyak, O., McCloskey, E., Sanders, K. M., Silverman, S., Solodovnikov, A., Tamulaitiene, M., Thomas, T., Kanis, J. A. (2018). Quality of life for up to 18 months after low-energy hip, vertebral, and distal forearm fractures-results from the ICUROS. *Osteoporos Int*, *29*(3), 557-566. https://doi.org/10.1007/s00198-017-4317-4

Talwar, A. (2023). Missed diagnosis on chest X-ray: auditing large volumes of data with the help of comprehensive artificial intelligence [Abstract] RSNA 2023, Chicago.

The Royal College of Radiologists. (2021). *Radiological guidance for the recognition and reporting of osteoporotic vertebral fragility fractures (VFFs)*. https://www.rcr.ac.uk/our-services/all-our-publications/clinical-radiology-publications/radiological-guidance-for-the-recognition-and-reporting-of-osteoporotic-vertebral-fragility-fractures-vffs/

Wiklund, P., Buchebner, D., & Geijer, M. (2024). Vertebral compression fractures at abdominal CT: underdiagnosis, undertreatment, and evaluation of an AI algorithm. *J Bone Miner Res*, *39*(8), 1113-1119. https://doi.org/10.1093/jbmr/zjae096

11. Appendices

Appendix A – Literature searching

Appendix A1: Search strategies (clinical and economic)

The search strategy was designed in Embase and translated to other databases; where accurate translation was not possible due to limited source functionality, simpler searches were conducted using terms relating to device names or vertebral fragility fractures. The same core strategy was used for the clinical effectiveness and economic searches, and a published filter designed to retrieve economic evaluations was used for the latter in MEDLINE and Embase (CADTH, 2016).

The searches were run between 17 and 19 of February on the following databases:

- Embase (OVID), 1974 to February 2025
- MEDLINE (OVID), 1946 to February 2025
- CENTRAL (Cochrane Library via Wiley), searched February 2025
- INAHTA (https://database.inahta.org/), searched February 2025
- RePEC IDEAS (https://ideas.repec.org) searched February 2025
- EconPapers (https://econpapers.repec.org) searched February 2025
- Google Scholar searched February 2025

A separate targeted search was subsequently carried out in Embase and Google Scholar to find economics studies explicitly addressing long term impact of treatment.

Records retrieved from clinical search (19 February 2025)

Database Name	Total number of records retrieved
MEDLINE	409
Embase	863
CENTRAL	47
INAHTA	93
Google Scholar	34
Total	**1446**
Duplicates	**326**

Total to screen	**1120**

Records retrieved from economics search (17 February 2025)

Database Name	Total number of records retrieved
MEDLINE	48
Embase	73
EconPapers	29
RePec IDEAS	50
INAHTA	61
Total	**269**
Duplicates	**81**
Total to screen	**188**

Searches were conducted as described in the Final Scope to identify evidence in clinical effectiveness and economic evaluations and economic models.

This search identified 1120 titles and abstracts for clinical evidence and 188 for economic evaluations after de-duplication.

Economics Searches

Ovid MEDLINE(R) and In-Process, In-Data-Review & Other Non-Indexed Citations <1946 to February 14, 2025>

#	Query	Records
1	("Annalise Enterprise" or "Annalise.AI").af.	17
2	(Annalise adj2 CXR).af.	1
3	(boneview* or gleamer).af.	31
4	(boneview adj gleamer).af.4	4
5	("BriefCase-Triage" or "Aidoc Medical").af.	2
6	(briefcase adj2 aidoc medical).af.	0
7	("CINA-VCF quantix" or "Avicenna.AI").af.	11
8	(c-spine or briefcase).af. and (ai or artificial intelligence).tw.	2
9	aidoc.af. and (spinal or spine or vertebr* or bone or osteop* or fractur*).tw.	5
10	(Healthost or Healthvcf or "nanox.ai" or nanox).af.	45
11	((IB lab or ImageBiopsy Lab) and flamingo).af.	0
12	"TechCare Spine (Milvue)".af.0	0
13	(Techcare adj2 Milvue).af.0	0
14	or/1-13	113
15	fractures, compression/	3585

#	Query	Records
16	osteoporotic fractures/	9093
17	spinal fractures/	18911
18	((compress* or spine or spinal) adj2 fracture*).ti,ab,kw.	12957
19	((vertebra* or osteoporo*) adj2 (fragil* or fracture*)).ti,ab,kw.	29353
20	(VFF or VFFS or VCF).ti,ab,kw.	1662
21	or/15-20	48420
22	diagnostic imaging/	47424
23	radiography/	330259
24	absorptiometry, photon/	27143
25	exp magnetic resonance imaging/	566109
26	exp tomography, emission-computed/	140765
27	x rays/	32659
28	(radiogra* or ct or (comput* adj4 tomogra*) or absorptiometry or dexa or dxa or magnetic resonance or mri or mrs or nmr* or x ray or xray).tw.	1975411
29	or/22-28	2514453
30	Algorithm*.ti,kf.	77655
31	(algorithm* adj2 (learn* or automate* or detect* or predict* or treatment* or therap* or radiolog* or AI or DL or ML or data or dataset* or base* or classif*)).ab.	109206
32	Artificial Intelligen*.ti,ab,kf.	64550
33	AI.ti,kf.	15768
34	(machine adj2 learn*).ti,ab,kf.	135901
35	machinelearn*.ti,ab,kf.	26
36	(deep adj2 learn*).ti,ab,kf.	78852
37	deeplearn*.ti,ab,kf.	30
38	neural network*.ti,ab,kf.	123062
39	(convolutional adj1 network*).ti,ab,kf.	4153
40	automate*.ti.	51038
41	(automate* adj3 (system* or score* or software* or analysis* or analyse* or risk* or evaluat* or tool* or detect* or process*)).ab,kf.	45749
42	(vector machine* or svm*).ti,ab,kf.	36271
43	radiomic*.ti,ab,kf.	13085
44	((supervised or unsupervised) adj3 (classifier* or prediction*)).ti,ab,kf.	1039
45	or/30-44	514816
46	Economics/	27545
47	Cost/	52178
48	exp Health Economics/	1741856
49	Budget/	11914
50	budget*.ti,ab,kf.	39040
51	(economic* or cost or costs or costly or costing or price or prices or pricing or pharmacoeconomic* or pharmaco-economic* or expenditure or expenditures or expense or expenses or financial or finance or finances or financed).ti,kf.	304396
52	(economic* or cost or costs or costly or costing or price or prices or pricing or pharmacoeconomic* or pharmaco-economic* or expenditure or expenditures or expense or	422729

#	Query	Records
	expenses or financial or finance or finances or financed).ab. /freq=2	
53	(cost* adj2 (effective* or utilit* or benefit* or minimi* or analy* or outcome or outcomes)).ab,kf.	236342
54	(value adj2 (money or monetary)).ti,ab,kf.	3297
55	Statistical Model/	101826
56	exp economic model/	16712
57	economic model*.ab,kf.	4538
58	Probability/	61762
59	markov.ti,ab,kf.	31896
60	monte carlo method/	33777
61	monte carlo.ti,ab,kf.	64964
62	Decision Theory/	974
63	Decision Tree/	12515
64	(decision* adj2 (tree* or analy* or model*)).ti,ab,kf.	46770
65	or/46-64	2433971
66	14 and 65	8
67	21 and 29 and 45 and 65	48
68	limit 67 to english language	48
69	limit 68 to (letter or historical article or comment or editorial or news)	0
70	68 not 69	48

Embase <1974 to 2025 February 28>

#	Query	Results from 3 Mar 2025
1	("Annalise Enterprise" or "Annalise.AI").af.	20
2	(Annalise and CXR).af.	10
3	(boneview* and gleamer).af.	14
4	("BriefCase-Triage" or "Aidoc Medical").af.	8
5	(briefcase and "Aidoc Medical").af.	3
6	("CINA-VCF quantix" or "Avicenna.AI").af.	13
7	(c-spine or briefcase).af. and (ai or artificial intelligence).tw.	12
8	aidoc.af. and (spinal or spine or vertebr* or bone or osteop* or fractur*).tw.	7
9	(Healthost or Healthvcf or "nanox.ai" or nanox).af.	50

#	Query	Results from 3 Mar 2025
10	"Zebra Medical".af. and (spinal or spine or vertebr* or bone or osteop* or fractur*).tw.	14
11	((IB lab or ImageBiopsy Lab) and (flamingo or bonebot)).af.	0
12	"TechCare Spine (Milvue)".af.	0
13	(Techcare and Milvue).af.	0
14	or/1-13	126
15	fractures, compression/	9,274
16	osteoporotic fractures/	26,296
17	spinal fractures/	35,943
18	((compress* or spine or spinal) adj2 fracture*).ti,ab,kw.	18,030
19	((vertebra* or osteoporo*) adj2 (fragil* or fracture*)).ti,ab,kw.	47,470
20	(VFF or VFFS or VCF).ti,ab,kw.	2,609
21	or/15-20	82,319
22	diagnostic imaging/	270,055
23	radiography/	206,444
24	absorptiometry, photon/	4,819
25	exp magnetic resonance imaging/	1,376,435
26	exp tomography, emission-computed/	354,007
27	x rays/	92,007
28	(radiogra* or ct or (comput* adj4 tomogra*) or absorptiometry or dexa or dxa or magnetic resonance or mri or mrs or nmr* or x ray or xray or x-ray).tw.	2,741,649
29	or/22-28	3,667,227
30	Algorithm*.ti,kf.	96,864
31	(algorithm* adj2 (learn* or automate* or detect* or predict* or treatment* or therap* or radiolog* or AI or DL or ML or data or dataset* or base* or classif*)).ab.	144,453
32	Artificial Intelligen*.ti,ab,kf.	78,358
33	AI.ti,kf.	19,454

#	Query	Results from 3 Mar 2025
34	(machine adj2 learn*).ti,ab,kf.	161,914
35	machinelearn*.ti,ab,kf.	332
36	(deep adj2 learn*).ti,ab,kf.	93,197
37	deeplearn*.ti,ab,kf.	178
38	neural network*.ti,ab,kf.	147,166
39	(convolutional adj1 network*).ti,ab,kf.	4,919
40	automate*.ti.	66,941
41	(automate* adj3 (system* or score* or software* or analysis* or analyse* or risk* or evaluat* or tool* or detect* or process*)).ab,kf.	68,648
42	(vector machine* or svm*).ti,ab,kf.	44,364
43	radiomic*.ti,ab,kf.	18,417
44	((supervised or unsupervised) adj3 (classifie* or predict*)).ti,ab,kf.	1,886
45	or/30-44	647,296
46	Economics/	245,473
47	Cost/	65,370
48	exp Health Economics/	1,115,453
49	Budget/	35,833
50	budget*.ti,ab,kf.	52,253
51	(economic* or cost or costs or costly or costing or price or prices or pricing or pharmacoeconomic* or pharmaco-economic* or expenditure or expenditures or expense or expenses or financial or finance or finances or financed).ti,kf.	376,106

#	Query	Results from 3 Mar 2025
52	(economic* or cost or costs or costly or costing or price or prices or pricing or pharmacoeconomic* or pharmaco-economic* or expenditure or expenditures or expense or expenses or financial or finance or finances or financed).ab. /freq=2	598,345
53	(cost* adj2 (effective* or utilit* or benefit* or minimi* or analy* or outcome or outcomes)).ab,kf.	328,820
54	(value adj2 (money or monetary)).ti,ab,kf.	4,486
55	Statistical Model/	179,526
56	exp economic model/	4,772
57	economic model*.ab,kf.	7,044
58	Probability/	162,899
59	markov.ti,ab,kf.	42,349
60	monte carlo method/	56,516
61	monte carlo.ti,ab,kf.	69,584
62	Decision Theory/	1,901
63	Decision Tree/	27,672
64	(decision* adj2 (tree* or analy* or model*)).ti,ab,kf.	63,247
65	or/46-64	2,195,697
66	14 and 65	11
67	21 and 29 and 45 and 65	126
68	67 not (editorial or letter).pt.	125
69	(conference abstract* or conference review or conference paper or conference proceeding).db,pt,su.	6,172,366
70	68 not 69	77
71	limit 70 to english language	73

Line 68: Used to exclude editorial and opinion pieces.

INAHTA

1	(("Annalise Enterprise" or "Annalise.AI" or Annalise adj2 CXR OR boneview* or gleamer OR
2	(boneview AND gleamer) OR
3	"BriefCase-Triage" or "Aidoc Medical" OR (briefcase AND aidoc medical) OR
4	"CINA-VCF quantix" OR
5	"Avicenna.AI" OR
6	((c-spine or briefcase) and (ai or artificial intelligence)) OR
7	Healthost or Healthvcf OR
8	"nanox.ai" OR
9	"IB Lab FlAMINGO" OR
10	IB lab or ImageBiopsy Lab OR
11	"TechCare Spine (Milvue)" OR
12	(Techcare AND Milvue))) AND
13	((((fractures, compression/[mh] OR
14	osteoporotic fractures/[mh] OR
15	spinal fractures/[mh] OR
16	((compress* or spine or spinal) AND fracture*)[Title] OR
17	((compress* or spine or spinal) AND fracture*)[abs] OR
18	((vertebra* or osteoporo*) AND (fragil* or fracture*))[Title] OR
19	((vertebra* or osteoporo*) AND (fragil* or fracture*))[abs] OR
20	(VFF or VFFS)[Title] OR (VFF or VFFS)[abs])) OR
21	((((opportunis* or incident* or uninten* or unsuspect* or fortuit*) AND
22	(detect* or present* or find* or identif*))[Title] OR
23	((opportunis* or incident* or uninten* or unsuspect* or fortuit*) AND (detect* or present* or
24	find* or identif*))[abs]) AND)) AND
25	((diagnostic imaging/[mh] OR
26	Radiography/[mh] OR
27	absorptiometry, photon/[mh] OR
28	magnetic resonance imaging/[mh] OR
29	tomography, emission-computed/[mh] OR
30	x rays/[mh] OR
31	(radiogra* or ct or (comput* AND tomogra*) or absorptiometry or dexa or dxa or magnetic
32	resonance or mri or mrs or nmr* or x ray or xray)[Title] OR
33	(radiogra* or ct or (comput* AND tomogra*) or absorptiometry or dexa or dxa or magnetic resonance or mri or mrs or nmr* or x ray or xray)[abs]) OR) AND
34	(((algorithm* or artificial intelligen* or automat* or radiomic*)[Title] OR
35	(algorithm* or artificial intelligen* or automat* or radiomic*)[abs] OR
36	(algorithm* or artificial intelligen* or automat* or radiomic*)[Keywords] OR (algorithm* AND (learn* or automate* or detect* or predict* or treatment* or therap* or
37	radiolog* or AI or DL or ML or data or dataset* or base* or classif*))[Title] OR
38	(algorithm* AND (learn* or automate* or detect* or predict* or treatment* or therap* or radiolog* or AI or DL or ML or data or dataset* or base* or classif*))[abs] OR
39	(algorithm* AND (learn* or automate* or detect* or predict* or treatment* or therap* or radiolog* or AI or DL or ML or data or dataset* or base* or classif*))[Keywords] OR
40	(machine or deep) AND learn*)[Title] OR

41	(machine or deep) AND learn*)[abs] OR
42	(machine or deep) AND learn*)[Keywords] OR
43	(neural or convolutional) AND network*)[Title] OR
44	(neural or convolutional) AND network*)[abs] OR
45	(neural or convolutional) AND network*)[Keywords] (automate* AND (system* or score* or software* or analysis* or analyse* or risk* or evaluat* or tool* or detect* or process*))[Title] OR
46	(automate* AND (system* or score* or software* or analysis* or analyse* or risk* or evaluat* or tool* or detect* or process*))[abs] OR
47	(automate* AND (system* or score* or software* or analysis* or analyse* or risk* or evaluat* or tool* or detect* or process*))[Keywords] OR
48	(vector machine* or svm*)[Title] OR
49	(vector machine* or svm*)[abs] OR
50	(vector machine* or svm*)[Keywords] OR
51	((supervised or unsupervised) AND (classifier* or prediction*))[Title] OR
52	((supervised or unsupervised) AND (classifier* or prediction*))[abs]) OR
53	((supervised or unsupervised) AND (classifier* or prediction*))[Keywords])) AND
54	((Economics/[mh] OR
55	Cost/[mh] OR
56	Health Economics/[mh] OR
57	Budget/[mh] OR
58	budget[Title] OR
59	budget[abs] OR
60	(economic* or cost or costs or costly or costing or price or prices or pricing or pharmacoeconomic* or pharmaco-economic* or expenditure or expenditures or expense or expenses or financial or finance or finances or financed)[Title] OR
61	(economic* or cost or costs or costly or costing or price or prices or pricing or pharmacoeconomic* or pharmaco-economic* or expenditure or expenditures or expense or expenses or financial or finance or finances or financed)[Keywords] OR
62	(cost* AND (effective* or utilit* or benefit* or minimi* or analy* or outcome or outcomes))[abs] OR
63	(cost* AND (effective* or utilit* or benefit* or minimi* or analy* or outcome or outcomes))[Keywords] OR
64	(value AND (money or monetary))[Title] OR
65	(value AND (money or monetary))[abs]OR (value AND (money ormonetary))[[Keywords] OR
66	Statistical Model/[mh] OR
67	economic model/[mh] OR
68	economic model*[abs] OR
69	economic model*[Keywords] OR
70	Probability/[mh] OR markov[Title] OR
71	markov[abs] OR
72	markov[Keywords] OR
73	monte carlo method/ OR
74	monte carlo[Title] OR
75	monte carlo[abs] OR
76	monte carlo[Keywords] OR
77	Decision Theory/ OR

78	Decision Tree/ OR
79	(decision* AND (tree* or analy* or model*))[Title] OR
80	(decision* AND (tree* or analy* or model*))[abs] OR
81	(decision* AND (tree* or analy* or model*))[Keywords])))
82	Decision Tree/ OR (decision* AND (tree* or analy* or model*))[Title] OR
83	(decision* AND (tree* or analy* or model*))[abs] OR (decision* AND (tree* or analy* or model*))[Keywords])))

RePec (IDEAS)

(("osteoporotic fracture" | "osteoporotic fragility" | "vertebral fracture" | "vertebral fragility" | "compression fracture" | "spinal fracture" | "spinal fragility" | "VFF" | "VFFS" | "VCF") AND (algorithm | algorithmic | "artificial intelligence" | "machine learning" | "deep learning" | "neural network" | "convolutional network" | "supervised learning" | "unsupervised learning" | "vector machine" | classifier | prediction | automate))

EconPapers

(("osteoporotic fracture" OR "osteoporotic fragility" OR "vertebral fracture" OR "vertebral fragility" OR "compression fracture" OR "spinal fracture" OR "spinal fragility" OR "VFF" OR "VFFS" OR "VCF") AND (algorithm OR algorithmic OR "artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR "convolutional network" OR "supervised learning" OR "unsupervised learning" OR "vector machine" OR classifier OR prediction OR automate))

Clinical effectiveness searches

Database(s): Embase 1974 to 2025 February 17

Search Strategy:

#	Searches	Results
1	("Annalise Enterprise" or "Annalise.AI").af.	19
2	(Annalise and CXR).af.	9
3	(boneview* and gleamer).af.	14
4	("BriefCase-Triage" or "Aidoc Medical").af.	8
5	(briefcase and "Aidoc Medical").af.	3
6	("CINA-VCF quantix" or "Avicenna.AI").af.	12
7	(c-spine or briefcase).af. and (ai or artificial intelligence).tw.	12
8	aidoc.af. and (spinal or spine or vertebr* or bone or osteop* or fractur*).tw.	7
9	(Healthost or Healthvcf or "nanox.ai" or "nanox ai").af.	19
10	"Zebra Medical".af. and (spinal or spine or vertebr* or bone or osteop* or fractur*).tw.	14
11	((IB lab or ImageBiopsy Lab) and (Flamingo or bonebot)).af.	0
12	"TechCare Spine (Milvue)".af.	0

13	(Techcare and Milvue).af.	0
14	or/1-13	93
15	fractures, compression/	9240
16	osteoporotic fractures/	26262
17	spinal fractures/	35817
18	(VFF or VFFS or VCF).ab,ti,kw.	2586
19	((compress* or spine or spinal) adj2 fracture*).ab,ti,kw.	17947
20	((vertebra* or osteoporo*) adj2 (fragil* or fracture*)).ab,ti,kw.	47336
21	(VFF or VFFS).ab,ti,kw.	300
22	or/15-21	82041
23	((opportunis* or incident* or uninten* or unsuspect* or fortuit*) adj2 (detect* or present* or find* or identif*)).ti,ab,kw.	42126
24	diagnostic imaging/	269250
25	radiography/	205267
26	absorptiometry, photon/	4789
27	exp magnetic resonance imaging/	1372451
28	exp tomography, emission-computed/	352760
29	x rays/	91667
30	(radiogra* or ct or (comput* adj4 tomogra*) or absorptiometry or dexa or dxa or magnetic resonance or mri or mrs or nmr* or x ray or xray or x-ray).tw.	2731270
31	or/24-30	3653303
32	Algorithm*.ti,kf.	96571
33	(algorithm* adj2 (learn* or automate* or detect* or predict* or treatment* or therap* or radiolog* or AI or DL or ML or data or dataset* or base* or classif*)).ab.	143974
34	Artificial Intelligen*.ti,ab,kf.	77670
35	AI.ti,kf.	19152
36	(machine adj2 learn*).ti,ab,kf.	160874
37	machinelearn*.ti,ab,kf.	332
38	(deep adj2 learn*).ti,ab,kf.	92605
39	deeplearn*.ti,ab,kf.	177
40	neural network*.ti,ab,kf.	146618
41	(convolutional adj1 network*).ti,ab,kf.	4892
42	automate*.ti.	66693
43	(automate* adj3 (system* or score* or software* or analysis* or analyse* or risk* or evaluat* or tool* or detect* or process*)).ab,kf.	68384
44	(vector machine* or svm*).ti,ab,kf.	44202
45	radiomic*.ti,ab,kf.	18298
46	((supervised or unsupervised) adj3 (classifie* or predict*)).ti,ab,kf.	1881
47	or/32-46	644224
48	(editorial or letter).pt.	2168939
49	animals/ not (animals/ and humans/)	1232353
50	(clinical conference or editorial or letter or news).pt.	2168939
51	or/48-50	3382589

52	14 not 51	91
53	(22 and 23 and 31 and 47) not 51	35
54	53 not 52	30
55	(22 and 31 and 47) not 51	810
56	55 not (54 or 52)	759

Database(s): Ovid MEDLINE(R) and In-Process, In-Data-Review & Other Non-Indexed Citations 1946 to February 17, 2025

Search Strategy:

#	Searches	Results
1	("Annalise Enterprise" or "Annalise.AI").af.	17
2	(Annalise and CXR).af.	3
3	(boneview* and gleamer).af.	9
4	("BriefCase-Triage" or "Aidoc Medical").af.	2
5	(briefcase and "Aidoc Medical").af.	0
6	("CINA-VCF quantix" or "Avicenna.AI").af.	11
7	(c-spine or briefcase).af. and (ai or artificial intelligence).tw.	2
8	aidoc.af. and (spinal or spine or vertebr* or bone or osteop* or fractur*).tw.	5
9	(Healthost or Healthvcf or "nanox.ai" or "nanox ai").af.	21
10	"Zebra Medical".af. and (spinal or spine or vertebr* or bone or osteop* or fractur*).tw.	7
11	((IB lab or ImageBiopsy Lab) and (flamingo or bonebot)).af.	0
12	"TechCare Spine (Milvue)".af.	0
13	(Techcare and Milvue).af.	0
14	or/1-13	74
15	fractures, compression/	3585
16	osteoporotic fractures/	9093
17	spinal fractures/	18911
18	((compress* or spine or spinal) adj2 fracture*).ti,ab,kw.	12958
19	((vertebra* or osteoporo*) adj2 (fragil* or fracture*)).ti,ab,kw.	29358
20	(VFF or VFFS or VCF).ti,ab,kw.	1663
21	or/15-20	48427
22	((opportunis* or incident* or uninten* or unsuspect* or fortuit*) adj2 (detect* or present* or find* or identif*)).ti,ab,kw.	26965
23	diagnostic imaging/	47424
24	radiography/	330259
25	absorptiometry, photon/	27143
26	exp magnetic resonance imaging/	566109
27	exp tomography, emission-computed/	140765
28	x rays/	32659

29	(radiogra* or ct or (comput* adj4 tomogra*) or absorptiometry or dexa or dxa or magnetic resonance or mri or mrs or nmr* or x ray or xray or x-ray).tw.	1975673
30	or/23-29	2514715
31	Algorithm*.ti,kf.	77668
32	(algorithm* adj2 (learn* or automate* or detect* or predict* or treatment* or therap* or radiolog* or AI or DL or ML or data or dataset* or base* or classif*)).ab.	109235
33	Artificial Intelligen*.ti,ab,kf.	64598
34	AI.ti,kf.	15782
35	(machine adj2 learn*).ti,ab,kf.	135958
36	machinelearn*.ti,ab,kf.	26
37	(deep adj2 learn*).ti,ab,kf.	78881
38	deeplearn*.ti,ab,kf.	30
39	neural network*.ti,ab,kf.	123095
40	(convolutional adj1 network*).ti,ab,kf.	4154
41	automate*.ti.	51044
42	(automate* adj3 (system* or score* or software* or analysis* or analyse* or risk* or evaluat* or tool* or detect* or process*)).ab,kf.	45760
43	(vector machine* or svm*).ti,ab,kf.	36278
44	radiomic*.ti,ab,kf.	13088
45	((supervised or unsupervised) adj3 (classifie* or predict*)).ti,ab,kf.	1449
46	or/31-45	515144
47	(editorial or letter).pt.	1992457
48	animals/ not (animals/ and humans/)	5269075
49	(clinical conference or editorial or letter or news or comment or historical article).pt.	2857576
50	or/47-49	8030462
51	14 not 50	67
52	limit 51 to english language	66
53	(21 and 22 and 30 and 46) not 50	14
54	limit 53 to english language	14
55	54 not 52	12
56	(21 and 30 and 46) not 50	442
57	limit 56 to english language	432
58	57 not (52 or 54)	409

CENTRAL via Wiley Search

IDSearch
#1 (("Annalise Enterprise" or "Annalise.AI")):ti,ab,kw
#2 (Annalise and CXR)
#3 (boneview* and gleamer)
#4 ("BriefCase-Triage" or "Aidoc Medical")
#5 (briefcase and "Aidoc Medical")
#6 ("CINA-VCF quantix" or "Avicenna.AI")
#7 (c-spine or briefcase) and (ai or artificial intelligence)
#8 aidoc and (spinal or spine or vertebr* or bone or osteop* or fractur*)
#9 (Healthost or Healthvcf or "nanox.ai" or "nanox ai")
#10 "Zebra Medical" and (spinal or spine or vertebr* or bone or osteop* or fractur*)
#11 ("IB lab" or "ImageBiopsy Lab") and (flamingo or bonebot)
#12 "TechCare Spine (Milvue)"
#13 (Techcare and Milvue)
#14 {OR #1-#13}
#15 MeSH descriptor: [Fractures, Compression] this term only
#16 MeSH descriptor: [Osteoporotic Fractures] this term only
#17 MeSH descriptor: [Spinal Fractures] this term only
#18 ((compress* or spine or spinal) NEAR/2 fracture*):ti,ab,kw
#19 ((vertebra* or osteoporo*) NEAR/2 (fragil* or fracture*)):ti,ab,kw
#20 (VFF or VFFS or VCF):ti,ab,kw
#21 {OR #15-#20}
#22 ((opportunis* or incident* or uninten* or unsuspect* or fortuit*) NEAR/2 (detect* or present* or find* or identif*)):ti,ab,kw
#23 MeSH descriptor: [Diagnostic Imaging] this term only
#24 MeSH descriptor: [Radiography] this term only
#25 MeSH descriptor: [Absorptiometry, Photon] this term only
#26 MeSH descriptor: [Magnetic Resonance Imaging] explode all trees
#27 MeSH descriptor: [Tomography, Emission-Computed] explode all trees
#28 MeSH descriptor: [X-Rays] this term only
#29 (comput* NEAR/4 tomogra*):ti,ab,kw
#30 (radiogra* or ct or absorptiometry or dexa or dxa or "magnetic resonance" or mri or mrs or nmr* or "x ray" or xray or "x-ray"):ti,ab,kw
#31 (Royal Osteoporosis Society, -#29)
#32 Algorithm*:ti,kw
#33 (algorithm* NEAR/2 (learn* or automate* or detect* or predict* or treatment* or therap* or radiolog* or AI or DL or ML or data or dataset* or base* or classif*)):ab
#34 (Artificial NEXT Intelligen*):ti,ab,kw
#35 AI:ti,kw
#36 (machine NEXT/2 learn*):ti,ab,kw
#37 machinelearn*:ti,ab,kw
#38 (deep NEXT/2 learn*):ti,ab,kw
#39 deeplearn*:ti,ab,kw
#40 (neural NEXT network*):ti,ab,kw
#41 (convolutional NEXT network*):ti,ab,kw
#42 automate*:ti
#43 (automate* NEAR/3 (system* or score* or software* or analysis* or analyse* or risk* or evaluat* or tool* or detect* or process*)):ab,kw
#44 ((vector NEXT machine*) or svm*):ti,ab,kw
#45 radiomic*:ti,ab,kw
#46 ((supervised or unsupervised) NEAR/3 (classifie* or predict*)):ti,ab,kw
#47 {OR #32-#46}
#48 #21 AND #22 AND #31 AND #47

#49 #21 AND #31 AND #47 #50 #21 and #22

Line 50: No results from line 48, this line checks for opportunistic fracture detection

INAHTA

Line	Query	Hits
32	#31 AND #17 AND #9	140
31	#30 OR #29 OR #28 OR #27 OR #26 OR #25 OR #24 OR #23 OR #22 OR #21 OR #20 OR #19 OR #18	201
30	((supervised or unsupervised) and (classifie* or predict*))	4
29	radiomic*	0
28	vector machine* or svm*	2
27	((automate* and (system* or score* or software* or analysis* or analyse* or risk* or evaluat* or tool* or detect* or process*))	85
26	(automate*)[title]	41
25	convolutional network*	0
24	neural network*	5
23	deep learn* or deeplearn	7
22	deep learn* or deeplearn	7
21	machine learn* or machinelearn*	9
20	(Artificial Intelligen*)	28
19	((algorithm* and (learn* or automate* or detect* or predict* or treatment* or therap* or radiolog* or AI or DL or ML or data or dataset* or base* or classif*))	186
18	(Algorithm*)	201
17	#16 OR #15 OR #14 OR #13 OR #12 OR #11 OR #10	1241
16	((radiogra* or ct or absorptiometry or dexa or dxa or magnetic resonance or mri or mrs or nmr* or x ray or xray or x-ray)) OR ((comput* and tomogra*))	917
15	"X-Rays"[mh]	44
14	"Tomography, Emission-Computed"[mhe]	251
13	"Magnetic Resonance Imaging"[mhe]	293
12	"Absorptiometry, Photon"[mh]	21
11	"Radiography"[mh]	33
10	"Diagnostic Imaging"[mh]	141
9	#8 OR #7 OR #6 OR #5 OR #4 OR #3	190
8	(VFF or VFFS or VCF)	2
7	(((vertebra* or osteoporo*) and (fragil* or fracture*)))	165
6	(((compress* or spine or spinal) and fracture*))	79

5	"Spinal Fractures"[mh]	60
4	"Osteoporotic Fractures"[mh]	30
3	"Fractures, Compression"[mh]	26
2	(("Annalise Enterprise" or "Annalise.AI" OR boneview* OR gleamer OR "BriefCase-Triage" or "Aidoc Medical" OR "CINA-VCF quantix" or "Avicenna.AI" OR c-spine OR Healthost or Healthvcf or "nanox.ai" or "nanox ai" OR "Zebra Medical" OR "IB lab" or "ImageBiopsy Lab" OR flamingo or bonebot)) OR ("TechCare Spine OR Milvue))	0
1	(("Annalise Enterprise" or "Annalise.AI")) OR ((Annalise and CXR)) OR ((boneview* and gleamer)) OR (("BriefCase-Triage" or "Aidoc Medical")) OR ((briefcase and "Aidoc Medical")) OR (("CINA-VCF quantix" or "Avicenna.AI")) OR ((c-spine or briefcase) and (ai or artificial intelligence)) OR (aidoc and (spinal or spine or vertebr* or bone or osteop* or fractur*)) OR ((Healthost or Healthvcf or "nanox.ai" or "nanox ai")) OR ("Zebra Medical" and (spinal or spine or vertebr* or bone or osteop* or fractur*)) OR ((("IB lab" or "ImageBiopsy Lab") and (flamingo or bonebot)) OR ("TechCare Spine (Milvue)") OR ((Techcare and Milvue))	0

Google Scholar
Search:

"Annalise Enterprise" OR "Annalise Container" OR boneview* OR "BriefCase-Triage" OR "CINA-VCF" OR Healthost OR Healthvcf OR "TechCare Spine” OR “Milvue Suite”
("IB lab" OR "ImageBiopsy Lab") +flamingo
("IB lab" OR "ImageBiopsy Lab") + bonebot
(opportunis* or incident* or uninten* or unsuspect* or fortuit*) AND (detect* OR present* OR find* OR identif*) AND (artificial intelligence OR AI OR machine learning OR deep learning) AND (spine OR spinal OR vertebra*) AND +fractur* - **gave no results**
(opportunistic OR incidental OR unintended OR unintentional OR unsuspected OR fortuitous) AND (detect* OR present* OR find* OR identif*) AND (artificial intelligence OR AI OR machine learning OR deep learning) AND (spine OR spinal OR vertebra*) AND +fractur* **34 results**

Targeted economics search

Database:

Embase <1974 to 2025 March 05>

#	Query	Results from 6 Mar 2025
1	("Annalise Enterprise" or "Annalise.AI").af.	21
2	(Annalise and CXR).af.	10

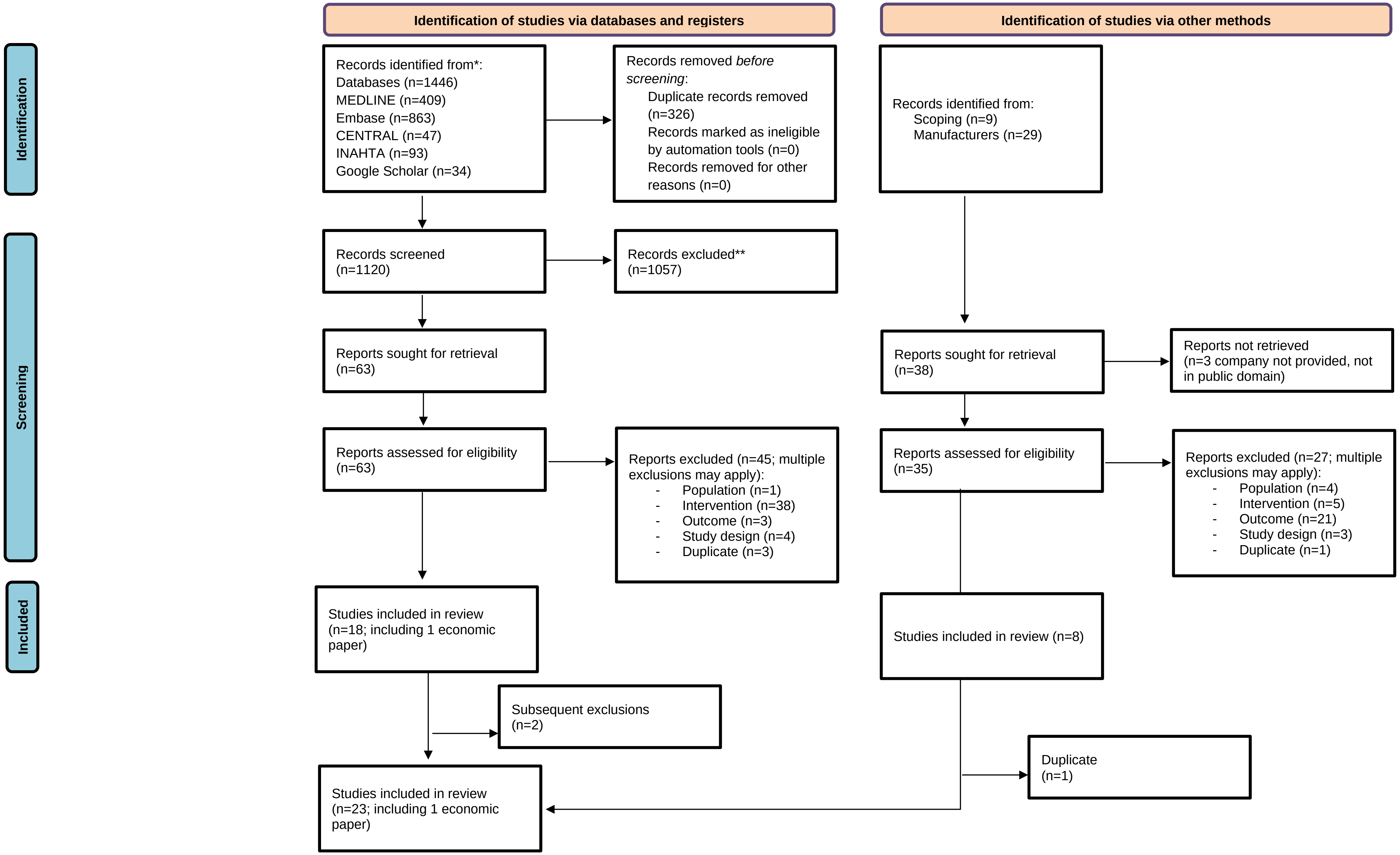
3	(boneview* and gleamer).af.	14
4	("BriefCase-Triage" or "Aidoc Medical").af.	8
5	(briefcase and "Aidoc Medical").af.	3
6	("CINA-VCF quantix" or "Avicenna.AI").af.	13
7	(c-spine or briefcase).af. and (ai or artificial intelligence).tw.	12
8	aidoc.af. and (spinal or spine or vertebr* or bone or osteop* or fractur*).tw.	7
9	(Healthost or Healthvcf or "nanox.ai" or nanox).af.	50
10	"Zebra Medical".af. and (spinal or spine or vertebr* or bone or osteop* or fractur*).tw.	14
11	((IB lab or ImageBiopsy Lab) and (flamingo or bonebot)).af.	0
12	"TechCare Spine (Milvue)".af.	0
13	(Techcare and Milvue).af.	0
14	or/1-13	127
15	fractures, compression/	9,281
16	osteoporotic fractures/	26,335
17	spinal fractures/	35,964
18	((compress* or spine or spinal) adj2 fracture*).ti,ab,kw.	18,048
19	((vertebra* or osteoporo*) adj2 (fragil* or fracture*)).ti,ab,kw.	47,520
20	(VFF or VFFS or VCF).ti,ab,kw.	2,611
21	or/15-20	82,394
22	diagnostic imaging/	270,364
23	radiography/	206,481
24	absorptiometry, photon/	4,826
25	exp magnetic resonance imaging/	1,377,587
26	exp tomography, emission-computed/	354,260
27	x rays/	92,063
28	(radiogra* or ct or (comput* adj4 tomogra*) or absorptiometry or dexa or dxa or magnetic resonance or mri or mrs or nmr* or x ray or xray or x-ray).tw.	2,744,548
29	or/22-28	3,670,486
30	Economics/	245,512
31	Cost/	65,372
32	exp Health Economics/	1,116,298
33	Budget/	35,874
34	budget*.ti,ab,kf.	52,316
35	(economic* or cost or costs or costly or costing or price or prices or pricing or pharmacoeconomic* or pharmaco-economic* or expenditure or expenditures or expense or expenses or financial or finance or finances or financed).ti,kf.	376,478
36	(economic* or cost or costs or costly or costing or price or prices or pricing or pharmacoeconomic* or pharmaco-economic* or expenditure or expenditures or expense or expenses or financial or finance or finances or financed).ab. /freq=2	599,150
37	(cost* adj2 (effective* or utilit* or benefit* or minimi* or analy* or outcome or outcomes)).ab,kf.	329,339
38	(value adj2 (money or monetary)).ti,ab,kf.	4,488

39	Statistical Model/	179,593
40	exp economic model/	4,784
41	economic model*.ab,kf.	7,051
42	Probability/	163,182
43	markov.ti,ab,kf.	42,398
44	monte carlo method/	56,559
45	monte carlo.ti,ab,kf.	69,649
46	Decision Theory/	1,903
47	Decision Tree/	27,718
48	(decision* adj2 (tree* or analy* or model*)).ti,ab,kf.	63,394
49	or/30-48	2,197,833
50	14 and 49	11
51	21 and 29 and 49	1,734
52	51 not (editorial or letter).pt.	1,708
53	limit 52 to english language	1,644
54	((long-term or long term or longstand* or last* or exten* or prolong* or endur*) adj impact*).ti,ab.	18,901
55	(treatment* adj2 cost*).ti,ab.	42,170
56	21 and 29 and 49 and (54 or 55)	35
57	56 not (editorial or letter).pt.	35
58	limit 57 to english language	34

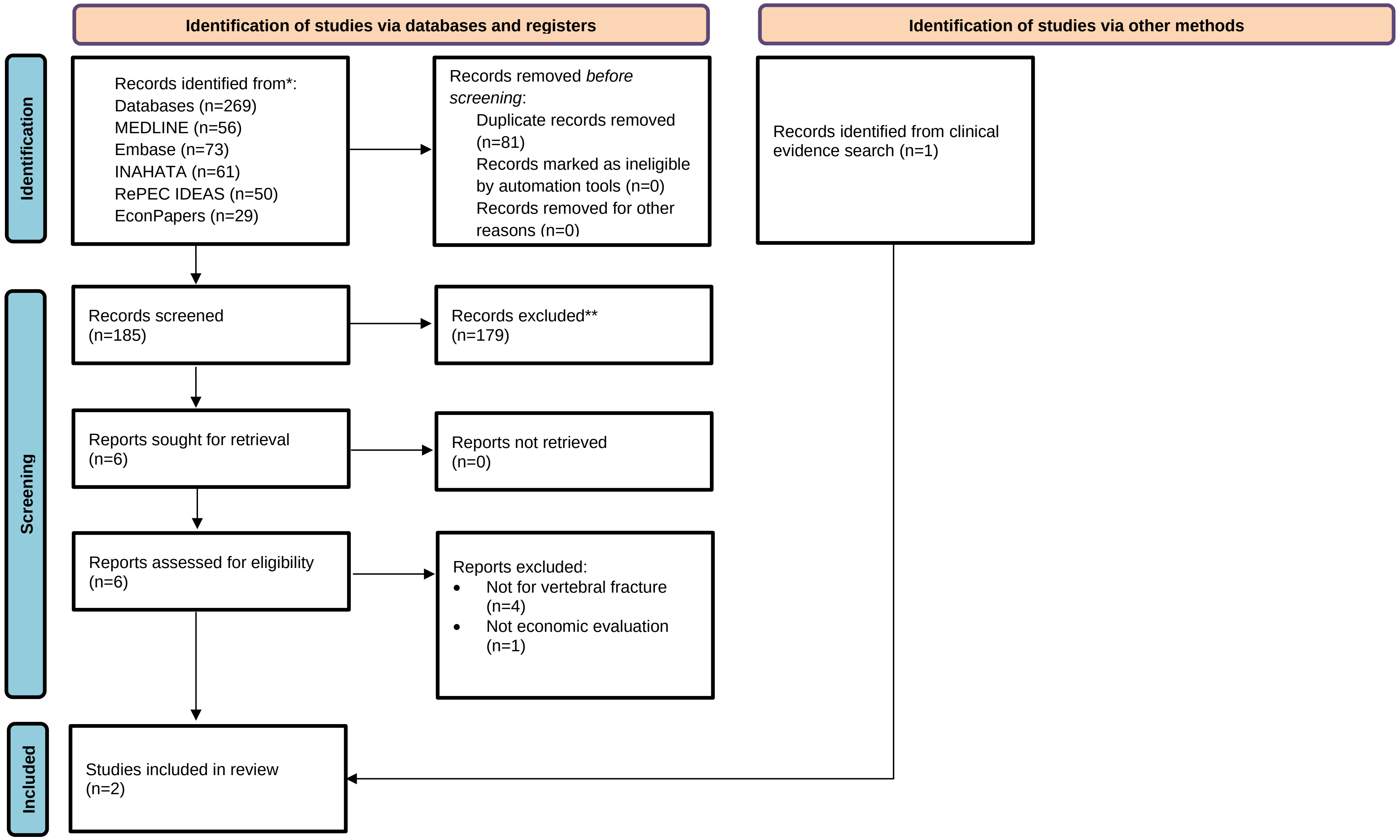
Google Scholar

((spine OR spinal OR vertebra*) AND +fracture*) AND (artificial intelligence OR AI OR machine learning OR deep learning) AND ("statistical model" OR "economic model" OR "cost-effectiveness" OR "cost effectiveness" OR utilit*) AND ("long term impact" OR "long term effect" OR "treatment cost" OR "cost of treatment"))

Appendix A2: PRISMA diagram: clinical



Appendix A3: PRISMA diagram: economics



Appendix A4: Study characteristics of included clinical evidence

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
1.	Annalise.AI [AiC] [redacted]	[redacted] GREEN [redacted] *Reference standard:* [redacted] AMBER	[redacted] AMBER [redacted]	[redacted] AMBER	[redacted] At stakeholder consultation the company stated that the population used to train and validate the AI for this study was reported in Seah et al. 2021.
2.	Avicenna.AI [AiC]	[redacted]	[redacted]	[redacted]	[redacted]

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
	Abstract [redacted] [redacted] [redacted]	[redacted] [redacted] GREEN [redacted] AMBER	[redacted] [redacted] GREEN [redacted] [redacted]	[redacted] GREEN	[redacted]
3.	Bendtsen & Hitz (Calc Tissue Int, 2024; 468-479) *Funding:* Amgen. Data collection, analysis and writing done without involvement of Amgen. *Declaration of interests:* Multiple authors with Amgen, Gedeon Richter Nordic, Novo Nordic, Novo Nordisk, UCB, UCB Nordic	Prospective diagnostic accuracy study *Intervention:* HealthVCF, v5.1.1 (n=10,012 scans); configured for the highest specificity GREEN *Comparator*: primary radiology report (produced by general radiologists who were not specialists in musculoskeletal imaging) *Reference standard:* Specialised	*Inclusion:* inpatients and outpatients referred for CT of thorax and abdomen for all indications except fractures, aged 50 years or older. *Exclusion:* Individuals without a social security number (tourist) or from other geographical regions, cases involving CT of thorax and abdomen as part of PET scans, where assessment of vertebral column not possible as judged by radiographer / radiologist, CT of thorax and abdomen for fracture identification, cases of incomplete exams and wrong region of interest	Diagnostic accuracy (including false positives, false negatives) in detection of vertebral fractures. Outcomes from sub-cohort who underwent 6 months follow-up: vital status, referral for DEXA scan, speciality of referring physician, initiation or adjustment of osteoporosis treatment, choice of treatment, and registration of osteoporosis diagnosis. Time taken for radiographer to analyse output of HealthVCF. AMBER	<u>Training</u>: NR <u>Validation</u>: Subset (n=1,000 scans) randomly selected. Unclear of overlap with follow-up cohort. <u>Sample size</u>: recruitment stopped when 10,000 images acquired. Authors reported that no performance error analysis was conducted within the study (out of scope).

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
		radiographers with 10-28 years experience (n=5), if inconclusive then senior radiologist consulted **GREEN**	Validation cohort used a 1000 random sample. **GREEN** *Image acquisition:* between 23 January 2022 and 25 October 2022. Image deidentification not reported *Setting:* Denmark (N=1); university hospital *Follow-up:* only true positive cases (identified by HealthVCF and reference standard) were followed for 6 months (n=538)		
4.	Chappell [2024] Poster ADOPT study *Funding:* NR *Declaration of interests:* NR	Retrospective diagnostic accuracy study (n=2,000 CT scans) *Intervention:* HealthVCF **GREEN** *Comparator:* original radiology report **GREEN** *Reference standard:* clinician with local radiologist adjudication **AMBER**	*Inclusion:* CT involving spine *Exclusion*: NR **AMBER** *Image acquisition:* All sagittal CT scans involving spine from 2017 (end date NR). Image deidentification not reported. Included abdomen and pelvic CT, and pulmonary angiogram CT. *Setting:* UK (N=4 sites), no further detail	Diagnostic accuracy (including false positives, false negatives) in detection of vertebral fragility fractures **GREEN**	Poster only, lack of detailed methodology. However, could represent the 'shadow' period of 500 consecutive scans per site described in the abstract by (Connor et al., 2024). Training: NR Validation: NR

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
5.	Connacher (Osteoporosis Int, 2019; S428) Abstract *Funding:* NR *Declaration of interests:* NR	Prospective diagnostic accuracy study (n=4,623 scans) *Intervention:* Zebra **AMBER** *Reference standard:* FLS nurses (n=2) **AMBER**	*Inclusion:* All thoracic or lumbar spine CT *Exclusion*: NR **AMBER** *Image acquisition:* All CT scans involving thoracic or lumbar spine over 4 weeks (dates NR) Image deidentification not reported *Setting:* UK (N=1); tertiary hospital	Diagnostic accuracy (including false positives, false negatives) in detection of vertebral fractures **GREEN**	Abstract only, lack of detailed methodology. No reference standard reported. <u>Training</u>: NR <u>Validation</u>: NR Limitation: only subset of 55 consecutive scans identified by Zebra as not having a vertebral fracture were checked. Subgroup analysis by age group (50 to 74 years, and over 75 years).
6.	Connor (J Bone Min Res, 2024; 296-297) Abstract ADOPT study *Funding: NR* *Declaration of interests:* One author with 3D-Shaper Medical	Prospective diagnostic accuracy study (n=36,620) *Intervention:* HealthVCF. One centre chose high specificity mode (to minimise false positives), while the other sites chose a balanced mode with higher sensitivity). **AMBER**	*Inclusion:* NR *Exclusion:* NR *Image acquisition:* CT over three months (dates not reported). Image deidentification not reported *Setting:* UK (N=4); FLS public hospitals	Number of analysed scans, number of flagged scans by AI, number of clinically reviewed scans, number of clinically positive scans. Issues raised by improvement teams. **AMBER**	Abstract only, lack of detailed methodology. Reference standard unclear. <u>Training</u>: NR <u>Validation</u>: NR Different AI technology settings and different reference standard across centres.

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
		Comparator: 2022 FLS database submitted spine records. *Reference standard:* 1 centre had synchronous augmented live reporting by radiologists, 3 centres forwarded eligible scans for asynchronous local clinical confirmation and FLS management (staff roles not defined) **AMBER**			
7.	Dai (J Comput Assist Tomogr, 2025) *Funding:* NR *Declaration of interests:* multiple authors with Avicenna, Bayer, Canon Medical GE, Novocure, Siemens (5 authors are paid employees of Avicenna)	Retrospective diagnostic accuracy study (n=474 scans); dataset source from 2 teleradiology networks. Purpose of the study was to validate tool across multiple sites and multiple vendors of CT scanner. *Intervention:* CINA-VCF 1.0 (Avicenna.AI) **GREEN** *Comparator:* Clinical radiology report	*Inclusion:* Patients aged 50 years or older, CT (non-enhanced or contrast-enhanced) for indications other than VCF, including chest or abdomen field of view, axial or sagittal acquisition with a homogeneous slice interval without a gap between successive slices, slice thickness of 3 mm or less, and soft and standard tissue reconstruction kernel. *Exclusion:* CT scans performed on patients under 50 years of age or not following the	Diagnostic accuracy (including false positives, false negatives) in detection of vertebral compression fractures **GREEN**	<u>Training</u>: 12,402 vertebrae, augmentation was used to enhance data sets and to improve model optimisation and generalisability. 5-fold approach (80% training, 20% testing) to mitigate training bias. <u>Validation</u>: Purpose of study. <u>Sample size</u>: minimum number to achieve AUROC of 95% with 5% significance and 80%

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
		(available in 317/474 cases) *Reference standard:* board certified neuroradiologists (N=2) with consensus for discordant cases determined by a third board-certified neuroradiologist **GREEN**	mandatory acquisition parameters, scans with the following parameters: only the cervical spine was visible, at least 3 consecutive vertebrae in the T1-L5 portion of the spine were not present and not completely visible, uninterpretable poor-quality images (significant motion, streak, or other artifacts impeding CT interpretation), and redundant cases (not further defined). **AMBER** *Image acquisition:* Validation data set of chest and abdominal CT was consecutively acquired across 2 different teleradiology networks: • between April and November 2021 (n=317), • between November 2022 and January 2023 (n=157). 38 different scanner models. Images were anonymised (except scan date).		power was 133 negative and 133 positive cases.

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
			Setting: US and France (N=NR); multiple teleradiology networks		
8.	Frias (RSNA, 2023; W5B-SPCH-4) Abstract *Funding:* NR *Declaration of interests:* Author reports nothing to disclose.	Retrospective diagnostic accuracy study (n=3,760 patients, 3,760 X-rays) *Intervention:* Annalise.AI **GREEN** *Comparator:* One of five participating physicians *Reference standard:* Discrepancies reviewed by one of two thoracic radiologists (blinded – not further defined) **AMBER**	*Inclusion:* Consecutive adults (aged over 18 years). *Exclusion:* NR **GREEN** *Image acquisition:* Chest X-rays (PA, AP or LAT views), date of acquired images and deidentification NR *Setting:* country NR *(*N=2); community and quaternary care hospitals	Diagnostic accuracy (including false positives, false negatives in the original radiology report) in detection of compression fractures **AMBER**	Abstract only, lack of detailed methodology. Unclear whether the identified compression fractures were also identified by the original radiographers as well (unable to quantify missed VFFs). At stakeholder consultation the company stated that the population used to train and validate the AI for this study was reported in Seah et al. 2021.
9.	Ghatak (J Am Coll Radiol, 2024; 220-229) *Funding:* funded by Annalise-AI *Declaration of interests:* all authors are employed by Annalise-AI	Retrospective diagnostic accuracy study (n=596 chest X-rays, 596 patients) *Intervention:* Annalise Enterprise CXR Triage Trauma (v2.2.0); the study reports that this uses the same AI as the Analise Enterprise (CXR module) device.	*Inclusion:* patients aged 18 years or older, chest X-rays performed at a hospital, including inpatient and outpatient, no limitations on the original clinical indication, only the first X-ray from a given patient was included. *Exclusion:* did not include a chest X-ray, or did not include	Diagnostic accuracy (including false positives, false negatives) in detection of vertebral compression fractures **GREEN**	Clinical Experts advised that it would be rare in the NHS to obtain a frontal and lateral projection (see Appendix D3). Unclear on generalisability of this evidence. Training: more than 750,000 chest X-rays labelled by three radiologists, identifying

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
		Assume balanced configuration applied (although not explicitly stated) **AMBER** *Comparator:* original radiology report (natural language processing of the initial report; ICD10 code presence checked for reporting of VCF) *Reference standard*: Manual review by board certified radiologists with training in thoracic radiology (N=3) agreement of 2 equalled consensus with 3rd only for disagreement adjudication **GREEN**	both a frontal (AP or PA) and lateral (LAT) projection. **AMBER** *Image acquisition:* consecutive radiology reports between 1 November 2015 and 31 October 2021. Radiographs required both a frontal (AP or PA) and lateral (LAT) projection. At least 8 scanner manufacturers were included. All radiographs were de-identified. *Setting:* US (N=4);		more than 100 different radiological findings (not specific to VFF). <u>Validation</u>: 596 chest X-rays (272 positive and 324 negative ground truth interpretations). <u>Sample size</u>: determined to ensure that the lower 95% CI for AUC was greater than 0.95. Subgroup analysis: sex, age, race, ethnicity, and X-ray machine manufacturer At stakeholder consultation the company stated that the population used to train and validate the AI for this study was reported in <u>Seah et al. 2021</u>.
10.	Guenoun (Clin Radiol, 2025; 106831) *Funding:* None received. *Declaration of interests:* Multiple authors	Retrospective diagnostic accuracy study (n=100 patients, 1700 vertebrae).	*Inclusion:* Consecutive chest-abdominal-pelvis CT scans in patients aged 50 years or older, acquired for various indications other than suspicion of VCF. All eligible CT scans were transferred	Labelling agreement (label each vertebra from T1-L5, quantification of the percentage of VHL for each visible vertebra from T1-L5 following Genant's methodology). Detection of	<u>Training</u>: on 12,402 vertebrae from 1,353 CT cases (from more than 64 scanner models from Siemens, Philips, GE, Canon) acquired in US

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
	employed by Avicenna.AI	*Intervention:* CINA-VCF Quantix v0.60 by Avicenna.AI. **GREEN** *Reference standard:* Two radiologists (with 19 and 7 years of experience) analysed same cases separately, blinded to the other, then VHL considered together and agreed by consensus). **AMBER**	from the hospital PACS and the first 100 cases included (first 50 potentially positive VCF cases and first 50 potentially negative VCF cases according to clinical reports). 20 cases were randomly selected for vertebral heigh loss measurements. *Exclusion:* significant artefacts, presence of material in the vertebrae. **GREEN** *Image acquisition:* Chest-abdominal-pelvis CT scans acquired between January 2019 and August 2020. CT scans with and without contrast were accepted. Single scanner. All patients scanned on 64-row scanner in helical model with 0.625mm slice collimation, tube voltage of 120kVp, and image matrix 512x512.Soft kernel used and images reconstructed at 1.25 mm slice thickness. Image deidentification not reported *Setting:* France (N=1 centre)	cases with grade 2 or 3 VCF (per patient). Measurement of the mean [redacted] vertebral bone attenuation within the first grade 0 or 1 vertebrae among L1-L4 starting with L1 (per patient). Diagnostic accuracy in detecting grade 2 or 3 vertebral compression fracture (sensitivity and specificity, accuracy) including correlation between AI and reference standard and differences between radiologists. **AMBER**	and French centres during 2021 and 2022. <u>Validation</u>: separate pilot dataset of 1,994 vertebrae from 152 cases. Sample size determined assuming 5% significance level, power of 80% for each metric: • Vertebral labelling expecting 95% accuracy with 95%CI not wider than +/- 5% (n=239 required vertebrae) • [redacted] measurement mean difference of -0.35, SD of 4.5 and max allowed difference of 10.5 (n=274 required vertebrae). • VCF screening for expected sensitivity and specificity of 90%, prevalence of 52%, confidence interval not wider

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
					than +/-10% (n=73 required cases). • Mean HU, correlation coefficient of 0.8 (n=9 required cases) Authors acknowledge that "*the algorithm may exhibit a different performance when applied to images from other scanners due to variations in image quality, noise characteristics, and scanning protocols.*" However, noted that the algorithm was trained on more than 64 scanner models from 4 different manufacturers.
11.	IB Lab Internal Report [CiC; 2023] [redacted] *Funding:* Provided by company *Declaration of interests:* Provided by company	[redacted] [redacted] GREEN [redacted]	[redacted] [redacted] AMBER	[redacted] GREEN	[redacted]

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
		[redacted] **AMBER**	[redacted]		
12.	Jones (BMJ Open, 2021; e052902) *Funding:* Annalise.AI *Declaration of interests:* Multiple authors employed, seconded or clinical consultant for Annalise.AI	Pilot (n=11 radiologists for 6 weeks) and then post-study survey (n=10 radiologists) *Intervention:* Annalise Enterprise CXR v.1.2 (modified version of commercially available tool was used; incorporated changes into user interface to provide feedback on the model recommendations). No clinical inputs (previous imaging, referral information, or patient demographics) included in AI analysis. *Comparator:* consultant radiologist (reporting minimum of 2000 chest X-rays per year,	*Inclusion:* All consecutive chest radiographs reported by radiologists (originating from inpatient, outpatient, emergency settings). *Exclusion:* Radiographs from patients aged less than 16 years. *Image acquisition:* Chest X-ray (with at least one frontal chest radiograph) acquired over 6-week period (between November and December 2020). Data from all sources were de-identified for analysis. *Setting:* Australia (large radiology network, N=106 sites contributing cases)	Number of spine wedge fractures missed by the AI (classed as non-critical findings). Radiologist reporting of subjective improved accuracy, reporting times and satisfaction with reporting times, attitude towards AI. **AMBER**	<u>Training</u>: 821,681 de-identified chest X-rays from 284,649 patients from inpatient, outpatient and emergency settings from Australia, Europe and US. Independent triple labelling of all images by 3 of 120 consultant radiologist. <u>Validation</u>: Validated for 124 clinical findings in multi-reader multi-case study (Seah et al. 2021); 34 of which were deemed priority findings. Ground truth labels were determined by 3 of seven independent subspecialty thoracic radiologists. The authors acknowledge the lack of review/adjudication of

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
		working in public/private hospitals and community clinical settings, with varying levels of experience). Radiologist included access to clinical information, referral and available patient history (as per standard workflow) **AMBER**			discrepancies between radiologist and AI, and that "*radiologists remained responsible for image interpretation and formulation of the report*". The EAG note that the abstract by Karunasena (J Med Image Radia Oncol, 2022; 95) also reports on this study.
13.	Kolanu (Journal of Bone and Mineral Research, 2020; 2307-2312) *Funding:* NR *Declaration of interests:* Multiple authors reported fees from Amgen. One author reported additional fees from Actavis and Bayer.	Retrospective diagnostic accuracy study (n=2,357) *Intervention:* Zebra (EAG notes that Nanox AI acquired Zebra Medical Vision in 2021; Company have confirmed evidence is generalisable to HealthVCF) **AMBER** *Comparator:* Original radiology report (standard care) *Reference standard*: One radiologist (2 radiologists and 1	*Inclusion:* CT scans performed as part of routine clinical care involving thorax or abdomen, patients aged over 50 years. *Exclusion:* NR **GREEN** *Image acquisition:* Abdomen or thorax CT performed, between January to May 2019 to achieve sample size of 1,751. Image deidentification not reported *Setting:* Australia (N=1); tertiary referral hospital	Failure to process image, diagnostic accuracy in detecting VCF. **GREEN**	Sample size: recruitment of images extended from 3 to 5 months to include planned sample size (target or 1,751) to achieve estimates of sensitivity and specificity, at 95% confidence interval with precision of 0.05, with estimated VCF prevalence of 20%.

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
		senior nuclear medicine physician used for adjudication) **AMBER**			
14.	Nicolaes (Osteoporosis International, 2024; 143-152) *Funding:* National Natural Science Foundation of China (grant no. 81971617), Beijing Hospitals Authority Youth Programme (QML20200402), the National Key R&D Program of China (grant no.2020YFC2004902), Beijing Hospitals Authority Clinical Medicine Development of Special Funding Support (ZYLX202107) and UCB Pharma/Amgen Inc. *Declaration of interests:* none declared. The manuscript was shared with UCB Pharma/Amgen Inc. for courtesy review prior to	Retrospective diagnostic accuracy study; external validation (5,195 CT scans); which included abdominal and chest CT from a prior community health screening study, subset of thoracic/lumbar spine CT from a hospital and random sample of abdominal CT from an epidemiology study. *Intervention:* BoneBot (n=4,810); confirmed by company. **AMBER** *Reference standard*: one sub-specialist radiologist with >20 years experience identified and graded fractured vertebra from lateral CT scout view, study cohort split into 3 groups each to a sub-	For images acquired from prior community screening *Inclusion:* Chest CT images where vertebrae could be identified by the naked eye. *Exclusion:* low resolution images. For images acquired from hospital in China *Inclusion:* aged 50 years or older at time of CT, with osteoporotic vertebral fracture or wedge deformity mentioned in the radiology report. CT with maximum slice thickness of 2mm with acceptable quality. *Exclusion:* <50 years old, with traumatic fracture, pathological fracture, or metal internal fixation of spine. For images acquired from previous epidemiology study *Inclusion:* Random sample *Exclusion:* NR **AMBER** *Image acquisition:*	Diagnostic accuracy (AUROC, Cohen's kappa, accuracy, sensitivity, specificity, PPV, NPV) reported patient-level and vertebral level. Time taken to run algorithm. Failure of algorithm to process image. **GREEN**	Training: VF detection model previously trained on private dataset of 666 CT scans (abdominal and chest CT, aged 50 years or older, maximum slice thickness 3mm; 55% VF patient-level prevalence), in Brussels. Validation: Purpose of study is external validation study of previously trained Convolutional Neural Network (CNN) model. Sample size (n=624 subjects) determined to measure sensitivity of 80% or more, assuming prevalence of 15%.

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
	submission. The other funders were not involved in the study.	specialist with <5 years experience, disagreement resolved by consensus with all 4 readers. **GREEN**	• Prior community lung cancer screening: 2,036 chest CT acquired between August 2013 and March 2014. • Epidemiology CASH study: 2,419 abdominal CT randomly sampled between June 2012 and March 2017. • Hospital in China: 740 thoracic/lumbar spine CT between January 2013 and December 2018. CTs from 16 scanners (3 manufacturers). All images anonymised prior to processing. *Setting:* Images obtained from China (N=8 institutions)		
15.	Nicolaes (J Bone Miner Res. 2023; 1856-1866) Additional details on eligibility criteria, methods, sample size calculations and baseline characteristics reported in Skjødt et al. (JMBR Plus, 2023; e10736)	Retrospective diagnostic accuracy; (n=2,000 CT scans); external validation of developed algorithm. *Intervention:* Not named (however supplied by IB LAB who confirmed AI model was previous version of IB Lab FLAMINGO)	*Inclusion:* Consecutive individuals aged 50 years or older at the time of the scan. Data obtained in the CT scan was linked (at patient level) to the Danish national registers, where information on demographics, medical history, pharmaceutical drugs were obtained.	VF detection (patient level and vertebral level), diagnostic accuracy (AUROC, Cohen's kappa, accuracy, sensitivity, specificity, PPV, NPV) **GREEN**	Training: 1011 routine CT scans of abdomen (potentially including pelvis) and chest from a non-interventional, prospective proof-of-concept study (Belgium; between January and August 2019) used to develop the AI for detection of VFFs.

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
	Funding: UCB Pharma and Amgen Inc. (EAG notes that the AI algorithm, BoneBot, was subsequently licenced by UCB to IB Lab). *Declaration of interests:* Multiple authors employees of UCB Pharma	AMBER *Reference standard:* Firstly, medical doctor triaged the scans (certain VF, potential VF, no VF). Secondly, blinded vertebral readings (of certain VF, potential VF and 5% subset of scans categorised as no VF – however the radiologist was blinded to the category) produced by a highly experience radiologist using Genant classification. AMBER	*Exclusion:* CT scans identified by the machine learning algorithm as not readable or not having registry data were excluded. GREEN *Image acquisition:* Abdominal or chest CT acquired from 01 January 2010 onward (no end date specified). Only the first eligible CT scan was included. Validation data was pseudonymized after extraction and subsequently lined to a national register for outcomes *Setting:* Denmark (N=1 hospital)		Sample size of validation determine assuming 15% prevalence of VFs in sample, sufficient to measure a sensitivity of 80%, requiring lower 95%CI to be greater than 70%. Subgroup analysis by sex, age and location (thoracic or lumbar vertebrae) were reported.
16.	Oppenheimer (Skeletal Radiol, 2024; 1563) *Funding:* Open access funding from Projekt DEAL. *Declaration of interests:* Two authors with conflicts (one author receiving research grants from 119 companies; however those grants had no role	Retrospective diagnostic accuracy study (n=512 X-rays) *Intervention:* BoneView (v1.2.0) GREEN *Reference standard*: Consensus of 2 radiologists with experience in musculoskeletal radiology.	*Inclusion:* Thoracic spine or lumbar X-rays, at least sagittal image acquired with inquiry for fracture. *Exclusion:* Spine X-rays acquired for other inquiries such as degenerative changes or pre- and post-surgery imaging. Cervical spine X-rays (not supported by software). Exams not including a sagittal image. AMBER	Failure of AI to process image, diagnostic accuracy. GREEN	<u>Training</u>: 60,000 images used in training; population characteristics and prevalence NR <u>Validation</u>: in 22 different institutions; population characteristics and prevalence NR Multiple scans may have been included for some patients.

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
	in the study design, data collection, analysis, decision to publish or preparation of the manuscript).	**GREEN**	*Image acquisition:* Lumbar and thoracic X-rays between February 2022 and June 2022 Image deidentification not reported. *Setting:* Germany (N=1); level 3 trauma centre		
17.	Page (JBMR Plus, 2023; e10778) *Funding:* NR *Declaration of interests:* One author is employed by and owns shares in Amgen, multiple received funding from Amgen.	Retrospective diagnostic accuracy study (n=1,200) *Intervention:* Zebra (EAG notes that Nanox AI acquired Zebra Medical Vision in 2021; Company have confirmed evidence is generalisable to HealthVCF) **AMBER** *Reference standard*: Two neuroradiologists (blinded) **AMBER**	*Inclusion:* Aged 50 years or older, CT of chest or abdomen or pelvis. *Exclusion:* NR **GREEN** *Image acquisition:* Consecutive (reverse chronological order) chest or abdomen/pelvis CT, between 2012 to 2017, until 550 of each acquired with an additional 50 CT scans of chest and 50 CT scans of abdomen/pelvis with previously defined VCF. Multiple CT scanners used (GE, Siemens, Canon). Images de-identified prior to processing by the software. *Setting:* US (N=1)	Failure to process algorithm, diagnostic accuracy **GREEN**	<u>Training:</u> Details of training population (number or type of images, patient demographics) not explicitly reported, but referenced made to separate paper published in 2017. <u>Validation:</u> aim of study is a blinded validation <u>Sample size:</u> 1000 patients required to determine if Zebra VCF detection algorithm had a positive likelihood ratio above 10 (sensitivity of 09% and specificity of 91%) and 95%CI to exclude a positive likelihood ratio of 8. consecutive scans used

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
					in reverse chronological order until 550 chest CT and 550 abdominal/pelvis CT were acquired; no additional detail provided. Authors acknowledge that the output of the software was binary and did not identify the location of the fracture. Limitations noted that chest and abdominal CT do not enable evaluation of the entire spine, and that false positives may be influences by intervertebral osteochondosis, Schmorl's nodes, congenital abnormalities requiring validation of findings by trained staff to ensure appropriate diagnosis.
18.	Pereira (Radiol Bras, 2024; e20230102) *Funding:* NR *Declaration of interests:* NR	Retrospective diagnostic accuracy study (n=964 scans) *Intervention:* HealthVCF **GREEN**	*Inclusion:* consecutive chest and abdominal CT scans for various indications which were not ordered specifically for assessment of the spine but included it within the field of view, age between 51 and 99 years	Failure to process the image, diagnostic accuracy. **GREEN**	Training: NR Validation: NR The study additionally reports diagnostic accuracy split by chest

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
		Comparator: General radiologists (N=NR) – not specialists in musculoskeletal imaging. *Reference standard*: Two radiologists specialised in musculoskeletal imaging. **GREEN**	*Exclusion:* presence of metastases (pathological features). **AMBER** *Image acquisition:* Chest and abdominal CT, 1 year period (not defined). All CT performed with volumetric acquisition in axial plane, slice thickness of 1.0mm for chest, and 3.0mm for abdominal scans, multi-detector CT scanners (Siemens), tube voltage 120 kVp, mean tube current 107 mAs, 3mm sagittal reconstructions were available for abdominal CT, sagittal reconstructions of the thoracic and lumber spine were performed. Image deidentification not reported *Setting:* Country NR (N=1); hospital		CT angiography, abdominal CT, chest CT.
19.	Roux (Rheumatology, 2022; 3269-3278) *Funding:* French Ministry of Health, Programme Hospitalier de Recherche Clinique.	Retrospective diagnostic accuracy study (n=173,720 patients, random sample of 500 patients half with VCF used in substudy).	*Inclusion:* 60 years and older, having a CT involving the lumbar spine performed regardless of medical indication. One scan per patient.	Diagnostic accuracy, radiodensity of lumbar vertebrae using Hounsfield Units, simulated DEXA scan T-scores **AMBER**	Training: NR Validation: NR Authors acknowledge that the analysis did not account for differences in

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
	Declaration of interests: One author is an employee of Zebra Medical Vision, other authors report no other conflicts.	*Intervention:* HealthVCF by Zebra Medical (EAG notes that Nanox AI was previously Zebra Medical) **GREEN** *Reference standard*: 2 blinded experts (no further detail provided) reviewed a subset of 500 scans (250 with VCF). **AMBER**	*Exclusion:* non-legibility of the CT scan by the software for measurement of bone parameters. **AMBER** *Image acquisition:* abdominal, thoraco-abdominal and spinal CT, number of different scanners used, between 1 January 2007 and 31 December 2013. Study used pseudonymised data; age and sex were recorded in meta-data of the CT scan. Image deidentification not reported *Setting:* France (N=35); hospitals and clinics		peak voltage or CT scanner manufacturer in its analysis. The study did not assess the thoracic spine, the number of fractures or their severity.
20.	Spangeus (Arch Osteopor, 2025; 42) *Funding:* AIDA/Medtech4Health, Vinnova, Sweden (open access finding Linkoping University) *Declaration of interests:* One author received fees from UCB, AMGEN, and Tromp Medical. Three authors	Retrospective diagnostic accuracy study; validation study (n=101 patients) *Intervention*: IB Lab FLAMINGO software (IB Lab GmbH, Vienna, Austria) **GREEN** *Reference Standard*: 2 radiologists (16 and 24 years' experience)	*Inclusion:* All patients identified were previously included in another study focused on falls and fall prevention during in-hospital care in the geriatric ward a hospital (between 2018 and 2020), patients had undergone a CT scan of thorax or abdomen for any reason within 6 months before or after the fall event. Image deidentification not reported	Diagnostic accuracy (sensitivity, specificity, PPV, NPV) in detection of moderate or severe VFF (grade 2 or 3) **GREEN**	Training: NR Validation: purpose of study The EAG notes that diagnostic accuracy is reported per scan (with 1 to 5 scans per patient). The study reports that cervical vertebral fractures are not analysed. The authors

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
	are employees of IB Lab GmbH.	blinded to each other's assessments. In cases of disagreement, a consensus decision was reached (method unreported). **AMBER**	*Exclusion*: NR **GREEN** *Image Acquisition*: CT scans for thoracic pathologies aortic assessment, spinal imaging, and abdominal pathologies performed with or without contrast enhancement. The scanners utilized were various Siemens models, and PET/CT scans from a GE machine. *Setting:* Sweden (N=1)		report the need to keep false positives to a minimum to reduce workload on radiologists, and need for prospective studies.
21.	Talwar (RSNA, 2023; W5B-SPCH-2) Abstract *Funding:* NR *Declaration of interests:* Author reported nothing to disclose.	Retrospective diagnostic accuracy study (n=1,559 chest x-rays) *Intervention:* Annalise Enterprise CXR **GREEN** *Comparator*: original radiology reporter (n=NR) *Reference standard: a* radiologist with 10 years or more experience reviewed discrepancies between	*Inclusion:* consecutive adults (aged 18 years and older) *Exclusion:* NR **GREEN** *Image acquisition:* Chest X-rays acquired in 2016 (de-identified). *Setting:* Australia *(*N=1)	Successful AI processing of image. 60 clinical findings (including pulmonary nodules, pleural effusions, spinal compression fractures, airspace opacities, acute rib fractures). **AMBER**	Abstract only, lack of detailed methodology. <u>Training</u>: NR <u>Validation</u>: NR At stakeholder consultation the company stated that the population used to train and validate the AI for this study was reported in [Seah et al. 2021]().

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
		intervention and comparator AMBER			
22.	Wiklund (J Bone Mineral Res, 2024; 1113-1119) *Funding:* Region Halland (funders had no influence over study design, data collection, analysis or preparation of the manuscript). *Declaration of interests:* Authors report nothing to disclose.	Retrospective diagnostic accuracy study (n=1,112) *Intervention:* Aidoc **GREEN** *Comparator:* original radiologist report **GREEN** *Reference standard*: all CT reviewed by general radiologist (10 years experience), all identified VCFs were reviewed by a musculoskeletal radiologist (over 30 years experience) for final classification. **AMBER**	*Inclusion:* Patients aged over 60 years, with at least 1 CT including the abdomen (e.g. abdominal, abdominopelvic, thoracoabdominal) with an indication for the examination other than suspicion of lumbar vertebral fracture. Only the first examination per patient in study period was included. *Exclusion:* additional CTs during the time period, pathologic VCFs, performed at an outside institution, CT biopsy examination. **GREEN** *Image acquisition:* Abdominal CT acquired between 18 January 2019 and 18 January 2020. CT scans with and without contrast were accepted. All examinations performed on 64-slice multidetector CT scanner (GE Healthcare). CT settings 100-140 kVp depending on protocol, automatic mAs with automatic tube current	Upload and analysis failure. Sensitivity, specificity, PPV, NPV. Instances of DEXA scan examination and healthcare visits for osteoporosis evaluation in the year following. **AMBER**	Training: Notes trained on tens of thousands of CT from a range of scanners from multiple centres worldwide to identify Grade 2 or 3 VCFs. Prevalence and patient demographics NR. Validation: NR

#	Author (year); Funding	Design and intervention(s)	Participants & Setting	Outcomes	EAG comments
			modulation, detector collimation 64x0.625mm, pitch ~1. All examinations included sagittal reconstructions of 5mm. All CT were pseudonymised. *Setting:* Sweden (N=1) centre		

Key: **GREEN** in scope; **AMBER** aspect not in scope; **RED** not in scope. Abbreviations: AI, Artificial intelligence; AP, Anterior-posterior; AUROC, Area under the receiver operating characteristic curve; CI, Confidence interval; CXR, Chest radiography; DEXA, Bone density X-ray scan (dual energy X-ray absorptiometry); FLS, Fracture liaison service; HU, Hounsfield Unit; LAT, lateral; NPV, Negative predictive value; NR, not reported; PA, Posterior-anterior; PACS, Picture archiving and communication system; PET, Positron emission tomography; PPV, Positive predictive value; VCF, Vertebral compression Fracture; VFF, Vertebral fragility fracture; VHL, Vertebral height loss.

Appendix A5: Excluded studies (n=71): clinical and economic

#	Technology	Source	Study	Publication type	Reason(s) for exclusion
1.	BoneView	Gleamer	Altmann-Schneider (2024, Pediatr Radiol; 136-145)	Full text publication	Population: radiographic image did not involve the spine Outcome: no mention of VFF [Note includes patients aged less than 2 years; off-label use]
2.	-	EAG literature search	Artyukova (Digital Diagnostics, 2024; 505)	Full text publication	Language: Only abstract available in English Language Intervention: device not named in abstract (cannot confirm intervention listed within scope)
3.	HealthVCF	EAG literature search	Beeri (SSRN, 2024; 4813440)	Pre-print (not peer-reviewed)	Full publications available with more information reported (unclear how many patients the technology was applied to).
4.	BoneView	Gleamer website	Boginskis (2023, Medicni perspektivi; 68-78)	Full text publication	Outcome: does not report specifically on spine fractures (results aggregated with fractures of other anatomical locations) or VFFs
5.		EAG literature search	Broadstock, (2000, New Zealand Health Technology Assessment (NZHTA), - Report)	Full text publication	Population: Cervical cancer screening not VFF
6.	BoneView	Gleamer website	Canoni-Meynet (2022; Diag Interven Imaging; 594-600)	Full text publication	Outcome: Mention of spine fractures, but no mention of VFFs/VCFs
7.	-	EAG literature search	Chen (Euro Spine J, 2025)	Correction	Study design: correction Intervention: Not listed in final scope (unclear if commercial product)
8.	-	EAG scoping	Chen (Eur Radiol, 2022; 1496-1505)	Full text publication	Intervention: Not listed in final scope (unclear if commercial product)
9.	-	EAG literature search	Chen (PLoS One, 2021; e0245992)	Full text publication	Intervention: Not listed in final scope (ResNet)
10	-	EAG literature search	Choi (European Radiology, 2024; 3750)	Full text publication	Intervention: Not listed in final scope
11	-	EAG literature search	Dagan (Nature Medicine, 2020; 77-82)	Full text publication	Intervention: multiple interventions (Zebra Medical involved in development of algorithm, but unclear if

#	Technology	Source	Study	Publication type	Reason(s) for exclusion
					intervention is a commercial product or prototype)
12	-	EAG literature search	Dalal (Aging Clin Exper Res, 2022; 1909-1918)	Full text publication	Intervention: Not listed in final scope (expert elicitation of standard care)
13	-	EAG literature search	Davy (Biomedical J Sci & Technical Res, 2021; 30889-30897)	Full text publication	Intervention: Not listed in final scope (no AI) Study design: Narrative review
14	-	Avicenna. AI	G. Chaix (2024), Thesis Manuscript, (Published). Sainte Marguerite Hospital, Marseille	Unknown	Full text not found in the public domain by EAG and not provided by company
15	-	EAG literature search	El Kojok (Comp Biol Med, 2025; 109446)	Full text publication	Intervention: Not technologies listed in scope
16	-	EAG literature search	Fanni (Diagnostics, 2023; 2020)	Full text publication	Study design: Narrative review Outcome: Not focused on VFFs/VCFs
17	-	EAG literature search	Ferrar (Osteoporosis Int, 2012; 59-65)	Full text publication	Population: Vertebral fracture at baseline Intervention: Not listed in final scope
18		EAG literature search	FranÃ§ois-Emery, (2009, . Medical Decision Making 29(1): 125-139).	Full text publication	Study design: Not an economic evaluation study for VFF/VCFs
19	-	EAG scoping	Gatineu (J Bone Mineral Res, 2024; 1-21)	Full text publication	Intervention: AI technologies not named
20	-	EAG literature search	Glessgen (medRxiv, 2022)	Full text publication	Intervention: Not technologies listed in scope (available on Github)
21	BoneView	Gleamer website	Guermazi (2022, Radiology; 627-636)	Full text publication	Outcome: Mention of spine fractures, but no mention of VFFs/VCFs
22	Zebra	EAG literature search	Gunasingam (Ger Orthopaed Surg, 2021; 51)_ABSTRACT	Abstract	Larger study with more detail included by the EAG
23	-	EAG literature search	Gunasingam (Int Med J, 2020; 28)_ABSTRACT	Abstract	Duplicate: of Gunasingam (Ger Orthopaed Surg, 2021; 51)_ABSTRACT
24	BoneView	Gleamer website	Hayashi (2022, Skeletal Radiol; 2129-2139)	Full text publication	Population: limbs Outcome: no spinal/vertebral data (images including spine explicitly excluded)

#	Technology	Source	Study	Publication type	Reason(s) for exclusion
25	-	EAG literature search	Hong (J Bone Mineral Res, 2023; 887-895)	Full text publication	Intervention: not listed in final scope
26	-	EAG scoping	Howlett (Clin Radiol, 2023; e1041-e1047)	Full text publication	Intervention: no AI tools (RCR audit of VFF reporting on CTs)
27	-	EAG literature search	Huber (Bone, 2024; 117176)	Full text publication	Intervention: not listed in final scope
28	-	Milvue	Internal study Paris	Unknown	Full text not found in the public domain by EAG and not provided by company
29		EAG literature search	Jin (2011, Medical Decision Making 31(2): 299-307)	Full text publication	Population: Observational prediction study for Hip fracture, not VFF/VCFs
30	-	EAG literature search	Karunasena (J Med Image Radia Oncol, 2022; 95)_Abstract	Abstract	Duplicate: full paper available Jones (BMJ Open, 2021; e052902)
31	-	EAG literature search	Kolanu (Osteopor Int, 2020; S179)_ABSTRACT	Abstract	Duplicate: of Kolanu (Journal of Bone and Mineral Research, 2020; 2307-2312)
32	-	EAG literature search	Krishnaraj (Journal of the American College of Radiology, 2019; 1473-1479)	Full text publication	Comparator, Outcome: Calculated BMD from CT compared with DEXA scan
33	-	EAG literature search	Lee (Diagnostics, 2024; 2477)	Full text publication	Intervention: Not device in final scope (deep-learning algorithm commercialised as ClariPi)
34	-	EAG Scoping	Lee (RSNA, 2023; S3B-SPMK-3)_abstract	Full text publication	Intervention: ClariVCF by ClariPi (not in scope)
35	-	EAG literature search	Liu (Infect and Drug Resist, 2025; 31)	Full text publication	Intervention: Not devices in final scope
36	-	EAG literature search	Matsumoto (Journal of the American College of Radiology, 2024)	Commentary	Study design: commentary on Ghatak (Jornal of the American College of Radiology, 2024; 220-229)
37	-	EAG literature search	Mehta (J Digit Imaging, 2020; 204-210)	Full text publication	Intervention: Not listed in final scope
38	-	Nanox AI	MK Javaid (2019), Oxford, UK, Using Artificial Intelligence Technology to Improve Case Finding for Vertebral	Unknown	Full text not found in the public domain by EAG and not provided by company

#	Technology	Source	Study	Publication type	Reason(s) for exclusion
			Fractures in the Fracture Liaison Service (FLS) Setting		
39	-	EAG literature search	Nadeem (Med Phys, 2024; 4201)	Full text publication	Intervention: Not listed in final scope
40	-	EAG literature search	NCT00388908	Trial registration	Intervention: Not listed in final scope (behavioural intervention)
41		EAG literature search	Nannan (PharmacoEconomics, 2023; 363-391).	Full text publication	Population: Interventions for osteoporosis, not VFF/VCFs
42	-	EAG literature search	Nissinen (Bone Reports, 2021; 101070)	Full text publication	Intervention: Not listed in final scope Outcome: No mention of VFFs/VCFs.
43	BoneView	Gleamer website	Novak (2024, BMJ; 086061)	Full text publication	Study design: Protocol (ISRCTN19562541; NCT06130397) Outcome: no mention of vertebral fragility fractures
44	-	EAG literature search	Oh (Endocrinol and Metabol, 2024; 500)	Full text publication	Intervention: Not listed in final scope
45	-	EAG literature search	Ong (Osteoporosis Int, 2021; 921-926)	Full text publication	Intervention: Not listed in final scope (Optasia)
46	BoneView	Gleamer website	Oppenheimer (2023, Life; 223)	Full text publication	Outcome: not specific to vertebral fragility fractures (any fracture). Mentions adverse event (AI mislabelling intervertebral spaces as a fracture).
47	-	EAG literature search	Page (Arthritis Rheumatol, 2020; 3948)_ABSTRACT	Abstract	Duplicate: of Page (JBMR Plus, 2023)
48	-	EAG scoping	Parreira (Spine J, 2017; 1932-1938)	Full text publication	Population, Intervention: Systematic review of clinical guidelines in management of VCF
49	-	EAG literature search	Pickhardt (Radiology, 2020; 64 -72)	Full text publication	Intervention: Not listed in final scope
50	-	EAG literature search	Poullain (Eur J Radiol, 2023; 110642)	Full text publication	Intervention: Not listed in final scope

#	Technology	Source	Study	Publication type	Reason(s) for exclusion
51	CINA-VCF	Avicenna AI	Quemeneur (ECR, 2024; C-1181)]	Abstract	Duplicate of Guenoun (Clin Radiol, 2025; 106831)
52	-	EAG literature search	Quenet (Applied Radiology, 2023; 22-23)	Commentary	Study design: Commentary
53	-	EAG literature search	Qui (Bone, 2025; 117330)	Full text publication	Intervention: Not listed in final scope (unclear if commercial product)
54	-	EAG literature search	Ramschutz (Osteoporosis Int, 2025)	Full text publication	Intervention: Not listed in final scope
55	-	EAG literature search	Ryu (Comput Struct Biotech J, 2023; 3452)	Full text publication	Intervention: Not listed in final scope
56	Annalise.AI	Annalise.AI	Seah (Lancet Digital Health, 2021; e496-e506)	Full text publication	Outcome: focus on chest findings
57	-	EAG literature search	Seo (Sci Report, 2021; 13732)	Full text publication	Intervention: Not listed in final scope
58	-	EAG literature search	Seol (Eur Spine J, 2024; 3221-3229)	Full text publication	Intervention: Not listed in final scope
59	-	Stakeholder consultation	Shelmerdine (Clinical Radiology, 2024; 665-672)	Full text publication	Study design: narrative review Population: not specific to vertebral fracture detection (broadly discussion of selecting AI in radiology).
60	-	EAG literature search	Shen (J Bone Mineral Res, 2023; 1278-1287)	Full text publication	Intervention: Not listed in final scope
61	-	EAG literature search	Silberstein (J Clin Med, 2023; 7730)	Full text publication	Intervention: Not listed in final scope (OfEye 1.0)
62	-	EAG literature search	Tomita (Computers in Biology and Medicine,2018; 8-15)	Full text publication	Intervention: Not listed in final scope
63	-	EAG literature search	Wang (BMC Med Imag, 2025; 41)	Full text publication	Intervention: Not listed in final scope
64	-	EAG literature search	Wang (Academic Radiology, 2025; 298)	Full text publication	Intervention: Not listed in final scope
65	-	EAG literature search	Wang (Quant Imag Med Safety, 2024; 800-813)	Full text publication	Intervention: Not listed in final scope

#	Technology	Source	Study	Publication type	Reason(s) for exclusion
66		EAG literature search	Willis, (2005, NIHR Health Technology Assessment programme)	Full text publication	Population: Cervical cancer screening not VFF/VCFs
67	-	EAG literature search	Xu (Frontiers in Endocrinol, 2023; 1025749)	Full text publication	Intervention: Not listed in final scope
68	-	EAG literature search	Yang (Osteoporosis Int, 2022; 2547)	Full text publication	Intervention: Not listed in final scope (AI-Rad Companion by Siemens Healthineers)
69	-	EAG literature search	Yildiz Potter (J Imag Informat Med, 2024; 2428-2443)	Full text publication	Intervention: Not listed in final scope
70	-	EAG literature search	Zhang (Frontiers Bioeng Biotech, 2024; 1397003)	Full text publication	Intervention: Not listed in final scope
71	-	EAG literature search	Zhang (Insights Imaging, 2024; 290)	Full text publication	Intervention: Not listed in final scope

Abbreviations: AI, Artificial intelligence; BMD, Bone mineral density, DEXA, Bone density X-ray scan (dual energy X-ray absorptiometry); VCF, Vertebral compression fracture; VFF, Vertebral fragility fracture.

Appendix B – Economic modelling

Appendix B1: Results from generic AI base case

Base case

Programme	Leaf	Probability	Cost	Benefit	Utility	QALY
AI	F.VF.managed	0.02567	137	0	0.01361	0.01361
AI	G.VF.surveillance	0.1502	5.047	0	0	0
AI	H.VF.unmanaged	0.1182	1.182	0	0	0
AI	I.VF.unmanaged	0.00297	0.0297	0	0	0
AI	J.NoVF	0.00703	0.0703	0	0	0
AI	K.NoVF	0.000696	0.02339	0	0	0
AI	L.NoVF	0.6953	6.953	0	0	0
SoC	A.VF.managed	0.01097	58.43	0	0.005814	0.005814
SoC	B.VF.surveillance	0.06417	1.515	0	0	0
SoC	C.NoVF	0.07663	1.809	0	0	0
SoC	D.NoVF	0.6264	0	0	0	0
SoC	E.VF.unmanaged	0.2219	0	0	0	0

Totals

Programme	Probability	Cost	Benefit	Utility	QALY
AI	1	150.3	0	0.01361	0.01361
SoC	1	61.75	0	0.005814	0.005814

Appendix C – Additional detail on AI technologies

Appendix C1: Additional technical details

Device (Company)	AI filters images	Are images de-identified before AI processing?	AI Training Datasets	Does the technology work on compressed or 'Lossy' images	Underlying VFF prevalence across training and validation data
Annalise Enterprise CXR and Annalise Container CXR, (Annalise.AI)	Yes – DICOM. Automatic fetch or forward from PACS, filtered by Study Description, Body Part Examined or Exam/Procedure Code, or a combination of these. Patient Age: ≥ 016Y, Modality Header: CR or DX DICOM SOP Class only. Study must contain at least one frontal projection (AP or PA) and may optionally contain two AP/PA and lateral projections.	[redacted]	Chest x-rays used for the training dataset were obtained from multiple datasets: I-MED Radiology Network (I-MED; Australia), MIMIC (Beth Israel Deaconess Medical Center,Boston, MA, USA), ChestX-ray (NIH Clinical Center, Bethesda, MD, USA), CheXpert (Stanford University Medical Center, CA, USA), and PadChest (Hospital San Juan, Spain; appendix p 4). Images of patients ≥16 years old & at least one frontal chest x-ray. DICOM tags were removed. Protected health information, excluding age and sex, was removed by automated process. Patient and case ID information were anonymised. Image data was preserved at original resolution and bit-depth	Yes, AI has been validated on both lossy and lossless imaging. AUC of 0.975 spine wedge fracture and 0.962 for osteopenia was observed in both formats	Training set: 782,146 studies 58,808 = Spinal Wedge 62,664 = Osteopenia. (No further breakdown available) Validation set: 12.83% VFF prevalence from 2,565 cases. 329 cases = Spinal wedge 292 cases = Osteopenia At stakeholder consultation the company shared that the Annalise CXR model was explicitly developed and validated to detect a wide range of findings visible on chest X-rays—including spine-related findings visible on CXR, including diffuse spinal osteophytes, kyphosis, osteopaenia, scoliosis, spinal fixation, spinal arthritis, and spine lesions, as well as technical factors indicative of poor image quality (Seah et al. 2021).
BoneView, (Gleamer)	Did not respond	Did not respond however RFI supplied from HTE20 indicates images are pseudonymised before being shared outside the NHS-environment.	Did not respond	Did not respond	Did not respond
TechCare Spine, (Milvue)	Did not respond	Unclear from RFI, no IFU supplied and company did not respond to EAG questions for confirmation	Did not respond	Did not respond	Did not respond
BriefCase-Triage, (Aidoc Medical)	Uses image-based orchestration process to identify the scan and choose relevant AI algorithms based on the detected anatomy before selecting the optimal image series for analysis	Yes, deidentified prior to being sent to Aidoc cloud environment	Trained using tens of thousands of images from various institutions to cover diverse range of patient populations and scanner makes/models to mitigate bias	Yes	Training set: ~5000 cases, 2,925 chest and 2,252 abdomen CT scans. Approx. 707 contained positive findings in lower lumbar spine and 1,170 positive findings in thoracic spine. Validation set: 318 cases, 184 positive cases and 134 negatives.
CINA-VCF Quantix (Avicenna.AI)	Yes, we perform a protocol check to ensure the technical	Yes, If the solution is deployed in the cloud. If it's deployed on-premise, it's not needed.	[redacted]	No	Training set: Grade 0 (VHL<20%) - 89.1% (N=10041)

Device (Company)	AI filters images	Are images de-identified before AI processing?	AI Training Datasets	Does the technology work on compressed or 'Lossy' images	Underlying VFF prevalence across training and validation data
	acquisition parameters are suitable		[redacted]		Grade 1 (20%≤VHL 25%) – 4.5% (N=510) Grade 2 (25%≤VHL 40%) – 4.7% (N=528) Grade 3 (VHL>40%) – 1.7% (N=189) Validation: Grade 0 - 52 (3.36%) Grade 1 - 14 (8.2%) Grade 2 – 51 (30.0%) Grade 3 – 53 (31.2%)
HealthVCF and HealthOST (Nanox AI)	Yes, AI checks and assures that scans comply with inclusion criteria otherwise they are rejected	Yes	NR	Model was trained with some Lossy images	Training set: 33% positive series, 67% of negative series No fracture 44.1%, mild 22.5%, moderate 24.9%, 8.5% Validation set: 50.3% positive 49.7%
IB Lab LAMINGO (IB Lab)	[redacted]	Yes	[redacted]	[redacted]	Training set: [redacted] Validation set: [redacted]

Abbreviations: AI, Artificial intelligence; AUC, Area under curve; HU, Hounsfield units; IFU, Instructions for use; RFI, Request for information; VF, Vertebral fracture; VFF, Vertebral fragility fractures; VHL, Vertebral height loss.

Appendix C2: Additional published outcomes (CINA-VCF)

The study by (Guenoun et al., 2025) also reported vertebral height loss measurements (%) in 20 patients and 336 vertebrae made by the AI technology (CINA-VCF) and two radiologists (reference standard). The study reported Bland Altman plots which showed similar 95% limits of agreement (LoA) between the two radiologists (mean difference: 1.6 [95% LoA: -8.1 to +11.3]), and between the AI and the reference standard (mean difference: -0.4 [95% LoA: -9.3 to 8.6]). The authors reported that 94.1% (317/337) of the differences between the AI and the reference standard lay within the limits of agreement of two radiologists (which contributed to the reference standard).

The study by (Guenoun et al., 2025) also reported Hounsfield Units (HU) on 93 patients. This was due to 7 cases being excluded as the mean HU did not include the same vertebrae; for example where the algorithm detected L1 as compressed, and therefore the mean HU was measured between L2 and L4; whereas the radiologist (reference standard) confirmed that L1 was not compressed and therefore calculated mean HU across L1 to L4 vertebrae. The Pearson correlation coefficient for mean HU measurement between the AI and reference standard was 0.89 (95%CI 0.84 to 0.92; $p<0.0001$) indicating a strong correlation between the two.

Appendix D – Correspondence log with Experts

Appendix D1: Questions sent to SCMs 12/02/2025 and to clinical experts 13/02/2025

#	Date responses received	Name, Affiliation
1	18/02/2025	[redacted]
2	18/02/2025	[redacted]
3	14/02/2025	[redacted]
4	12/02/2025	[redacted]
5	14/02/2025	[redacted]
6	18/02/2025	[redacted]
7	19/02/2025	[redacted]
8	14/02/2025	[redacted]

Question 1	The EVA focuses on opportunistic detection of vertebral fragility fractures (assessing radiographic images that include the spine taken for reasons other than a suspected vertebral fracture). How common does opportunistic detection of VFFs currently occur without the assistance of AI?
Expert 1:	As a Rheumatologist, it is difficult to comment on personal experience related to the incidence of opportunistic VFFs without AI assistance. However, nationally there have been multi-centre audits that have demonstrated improvement in practice following introduction of the RCR guideline with CT studies (Howlett DC, Drinkwater KJ, Mahmood N, Salman L, Griffin J, Javaid MK, Retnasingam G, Marzoug A, Greenhalgh R; Collaborating Author Group. Radiology reporting of incidental osteoporotic vertebral fragility fractures present on CT studies: results of UK national re-audit. Clin Radiol. 2023 Dec;78(12):e1041-e1047. doi: 10.1016/j.crad.2023.09.004. Epub 2023 Sep 29).
Expert 2:	No
Expert 3:	Whilst not explicitly opportunistic, the FLS national database KPI3 is a good indicator of the variability in identifying spinal fractures across the UK. This website also has data going back to 2016 which suggests we are getting better at VFF detection: https://www.fffap.org.uk/FLS/charts.nsf/benchmarks?readform&yr=2024&vw=&org1= However, it is generally accepted that more than half of vertebral fractures are missed in current UK practice: /https://strwebprdmedia.blob.core.windows.net/media/clvbclel/sotn-report-2021-v6-2-final-1.pdf
Expert 4:	I don't think this is known. There will be national variation based on local clinical guidelines, and based on interest of clinicians/presence of a clinical champion.
Expert 5:	Unable to provide any insight into this question
Expert 6:	Vertebral fragility fractures (VFFs) are difficult to detect without AI assistance, especially as radiologists face high workloads and fatigue. Since VFFs are not typically the primary focus of CT scans, X-rays, or MRIs, they are often identified opportunistically. The publication of the RCR Guidelines on VFF in 2021 has led to improvements in both practice and detection rates. Osteoporosis leads now promptly identify and report VFFs in line with established policies and service-level agreements (SLAs). Additionally, the adoption of AI for VFF detection has been a significant advancement, gradually reducing missed diagnoses and improving patient outcomes.

Expert 7:	[redacted] This is in part because many patients do not have symptoms or there is a delay in patients presenting due to managing their own symptoms. Those that come to clinical attention are most likely to be associated with back pain that is sufficiently severe at the acute phase, or associated with a notable trauma. In many patients however, back pain may not alert health professionals or patients to the presence of a vertebral fracture, and it will remain undiagnosed. Vertebral fractures are typically diagnosed radiographically (X-ray, CT or MRI) using recognised and validated methods. They can also be diagnosed via vertebral fracture assessment completed at the time of a bone mineral density assessment with dual-energy X-ray absorptiometry (DXA). VF's can be identified occasionally on a chest X-ray examination including the lateral projection but this is dependent on the imaging parameters (exposure factors). The lateral chest X-ray projection is not a standard projection in many imaging departments due to most patients being referred for CT rather than acquiring a lateral chest projection. [redacted] There is more engagement and recognition of the importance of identifying VF's within Radiology following the RCR publication on all imaging and following these patients up appropriately. Since this publication, we have appointed an osteoporosis lead within Radiology to develop a protocol and lead on regular audit of the identification of VF's. We have also developed a standard reporting code that can be used in plain film reporting and cross-sectional imaging to ensure that the referrer is alerted to the presence of VF's and is asked to assess the patients fracture risk and refer for a DXA scan if appropriate. We have also agreed on using unambiguous language in reports to make it clear that when a vertebral fracture is identified on imaging. 3. Cooper C, Atkinson EJ, Jacobsen SJ, O'fallon WM, Melton LJ. Population-based study of survival after osteoporotic fractures. Am J Epidemiol [Internet]. 1993 May 1 [cited 2022 Jan 6];137(9):1001–5. Available from: https://pubmed.ncbi.nlm.nih.gov/8317445/ Royal Osteoporosis Society (ROS). Clinical Guidance for the effective identification of vertebral fractures [Internet]. 2017 [cited 2022 Jan 7]. Available from: https://theros.org.uk/media/3daohfrq/ros-vertebral-fracture-guidelines-november-2017.pdf Fink HA, Milavetz DL, Palermo L, Nevitt MC, Cauley JA, Genant HK, et al. What proportion of incident radiographic vertebral deformities is clinically diagnosed and vice versa? Journal of Bone and Mineral Research. 2005 Jul;20(7):1216–22.
Expert 8:	I don't think that VFF's are commonly detected opportunistically. I don't think that the RCR VFF guidelines have been widely implemented.
Question 2	Are vertebral fragility fractures the same as vertebral compression fractures? Or are compression fractures a subset (if so, are any other subgroups of fragility fractures we should be aware of)? Are there other terminology used to describe fragility fractures that we need to be aware of when looking at the literature?

Expert 1:	• The term “fragility” within VFFs denotes the presence of an underlying weakened bone state secondary to conditions such as Osteoporosis or Osteopenia. Therefore, VFFs relate to fractures of the vertebral body secondary to this weakened bone state. Although often referred to in the same context, the term “compression” within vertebral compression fractures denotes the process of the vertebral body “collapsing” or reducing in height. Although compression fractures often occur due to conditions such as Osteoporosis, they can also be caused by other pathological states such as infections, cancers and trauma. This is the distinction between the two, and therefore compression fractures should be seen as a subset. • -Other “subsets” or “terms” to consider include: “wedge fractures”, “crush fractures”, “biconcave fractures”, “loss of vertebral height”, “insufficiency fractures”, “osteoporotic fractures”, “low-energy fractures”
Expert 2:	In principle yes. Vertebral compression fractures commonly include fractures due to cancer, therefore vertebral ‘fragility ‘fractures are often used to denote compression fractures due to osteoporosis.
Expert 3:	Others on the committee will be better at this than me, my understanding is that vertebral compression fractures are essentially synonymous with VFFs. There are categories of compression fracture (wedge, biconcave, crush). A real problem in clinical reporting is that the word “fracture” is not always used in radiology reports which means VFFs may not be acted upon or taken as seriously, such as when terms like “wedging” or “height loss” are used.
Expert 4:	Yes – multiple different terminologies all meaning vertebral fragility fracture is a very difficult area and is one explanation for why they are easily missed. All these terms indicate a vertebral fragility fracture: vertebral fractures, vertebral fragility fractures, crush fracture, wedge fracture, compression fracture, depression of the superior endplate, vertebral collapse. These terms may also indicate a vertebral fragility fracture is present, or that there might be degenerative change: height loss, anterior height loss. . [redacted]
Expert 5:	As a physiotherapist and in my experience the terminology used to describe a Fragility fracture or compression fracture of the spine is interchangeable. As holistic practitioners though we would be taking a history to determine the history of the presenting case and building a picture of how the injury occurred.
Expert 6:	Vertebral fragility fractures (VFFs) and vertebral compression fractures (VCFs) are similar but not identical. VFFs are typically caused by low-trauma events due to weakened bones, often linked to osteoporosis. VCFs can also result from weakened bones but may occur from higher-trauma events, osteoporosis, injuries, or infections. In some cases, fragility fractures can progress to become compression fractures, making VCFs a subset of VFFs. In addition to vertebral fragility fractures, other common fragility fractures include: • Wrist Fragility Fractures (WFFs) • Hip Fragility Fractures (HFFs) • Pelvic Fragility Fractures (PFFs) • Proximal Humerus Fragility Fractures (PHFFs) • Rib Fragility Fractures (RFFs) Common terminologies used in the literature to describe fragility fractures include 'Osteoporotic,' 'Compression,' 'Spontaneous,' 'Low-Trauma,' 'Stress,' 'Insufficiency,' 'Pathological,' 'Senile,' and 'Postmenopausal' fractures.

Expert 7:	In the absence of a significant mechanism of injury/trauma, for example road traffic collision (RTC)/ fall down a full flight of stairs vertebral fragility fractures are the same as vertebral compression fractures. Vertebral fractures attributed to fragility/ osteoporotic are usually sustained with minimal trauma/injury such as a fall from a standing height etc. The most widely used technique is the semi-quantitative method described by Genant et al. This system involves the visual recognition of a loss of vertebral body height on a lateral projection combined with careful assessment of the vertebral endplates to diagnose a fracture. The fracture can be graded using the Genant semi quantitative method, within the report content Grade 1 (mild) fracture 20–25% loss of height. Grade 2 (moderate) fracture 25–40% loss of height. Grade 3 (severe) fracture >40% loss of height. Vertebral compression is a descriptive term similar to wedge vertebral fracture, to imply that there is loss of the normal vertebral contours and loss of height. Generalised vertebral compression can be seen in the elderly population with generalised degeneration throughout the spine so it cannot be assumed that all vertebral fractures are related to fragility/osteoporosis and other causes of VF must be excluded included metastatic bone disease and multiple myeloma. Low trauma fracture/ stress fracture/ insufficiency fracture/ pathological fracture /atraumatic fracture / fatigue fractures are other terms used to describe fragility fractures. Other fragility fractures to be aware of that are associated with a high mortality/morbidity are fractures of the hip/pelvis/humerus. Other terms used to describe VF's - Fracture-shaped vertebral deformity (FSVD)/ osteoporotic vertebral fracture (OVF) / osteoporotic-like vertebral fracture (OLVF)[1] Skeletal Radiology (2024) Yì Xiáng J. Wáng et al 53:2563–2574 https://doi.org/10.1007/s00256-024-04678-4
Expert 8:	If a patient develops a compression fracture and they have osteopaenia or any other bone quantity/quality disease (low BMD) then it is a fragility fracture. But if a 20 year old healthy person with normal BMD develops a compression fracture, it is NOT a VFF. Other terminology used to describe fractures; wedge, compression, wedging, endplate depression, loss of height, plana, anterior wedge, biconcave, fragility fracture.
Question 3a	There may be direct evidence for the diagnostic accuracy of the AI; for example comparing MSK-trained reporter + AI with MSK-trained reporter alone, where the reporter is asked to detect VFFs. However, the use-case of the scope is opportunistic detection of VFFs by any reporter who is reviewing a scan for a purpose other than detecting VFFs. a. To what extent is the direct evidence for diagnostic accuracy of an AI applicable to the use-case in the scope, given the different comparators?

Expert 1:	- MSK-trained reporters (and in particular MSK-trained radiologists) should be considered the 'gold-standard' to provide the highest diagnostic benchmark against which accuracy of AI tools should be evaluated. - Despite this, the real-world situation of non-MSK trained reporters reviewing scans requested for an alternative indication is important to consider, and this is where AI assistance will likely have the greatest incremental benefit within the NHS. - Therefore, it is important when considering the evidence to factor in the comparator and weigh the evidence appropriately. For example, in two studies with equivalent characteristics and diagnostic accuracies but with differing comparators (one using MSK-trained reporters, one using non MSK-trained reporters): o Greater positive weighting should be given to the study using MSK-trained reporters, as it more robustly demonstrates the AI's performance against the most experienced human standard. o A strong AI performance against MSK-trained reporters implies that it will likely outperform or substantially assist non-MSK-trained reporters, where the clinical impact is greatest.
Expert 2:	Unsure
Expert 3:	think given the wide remit of imaging modalities in the EVA, it will be important to gauge increases in performance based on the wide range of reporting experience which could vary from a reporting radiographer who only reports chest x-rays, to a consultant MSK radiologist looking at CT. A big concern with studies has to be a participant bias. If reporters know they are in a study based on VFF detection, there performance will be exaggerated compared to typical practice. Similarly, given the low pick up rates of VFF, a straight DTA study may not consider clinical nuances. If a CT has evidence of terminal metastatic cancer, a radiologist may not report an opportunistic VFF (rightly or wrongly), but AI would. There are also ultimately subjectivities in the perception and categorisation of vertebral fractures as mild, moderate, severe (or even present at all) which will effect clinical decision making
Expert 4:	I don't think it is relevant. The point is that the in the opportunistic scenario, the reporter is focused on finding the cancer or whatever, not looking for vertebral fragility fractures. So, although the diagnostic accuracy of the AI output needs to be good, the very action of reminding the reporter is critically important and is independent of the accuracy.
Expert 5:	Unable to comment
Expert 6:	Since the MSK-trained reporter is already specialized in detecting musculoskeletal conditions, they are more likely to notice VFFs even if it is not the primary reason for reviewing the scan. AI, designed specifically for detecting VFFs, will act as an extra layer of support. This collaboration would reduce errors and improve accuracy by catching subtle findings or mitigating cognitive fatigue and bias. While the reporter is trained for MSK conditions, human error (e.g., distractions, fatigue, or missing subtle signs) can lead to overlooked VFFs, especially when VFF detection is not the main focus of the scan review. Without AI support, there is a higher risk of missed detections or false negatives in opportunistic scenarios. The combined use of AI and MSK-trained reporters is more likely to lead to accurate opportunistic detection of VFFs because the AI provides a "second set of eyes," tailored for VFF identification. In contrast, relying solely on the MSK-trained reporter, especially in opportunistic settings, may result in lower accuracy due to human limitations and competing priorities during the scan review process.

Expert 7:	AI technologies cannot currently be used autonomously without human interpretation. Therefore, the images still need to be interpreted by somebody competent and confidant to do so – The gold standard is an MSK Consultant Radiologist / MSK Reporting Radiographer. I would have reservations of other health professionals issuing opinions directly to the patient based on the AI findings with the absence of a formal radiological report. In particular with having experience of normal variants of the spine which can have the appearance of vertebral deformities but are not true fractures. To diagnose a patient with VFF’s incorrectly has huge implications from the patient’s perspective. The comparator is standard care where the radiologist or radiographer interprets the radiograph without AI assistance, usually within 24 hours of the image being taken – Not all reports can be issued within 24 hours of acquisition it will be determined by the referral type i.e. ED, GP etc.
Expert 8:	I am not aware that there is any direct evidence
Question 3b	Would any other evidence be considered generalisable to the decision problem?
Expert 1:	o External Validation Studies: Multicentre studies evaluating AI accuracy against human reporters at different NHS trusts, using different imaging protocols or across different regions/countries are essential to support understanding of model generalisability. o Real-world Evidence: Any studies using an AI tool as part of a clinical workflow (perhaps as part of a retrospective or prospective cohort) would be important to understand direct applicability to NHS clinical pathways.
Expert 2:	Unsure

Expert 3:	the EVA scope looks to investigate clinical and cost effectiveness. It will be important to consider the criteria that might apply to a business case for purchasing an AI VFF product. A cynic (not necessarily me!) may argue: 1) Does it increase productivity? It may not if it increases reporting times and leads to more FLS referrals or extra burdens on GP services who need to act on bisphosphonate treatments. 2) Does it reduce cost by improving diagnostic accuracy?: whilst it may spot more fractures, it may not reduce costs (at least in the short term) as many VFFs are currently often not treated, or treated with relatively inexpensive pain management, and the primary risk is further risk of fracture, which may be further down the line. So it may actually cost more in the short term. 3) Cost benefits are not realised quickly, they are primarily from reduced fracture rates and associated treatment (such as expensive NoF fractures) into the future. As such, wider questions are important such as: even if fracture detection is improved does it lead to an increase of patients on osteoporosis treatment? Of those treated how many stick with treatment to actually gain benefits? The answers will varied a lot depending on whether an fracture liaison service (FLS) is present. [redacted] opportunistic VFFs were as likely to stick with treatment as symptomatic VFF, but there is still room for improvement. https://www.sciencedirect.com/science/article/pii/S1078817424003808 This economic analysis of opportunistic bone health screening may also be of interest (non-VFF though): https://pmc.ncbi.nlm.nih.gov/articles/PMC11792445/ It will be important that results refer to "per-vertebra" accuracy, as well as "per-patient" in cases where there are multiple VFF. An application or human may miss a single fracture, but the patient may still go down the correct treatment pathway if there are multiple VFFs.
Expert 4:	One good scenario: the rate of diagnosis/reporting of vertebral fragility fractures on CT scan reports in hospitals with/without the AI
Expert 5:	My thoughts turn to Real world workflow and integration with other digital technologies (PACS/EPR's). Evaluating how the new process and workflows compared to the As-is
Expert 6:	Yes, evidence such as those involving the use of AI to opportunistically detect VFFs from diagnosis could be generalized. Also, in cases where the reporters are not even MSK-trained or specialists in this area but are still able to use AI to detect VFFs because they are skilled in the use of AI and can learn on the job details involving VFF detection opportunistically from diagnosis.

Expert 7:	Not all referrers have access to imaging and in particular, facilities in the community, these centres therefore rely on the formal radiological report to help with patient management/ treatment. Subgroups – Patients who are on long-term steroid treatment (steroids longer than 3 months duration) are at very high risk of VFF's should these not be included in a subgroup? I think this should be clearly stated in the information about the subgroups. I am unsure how useful the AI tool will be on a CXR particularly if the CXR has been performed on a very acutely unwell patient, which are quite often performed anteroposterior (AP) as opposed to the gold standard posteroanterior (PA) projection. The quality of the resultant X-ray image will have a big impact of the accuracy of the AI finding therefore using the AI tool on some types of acute patients may not be appropriate for example rotated patients that cannot sit or stand due to presenting symptoms. [redacted]
Expert 8:	Generalisable evidence would be the general accuracy of the AI programme in detecting vertebral fractures. It would also be possible for an AI programme to retrospectively review x-rays/scans from the past 12 months eg put all chest, abdominal, shoulder x-rays through the AI programme and it will flag any that have vertebral fractures. This would then require manual review.
Question 4a	The NICE scope mentions AI tools that can be used in Xray, CT, MRI and potentially Ultrasound, and explicitly states the following technologies: o Annalise Enterprise by Annalise.AI [X-ray] o Annalise Container by Annalise.AI [X-ray] o BoneView by Gleamer [X-ray] o TechCare Spine by Milvue [X-ray] o BriefCase-Triage by Aidoc Medical [CT] o CINA-VCF Quantix by Avicenna.AI [CT] o HealthVCF and HealthOST by Nanox AI [CT] o IB Lab FLAMINGO by IB Lab [CT] Do you have any experience of using the above AI technology (or similar) and how does it fit into the workflow of imaging?
Expert 1:	No direct experience of using the above AI technologies.
Expert 2:	Yes Optasia medical ltd, who I understand are now liquidated. Work flow please see details from publication (provided)
Expert 3:	sorry I do not have any meaningful direct experience
Expert 4:	Yes – I have worked with a similar technology produced by Optasia medical for research purposes. But it was almost impossible to use within the NHS because of (lack of) agreement from IT that the software could be 'housed' within the hospitals IT system. They were very risk averse.

Expert 5:	[redacted] I have worked with other imaging AI Companies and implemented this technology into Clinical workflows at a single site. The imaging would be requested and completed as per the standard care pathway. The image once uploaded onto PACS was then reviewed by the Assigned Consultant and if they wanted this image to be reviewed by AI they would send the selected images in PACS and push them out to the AI Company. The image would then be processed by the AI Company and the referring consultant would be notified when this was completed. The consultant would then log onto the AI Platforms Dashboard and review their AI Read Image. The AI processed image did not feedback directly into PACs. To note- the consultants chose not to allow permission to send AI reads to the AI provider by other members of their team (SHO, Registrars, Radiologists ect), however this was an option offered by the AI provider. Similarly, the extended team did not have access to review the images on the AI Dashboard. The lack of integration back to PACS was a limitation of this product in the workflow, however the notifications were an acceptable short-term workaround. The turnaround time of AI Images was deemed acceptable due to the nature of the workflow and surgical planning timelines. The ‘number of clicks’ to send images for AI reads was deemed acceptable, however the frequency of these requests over time fluctuated, and it was noted some consultants were more “bought into” the technology than others. N.B I would consider if the technology is being implemented to A) solve a problem that currently exist in the clinical pathway B) Significantly improves clinical or operational outcomes, or C) Is considered as a ‘Nice to have’ The category into which the AI technology falls into will play a part in the benefits case and help to build a business case for the procurement of such a technology, providing the foundations of expected outcomes, stakeholder impact assessment and individual stakeholder benefits.
Expert 6:	Although I don’t have direct experience with the listed AI technologies, my familiarity with similar imaging tools provides insight into how they typically function: • The AI systematically analyses all MSK features in the image using advanced classification algorithms, identifying patterns and features indicative of vertebral fragility fractures (VFFs). • Once the analysis is complete, the AI predicts and flags cases that are likely to have VFFs. This allows radiologists to prioritize these scans for review. • With flagged images, the radiologist can focus their attention on the AI-marked regions, ensuring faster and more accurate diagnosis. • The tool acts as a "second opinion" and helps mitigate the risk of missed fractures, especially in opportunistic settings where VFF detection is not the primary goal. • AI systems often improve over time through feedback from radiologists, ensuring continuous enhancements in diagnostic accuracy and efficiency. Even without direct use of these specific tools, AI-based imaging systems generally follow a workflow that enhances the radiologist’s diagnostic process by identifying and flagging potential VFFs with high accuracy. This type of tool is essential for opportunistic fracture detection and reduces human error in clinical imaging settings.

Expert 7:	Yes, I have led a recent project at my trust implementing an AI tool in the reporting of chest X-ray images. This included educating staff within Radiology during the first phase of the project to ensure familiarity with the tool when reporting. This extended to staff outside of radiology during the second phase of the project. The AI tool has helped in triaging certain images for an urgent radiological report and helped to inform clinical decision-making. The implementation of the AI tool has had an initial impact on workflow until everyone was familiar with the use of AI and how to interpret the secondary capture images. [redacted]
Expert 8:	I have used Annalise (not sure which one), and I've also used Boneview by Gleamer. They both provide an additional image on PACS with a diagnosis and also an image with the abnormality highlighted.
Question 4b	If you have any experience with AI tools, to what extent is the AI model (the data it is trained on) information made available or explained to you as a user?
Expert 1:	N/A
Expert 2:	Again please see details from our paper below
Expert 3:	sorry I do not have any meaningful direct experience
Expert 4:	I use AI tools as a clinical researcher, and make sure I understand the data it is trained on and how it has been tested, trained and validated. I think clinicians would want reassurance that the tool is valid
Expert 5:	The original data sets were not made available to the Data Scientist who supported the technical implementation of these technologies. However, an overview of the demographic data and the performance metrics were provided (True Positives, True Negatives, False Positives, False Negatives and Receiver Operating Characteristics). The site I was working at though had a dedicated 'Sandbox' environment where it would initially deploy the AI and run through it own locally held anonymised data sets to determine if the technology performed as stated in our own local environment. The companies (SME's) would use provide us with any research completed/published for the team to review and evaluate. The would also be invited to present at the AI Board at the Trust where members would have the opportunity to ask questions and make recommendations prior to implementation.

Expert 6:	As a user, AI modelling datasets are typically informative, containing essential features required for accurate predictions. In the case of imaging for opportunistic detection of vertebral fragility fractures (VFFs), key features such as fractures and subtle structural changes are critical for effective AI predictions. The presence of detailed fracture-related features and other musculoskeletal (MSK) features should be made transparent and easily accessible to users. This ensures radiologists or reporters understand how the AI is flagging or highlighting abnormalities in the scans. Clear information about the AI's understanding of all MSK scans (e.g., anatomical regions analysed, types of fractures detected, and limitations) will greatly improve user efficiency. When users are aware of the full range of features being analysed, they can make better-informed decisions, reducing false negatives or unnecessary follow-ups. Providing such transparency enables a more collaborative relationship between the user and AI, where radiologists or reporters can trust the AI's predictions and work more efficiently to confirm or reject findings based on clinical expertise. Transparent and detailed information about the AI's dataset and the key features it is trained on is vital for effective user engagement. This approach ensures that users can fully leverage AI's capabilities while maintaining accuracy and clinical confidence in opportunistic VFF detection.
Expert 7:	The AI tool I am currently using is delivered into live clinical use, using the DICOM secondary capture methodology. This means that after the acquisition of a study, it will be analysed by the AI software and delivered alongside the original study to both the radiology PACS and Xero viewer. The AI tool is based on human input, its detection of an abnormality is based on a confidence bar, and the algorithms can be adjusted if there is an issue with false negatives / positives/sensitivity and specificity.
Expert 8:	It is not explained at all.
Question 4c	As this EVA is considering multiple technologies the EAG will attempt to summarise key technical aspects of each AI technology. Broadly speaking what technical features of the AI technology would be helpful to you clinically when considering their use in the opportunistic detection of vertebral fragility fractures?

Expert 1:	**1. Diagnostic Accuracy Metrics:** - **Average Accuracy Across Populations:** An overall measure of the AI tool's performance (e.g. sensitivity, specificity, PPV, NPV, and AUC) can help clinicians to assess its potential fallibility within the clinical workflow. However, similarly, this information needs to be presented to clinicians in a format which is easy to understand and apply to their clinical practice. - **Accuracy in Specific Populations, Subgroups and Infrastructures:** Stratified performance data across different: ▪ **Patient Demographics:** Age, sex, ethnicity, BMI ▪ **Patient Disease States:** Disease cohorts (e.g. osteoporosis, cancer patients etc) ▪ **Imaging Modalities:** CT, MRI, X-ray. ▪ **Imaging Protocols/Scanners:** Different scanner types, protocols etc. This information helps identify performance disparities and potential biases, enabling appropriate clinical interpretation and mitigation of false positives/negatives. **2. Interpretability and Explainability (Clinical Transparency):** Information related to an AI tool's ability to provide explanations of the decision-making process. This may include reporting of measured metrics such as o Vertebral height measurements (e.g. Genant classification) o Pixel Density or Texture o Comparison with Prior Imaging Probability metrics or uncertainty measures alongside AI outputs would be useful to help clinicians to gauge the reliability of results. **3. Workflow Integration and Usability** o **Alert Mechanism:** The ability of an AI tool to provide clear, actionable alerts for potential VFFs with severity grading (e.g. mild, moderate, severe) to prioritise further review. Ideally, this would provide feedback to the user regarding ongoing referral or further investigations. o **Integration with Reporting Systems:** Compatibility with PACS, EHRs and other NHS systems for seamless incorporation into radiology workflows. o **Report Enrichment:** Structured AI-generated comments that can be easily incorporated into clinical reports. **4. Robustness, External Validation, and Generalisability** o **Performance Across Imaging Protocols:** Data pertaining to accuracy across different scanners and protocols. **5. Technical Safety and Reliability** o **Error Management:** Data related to the availability of 'error modes' (e.g. recognition of false positives in degenerative disease) and strategies for human-AI disagreement resolution. o **Need for Ongoing Algorithm Training & Accuracy Consequences:** Frequency and mechanism for model

	retraining, ‘updates’ to the tools and whether either of these have an impact on model accuracy over time.
Expert 2:	Accuracy compared to human identification and different grades of fractures (grad 1-3).
Expert 3:	The effect on reporting time is important. It is important if/how the application indicates its confidence in its decision-making, the thresholding of this confidence as a positive or negative finding will alter the sens/spec profile of the software and so it is important the approach is reported in the literature. It would be helpful for software to categorise the severity or type of VFF to aid reporting. It is underestimated how effective the usability of the software is. If it is clunky or takes extra “clicks” engagement will be lower. Users will want the ability to switch it on or off easily. Some radiologists may not engage with it, for various potential reasons, which damages the business case for its use It will be important that referrers can also see the AI output/transparency/segmentation of vertebrae to support the radiology report.
Expert 4:	(1) the method of vertebral fracture identification the original manual annotations used; (2) some broad description of the technological method eg machine learning/neural network; (3) some quantification of diagnostic accuracy; (4) importantly, what the output is e.g. highlighting there is likely to be a fracture at a specific vertebrae/highlighting there is likely to be a fracture somewhere in the spine/highlighting there may be an abnormality at a specific vertebral level/writing text in the radiology report directly to the requester; (5) has it been user-tested within PACS; (6) the proportion of scans it identifies as having a vertebral fragility fracture. We think somewhere between 6 and 12% of older people have a vertebral fragility fracture. Some AI methods have a high false positive rate. Quantifying the number of scans it identifies as having a fracture would help understand the burden of work needed to second read and exclude the false positives.
Expert 5:	1) Integration with existing systems: The technology neds to integrate with existing workflows e.g. PACS’s systems or EPRS. This would make it easier for clinicals to access and review. Having it as a separate system with new user names, passwords logins etc, is not making the operational workflow process any more efficient and potentially mean the technology will not be utilised longer term. 2) Time of Referrals: if these opportunistic vertebral fractures are diagnosed sooner, patients will access services and therapy sooner. The AI read that has a postie finding of VFF may generate a better MDT approach and pathway for this category of patients. This in turn will potentially have a positive impact on patient outcomes including pain and returning to work (Social Economic benefits), Activities of Daily Living (ADL’s) and keep them active and thus potentially reducing the risk of other health implications or impacts of a sedentary lifestyle due to pain for example. 3) Educational Support, the technology can potentially be used to show the patient their images that might have been annotated by the AI. Many patients will not have seen an X-Ray and therefore showing them an image of their Vertebra that has potentially been annotated/highlighted/marked-up to show the area of the VFF may help patients envisage and understand their diagnosis. 4) If the AI technology can also provide information to clinical teams about the bone mineral density for example, it could also be used to help make clinical decisions around preventative measures and management moving forward.

Expert 6:	The technical features helpful to me clinically would be; Accuracy: The percentage of correct classifications (both true positives and true negatives) across all MSK scans, whether or not they contain VFFs. High accuracy ensures that the AI is effective at distinguishing between scans with and without VFFs, making it clinically reliable for general use in identifying fractures opportunistically. Sensitivity (Recall / True Positive Rate): The ability of the AI to correctly identify VFFs (true positives) out of all MSK scans that actually contain VFFs. High sensitivity is crucial for reducing the number of missed VFFs. In clinical practice, missing fractures can lead to delayed treatment and adverse outcomes, so the AI should detect as many fractures as possible. Specificity (True Negative Rate): The ability of the AI to correctly classify scans without VFFs (true negatives) out of all MSK scans that do not contain fractures. High specificity reduces false positives, preventing unnecessary follow-up procedures for scans that don't actually contain fractures. This helps clinicians avoid false alarms and conserve resources. Precision (Positive Predictive Value): The ability of the AI to correctly classify VFFs from all MSK scans it predicts as having VFFs. It tells you the proportion of true VFFs among all those flagged as positive by the AI. High precision ensures that when the AI identifies a VFF, it's likely to be correct, reducing the risk of unnecessary investigations or treatments for fractures that don't actually exist. F1-Score: The harmonic mean of precision and sensitivity (recall). This metric provides a balance between detecting VFFs (sensitivity) and ensuring that the detected VFFs are truly positive (precision). The F1-score is a crucial measure of overall performance, especially in cases where both false negatives and false positives can have significant clinical implications. A high F1-score indicates that the AI performs well in both detecting VFFs and reducing errors in classification. Robustness: The AI's ability to maintain consistent performance across a wide range of MSK scan types, including variations in image quality, position, and modality (CT, X-ray, MRI). Robustness is vital in clinical environments where image quality can vary significantly. AI systems that can accurately detect VFFs across diverse conditions (e.g., different types of scans or suboptimal image quality) are more valuable in real-world practice, ensuring consistent results regardless of the imaging challenges. Safe to use and complies with the regulatory bodies. It should be seen as an assistive tool, not an autonomous diagnostic tool, with human oversight being a key component of the clinical workflow.
Expert 7:	Confidence bar of how certain AI is of what it is has detected. A comment on the findings issued by AI which makes it clear to the clinician/ referrer that this does not replace clinical decision-making and is only a supplementary decision tool. A comment on limitations of the technology in differentiating between VFFs and non-fracture deformities such as Scheuermann's disease or Schmorl's nodes. This is important to help referrers to decide if the appearance are due to a true fracture.
Expert 8:	We just need to know that its accurate and it also needs to be able to highlight the area of concern on the image.

Question 5	The EAG plans on applying subgroups (where possible using the published literature) by image type and location/setting of imaging; this is due to potential differences in diagnostic accuracy and different pathway/costs in economics modelling. Is this clinically appropriate?
Expert 1:	Yes, this is clinically appropriate and recommended. It is important not to exacerbate health inequalities through use of AI tools in this context, particularly as different populations may experience variations in access to specific services or imaging modalities.
Expert 2:	Yes
Expert 3:	By “image type” I assume you mean imaging modality (such as x-ray, ultrasound, CT etc.). If so, yes this is definitely appropriate. I am less convinced that different referrers or imaging settings are important, as opportunistic VFFs should be followed up and at least considered for bone health screening in most cases regardless of imaging locations/setting/referral pathways
Expert 4:	Yes for image types e.g. CT or plain radiographs. I’m not clear the benefits of looking at where the setting of the image e.g. community/hospital is needed – they will be reported by the same clinicians
Expert 5:	Yes Location /Setting: The setting in which imaging is taken e.g. emergency departments, outpatient clinics, inpatient settings (spanning form ITU to ward-based care) will some variation in who will be reading of the report and therefore making clinical decisions regarding the patients care. For example, in ITU the report read by an intensive care consultant may focus on the text related to the lung fields when considering ventilation weaning plans and they may refer to VFF to other teams to review and Ax for a management plan, however this may be delayed due to other priorities in relation to the stabilisation and medical management of the patient. From a physiotherapy perspective in intensive care we would be waiting for the assigned named consultant to document they are happy for us to start mobilisation. In contrast a radiology report viewed in an outpatient setting in in isolation to the Ward MDT and thus referrals to Physiotherapy Outpatients will have their own time delays and waiting lists. In short, the AI may facilitate timely referrals and interventions, however the method and into which care pathways will vary depending on the location of the patient. Implementation: Approximately 90% of Trusts in England are using and EPR to hold patient notes, integrations of third-party technology providers however vary vastly and therefore there may be some differences in how the AI Reports are accessed and reviewed. If the implementation is in aligned with local PCS’s systems and EPRS the reports might be viewed in line with the clinical professionals reviewing all the patient notes and documents. However, if the implementation required the end user to access the AI reports on a cloud-based platform or via an alternative platform with user name/password etc, this could impact the frequency in which these AI reports are viewed if at all.

Expert 6:	The quality of imaging, determined by factors like pixel resolution and density, affects sensitivity and specificity for detecting VFFs. Since CT and MRI scans tend to provide more detailed images compared to X-rays, better detection rates may be achieved with these modalities. Subgroup analysis by image type allows for more tailored evaluations of AI performance for each imaging method. Different clinical settings (e.g., emergency departments vs. outpatient clinics) may influence how AI is used in practice. In emergency settings, faster image processing and quicker decision-making may be prioritized, while outpatient settings may allow more detailed analysis. By categorizing results based on setting, clinical recommendations can be fine-tuned for each scenario, leading to better patient outcomes. Subgroup analysis helps ensure that the evaluation is both clinically relevant and aligned with real-world practices. This comprehensive approach enhances the accuracy and reliability of recommendations for AI adoption across diverse clinical workflows, ultimately supporting better decision-making and more effective care
Expert 7:	Yes this is clinically appropriate as some image types will be more accurate at aiding AI to identify VF's. For example, a lateral CXR or spinal imaging will allow for better visualisation of the vertebrae whereas the CXR could potentially limit AI in recognising VF's due to exposure parameters and radiographic quality.
Expert 8:	Yes
Question 6a	If a vertebral fragility fracture is detected opportunistically, - what happens next? Does another staff member review the radiographic image (staff role, band, time taken to review image)?
Expert 1:	o There is considerable variability in approaches across NHS trust, radiology department, and imaging reporter. o For MSK-trained reporters, it is unlikely that the images would be reviewed by a second reporter. In this case, when the presence of VFFs is clear in the opinion of the MSK-trained reporter, the VFF would be included in the report and ideally flagged to the requesting clinician for them to action. o For non MSK-trained reporters, or in the case of uncertainty, it is more likely that the images would be reviewed by a second reporter (likely an MSK-trained senior reporter such as a specialist MSK radiologist). o A further possible pathway is that the report is issued to the requesting clinician with a VFF highlighted, and the requesting clinician (such as a Rheumatologist or other hospital specialist) then organises further discussion (perhaps as part of an MDT meeting) to gain further perspectives on the validity of the VFF and further imaging requirements.
Expert 2:	Depends on the AI system used. Some use radiologists to then 2nd read, which is the preferred option until we can become confident with the software.
Expert 3:	I don't think so, it is typically the responsibility of the reporting clinician to identify and mention the VFF in their report, and then up to the referring clinician to act on the report that a VFF has been observed. The exception is where the Trust has a FLS, in which case the patient details may go to the FLS, where an FLS staff member may contact the patient or refer them to DXA. FLS services may have varied protocols on which VFF cases they follow up on. Around 50% of Trusts in England have a FLS, but it 100% in Scotland.

Expert 4:	It depends. For example: if it was a consultant radiologist who spotted a vertebral fragility fracture on a CT done for suspected malignancy, for example and they are confident it is a vertebral fragility fracture, no-one else needs to review the image. If it was a consultant radiologist who is uncertain they might discuss with colleagues. If it was a trainee they would need to discuss with their senior team.
Expert 5:	Unable to comment
Expert 6:	The radiographic image is reviewed by a senior clinician or consultant radiologist, usually of band 6 or higher. The review typically occurs within minutes, though the time may vary based on the image's complexity and the reviewer's workload.
Expert 7:	If the patient is acutely in pain due to the VFF the radiographer will escalate the images for an immediate report (hot reporting – a Advanced Radiographer Practitioner provides hot reporting 12 hours 7 days at this trust) The Reporter will then decide if the patient needs onward referral to ED for example if pain is not controlled or patients has appearances of an unstable VFF which requires input from the spinal specialists/ trauma and orthopaedic. If the patient is relatively mobile and they are not complaining of high levels of pain/managed pain with analgesics, we refer the patient back to the referrer with a reporting code that alerts the referrer that we have detected VFF's which could be osteoporotic. We then recommend that the referrer investigates for other causes of fracture and excludes more sinister causes such as multiple myeloma or metastatic bone disease. The reporter will also add a code which asks the refer to FRAX the patient and assess their future risk of fracture and consider if a DXA scan would benefit the patient. This escalation process is for all diagnostic radiographers band 5 and above. They will review their images after acquirement for diagnostic quality and assess for abnormality – this will then determine if the patients is escalated for immediate report.
Expert 8:	It would presumably have to be reviewed at the same time as the examination is reported. If someone had to review the imaging a second time it would complicate matters significantly (additional resources to report a second time, report would need an addendum which would then have to be read a second time by whoever actions the report).
Question 6b	Would we broadly assume the same staff band, time is applied for re-review of an opportunistic detection of a potential VFF identified on X-ray or CT?
Expert 1:	Not in many cases. For example, in unclear cases or with those that are inexperienced with detecting VFFs, it is more likely that escalation will occur with re-review of VFFs being conducted by a reporter at higher staff band or with specialist experience (i.e. an MSK-reporter). It is also likely that a greater amount of time is used to review images in this case.
Expert 2:	Yes consultant radiologist
Expert 3:	I am not a reporting clinician, but again I would think second reporting of VFFs is rare. The referral team may re-review imaging and make clinical decisions in support with, or against, the radiology report.
Expert 4:	I think it would be critically important for any possible vertebral fragility fracture identified by an automated method to be reviewed by a clinician. This is mainly because some AI methods have a high false positive rate because they are based on a quantitative morphometric approach where degenerative change is erroneously classified as a mild vertebral fracture.

Expert 5:	Unable to comment
Expert 6:	The need for a re-review varies depending on the imaging modality. CT scans typically provide higher resolution and clearer images, which may reduce the need for additional reviews or processing time. X-rays, on the other hand, may produce less detailed images, often requiring more time for interpretation and potential escalation for a second opinion. This variability may influence the total time spent on reviewing each case and could potentially affect the staff band required depending on the complexity of the case
Expert 7:	Yes from an X-ray perspective as the radiographers, assess the images immediately after acquisition. The AI tool would need to be visible to the radiographers on acquisition of the image not just at reporting. CT images are reviewed by a CT radiographer but usually reported by a Consultant Radiographer. So the process in CT maybe different as not sure if the scanning radiographers will have time to assess all the scanned images/slices.
Expert 8:	Would need to be the same staff band yes. You do then have a dilemma as to whether the re-review is re-reviewing the whole image or just the vertebral fracture. If just the vertebral fracture it would be less time consuming than the initial report time.
Question 6c	Is the patient referred for another scan to confirm the presence of the VFF?
Expert 1:	This depends on the complexity of the imaging and certainty of the VFF presence. In my experience, presence of VFFs are stated by the reporter on the imaging report with recommendations for further imaging if required. However, further imaging usually relates to detecting the underlying process (e.g. a DEXA for BMD assessment in Osteoporosis), rather than confirming the VFF. In some cases, such as with multiple VFFs where there is concern over complications or need for surgical intervention (such as with vertebroplasty), further imaging may be suggested to provide further information to inform ongoing management (such as an MRI/CT in someone who has only undergone an x-ray). It is possible that this is requested automatically by the reporter, but often the reporter would relay recommendations to the requesting clinician to action, or they may be referred to specialist services such as Fracture Liaison Services (FLS).
Expert 2:	No, but referred for a bone density scan to confirm the severity of osteoporosis
Expert 3:	The patient should ideally be referred for a DXA scan, including a vertebral fracture assessment (VFA scan) as part of this examination. The patient should also have bone health screening using clinical algorithms such as FRAX (QFracture less often used but NICE approved). This is far less likely to happen in areas that do not have FLS services. In theory, a VFF identified on x-ray could be sent for higher fidelity CT (or less likely MRI), but I think this would be rare for an asymptomatic opportunistic VFF.
Expert 4:	Unlikely to be needed if it was originally found on a CT. If it was found on a plain radiograph, then maybe.
Expert 5:	Unable to comment

Expert 6:	The decision to refer for additional imaging depends on the initial findings. CT scans are generally sufficient for confirming VFFs due to their high resolution. However, if X-rays are unclear or inconclusive, a referral for further CT or MRI imaging may be necessary to confirm the diagnosis and assess the full extent of the fracture. If the AI has flagged a potential VFF and the clinician confirms it with confidence, additional imaging may not be required. Nevertheless, clinical protocols may still recommend follow-up imaging in specific cases to: • Ensure accurate assessment of fracture severity. • Identify additional fractures that may not have been detected initially. • Plan appropriate treatment, particularly in complex or high-risk cases. Ultimately, the decision is guided by the patient's clinical condition, the quality of the original images, and the clinician's judgment.
Expert 7:	Not necessarily, it depends if there is a concern that the fracture could be unstable/relatively acute, in which case we will perform a CT scan to assess the fracture characteristics. We would recommend a DXA scan if appropriate and if the patient can tolerate the scan to assess the quality of their bone density.
Expert 8:	It depends on how confident the reporter is on the first images. I would say that most cases will be clear cut. If there is any doubt, referral for dedicated x-ray of that area would be necessary.
Question 7	Are there any published UK audits or service evaluations of AI technologies used in opportunistic detection of VFFs you are aware of that would support the EAG assessment? Can you share links to any you think might be of interest please?
Expert 1:	I am not aware of any published UK audits or service evaluations of AI used in opportunistic VFF detection
Expert 2:	Yes : Ong T, Copeland R, Thiam CN, Cerda Mas G, Marshall L, Sahota O. Integration of a vertebral fracture identification service into a fracture liaison service: a quality improvement project. Osteoporos Int. 2021 May;32(5):921-926. doi: 10.1007/s00198-020-05710-8. Epub 2020 Nov 10. PMID: 33170309.
Expert 3:	The UK ADOPT study is the main one that I am aware of: https://theros.org.uk/blog/how-ai-is-helping-to-speed-up-spinal-fracture-diagnosis/ Again, it may be worth enquiring with the ROS and those who run the FLS database, whilst it does not explicitly indicate AI opportunistic VFF detection, they may know more.
Expert 4:	No response
Expert 5:	Unknown

Expert 6:	A national UK re-audit was published in 2022 by the Royal College of Radiologists (RCR) regarding radiology reporting of incidental osteoporotic vertebral fractures on CT studies. This provides insights that may support the EAG assessment. You can find the report at the following link: UK National Re-Audit - RCR Report (2022). Additionally, the National Library of Medicine published a study in May 2024 titled "Accuracy of an artificial intelligence algorithm for detecting moderate-to-severe vertebral compression fractures on abdominal and thoracic computed tomography scans." This study may also be useful for evaluating AI's accuracy in detecting VFFs. You can access the publication here: National Library of Medicine (2024) - AI Algorithm Accuracy Study.
Expert 7:	The Adopt study https://www.nanox.vision/wp-content/uploads/2024/11/Poster-for-WCO-Conference-UP-TP-10-MEGA2.pdf ASPIRE™: Using machine learning to detect undiagnosed fractures in patients with osteoporosis – Case study The University of Manchester RAD Magazine, 'AI detects previously undiagnosed vertebral fractures.' 11 December 2019. https://twitter.com/RADMagazine/status/1204750239753809920
Expert 8:	I am not aware of any
Question 8	Are you aware of any audits reporting compliance of NHS Trusts with The Royal College of Radiology guidance (Radiological guidance for the recognition and reporting of osteoporotic vertebral fragility fractures, May 2021).
Expert 1:	I am not aware of any audits reporting compliance of NHS Trusts with The RCR guidance.
Expert 2:	No
Expert 3:	No I am not.
Expert 4:	[redacted] [figure redacted] [redacted]
Expert 5:	No
Expert 6:	The most recent guidance from The Royal College of Radiologists (RCR) was issued in May 2021, with a follow-up re-audit conducted in 2022. These documents highlight standards for recognizing and reporting osteoporotic vertebral fragility fractures. However, I am not aware of any additional audits or evaluations beyond those already mentioned by the RCR.

Expert 7:	The Royal College of Radiologists (RCR) undertook a UK-wide audit in 2019 to evaluate organisational reporting infrastructure and specific patient-related reporting data in the diagnosis of VFFs on CT, where the thoraco-lumbar spine was included in the field of view but was not the area of clinical interest.[1] This audit was undertaken as a collaboration with the Royal College of Physicians (RCP) and the Royal Osteoporosis Society (ROS). [10] The audit involved 6,357 patients with an overall radiology departmental response rate of 63%; 1,362 (21.4%) of patients had a VFF on auditor review. The audit confirmed a lack of compliance with all audit standards – in particular pertaining to report comment on bone integrity, severity of fractures, use of recommended terminology (‘vertebral fracture’) and appropriate recommendations for further investigation/referral. We have conducted a local audit at this trust and discovered that we had poor compliance with a number of the recommendations. Assessment of compliance with the radiological Guidance for the Recognition and Reporting of Osteoporotic Vertebral Fragility Fractures (VFFs) in plain film (May 2022). 1. Howlett D C, Drinkwater K, Mahmood N, Illes J, Griffin J, Javaid K. Radiology reporting of osteoporotic vertebral fragility fractures on computed tomography studies: results of a UK national audit. Eur Radiol 2020; 30(9): 4713–4723. 2. Adams J, Clark E, Clunie G et al. Clinical guidance for the effective identification of vertebral fractures. London: National Osteoporosis Society, 2017.
Expert 8:	I am not aware of any.

Appendix D2: Questions sent to SCMs and experts 06/03/2025

#	Date responses received	Name, Affiliation
1	10/03/2025	[redacted]
2	No response	[redacted]
3	No response	[redacted]
4	08/03/2025	[redacted]
5	13/03/2025	[redacted]
6	No response	[redacted]
7	07/03/2025	[redacted]
8	07/03/2025	[redacted]
9	No response	[redacted]
10	12/03/2025	[redacted]
11	No response	[redacted]
12	14/03/2025	[redacted]
13	13/03/2025	[redacted]
14	No response	[redacted]
15	10/03/2025	[redacted]
	Meeting on 14/03/2025	[redacted]

Question 1	Do you think the above model structure makes sense?
Expert 1:	Broadly this structure makes sense. Some points to consider: The outcomes from AI analysed images aren't defined. Presumedly, AI would screen and provide information to the reporting radiographer/radiologist who would then review the images (if positive) and the pathway would follow the top flow in the model. It's not entirely clear to me what the two outcomes from the "Sensitivity (Test (true) Positive)" box are. Currently, the top outcome box says "True VFF Care Refer" which makes sense (i.e. true positive of fracture, refer to relevant VFF services). The lower box says "True No VFF Care Refer/further Investigation". Does this mean that one potential outcome would be that the radiologist/radiographer does nothing after the VFF is identified? There should always be some sort of response from this, either a referral or clear correspondence to a relevant clinician to action, or else there is a risk of a VFF being under/mistreated. It is likely this just needs to be made clearer with the format/wording.
Expert 2:	-
Expert 3:	-
Expert 4:	The standard care arm and sub-branches make sense, the AI arm does not. In my head this would mirror the standard care arm but just be "standard care with AI". I am not sure what the "[+]" next to standard care and AI boxes is meant to indicate.
Expert 5:	Partly, but… AI can still have false positives and false negatives
Expert 6:	-
Expert 7:	Not entirely. I would like to see PPV and NPV included as these values give a clearer idea of how the AI performs. Specificity in Radiologists will the majority of the time be 100% because a Radiologist will only call a VFF when they are certain it is a VFF and no other pathology. This is remembering they are the clinical experts and are the only clinicians who can formally diagnose a VFF in radiology imaging.
Expert 8:	To some degree.The AI Route ends with Image ineligible for AI Analysis, but wouldn't there also be a) Patients have Fracture, b) Patients do not have Fracture, and then the ongoing Positive/Negative analysis. Apologies if this is shown in a different image that has not be shared. The standard care route 'Patients do not have underlying fracture -> Test (false)Positive-> False VVF Care refer/further investigation' I think should also have '*' to indicate Delayed or missed presentation of VVF
Expert 9:	-
Expert 10:	Yes, it is straightforward in terms of reporting and onward referral A major advantage that AI can bring to clinical practice is its potential to function as a triage system. If the AI can become prioritized on the work list. Putting most urgent cases at the top of the reporting list, this will benefit patients. Guermazi et al (2022) Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence Radiology: Volume 302: Number 3—March 2022 n radiology.rsna.org
Expert 11:	-

Expert 12:	Yes, the model structure appears to follow a logical flow for opportunistic detection of VFF. It effectively distinguishes between standard care (radiologist/radiographer interpretation) and AI-assisted analysis. The sensitivity, specificity, and prevalence considerations are appropriately mapped. However, the model could clarify how AI influences the decision-making process—whether AI acts as a triage tool or provides a second opinion. Additionally, the handling of AI-ineligible images could be expanded upon.
Expert 13:	The diagram does not represent the AI pathway correctly for the opportunistic detection
Expert 14:	-
Expert 15:	The standard care branch of the algorithm above is comprehensive. However, the AI branch appears to be incomplete.
Meeting on 14/03/2025	[redacted]– Does the model identify VFF (refer or investigate). Referral pathway afterwards. Need more information to described. [redacted]- Patients with VFF (fragility) – implied diagnosis, recent fracture or not, cancer or not. What is it? Vertebral deformity plan – can't make diagnosis of fragility. Human interpretation Schmorl's node (previously there) exclude those. Alot of scoliosis/tilt (machine is not good at). Small proportion for cancer. What is causing it later down the line? [redacted]– deformity (don't know what's causing). Cancer pathological fracture. [redacted]– "identification of vertebral deformity". What pathways. [redacted]– deformity Yes/No. Other diseases diagnosed. Consensus: Consider a change to pathway to include confirmatory testing for VFF, consider adding another line coming out of do not have deformity for VFF for "other diseases" and true positive for underlying
Question 2a	In terms of opportunistically detecting Vertebral Fragility Fractures (VFFs) can we broadly assume that the prevalence of VFF is likely to be the same for patients referred for X-ray or CT?
Expert 1:	I don't think we can assume this for the following reasons: Different populations and numbers of scans: X-rays more likely to be completed for those with MSK complaints (hip pain, MSK chest pain, shoulder pain, back pain) and therefore more likely to capture the population at risk of VFFs. Similarly, CT is more likely to be completed in those being investigated for cancers. CT would be more sensitive at detecting subtle VFFs compared with x-rays, particularly with certain sequences/approaches. Using both modalities in the same population, you'd expect a higher prevalence of VFFs detected on CT.
Expert 2:	-
Expert 3:	-
Expert 4:	I personally feel like this is a reasonable assumption given you are looking for opportunistic findings for referrals for other clinical reasons. I guess you could argue that a CT population has a higher threshold for accepted referrals (due to higher radiation burden) and so may have higher general prevalence of co-morbidities/illness/age, and therefore increased risk of VFF. But on the flip side you may have more x-rays specifically for back pain. I still think it is a reasonable assumption for a model to assume uniform prevalence between modalities.
Expert 5:	I think this would be a pragmatic assumption, but it's likely that those who get referred for CT are more unwell so are likely to have a higher prevalence of VFF than those referred for X-ray
Expert 6:	-

Expert 7:	No, there tends to be younger patients attending standard x-ray departments compared to CT. Therefore we can expect VFF to be more prevalent in CT departments than X-ray.
Expert 8:	Broadly, however a CT scan and the planes that these images can be reviewed in will allow the reader to review the spine in 'finer slices' than an X-Ray Image, therefore the prevalence of VVF might be marginally higher.
Expert 9:	-
Expert 10:	This is a difficult one to answer as the most common examination in diagnostics is the chest X-ray (CXR) and quite often, the CXR answers the clinical question, so no further imaging is required. CT scans are usually used for patients that have other co morbidities and therefore these patients are more likely to have vertebral fractures. For example, patients with established pulmonary disease often have surveillance CT scans and are therefore more likely to get VFF's identified as they are also on medications such as steroids, which puts the patient at higher risk of sustaining VFF's. . A chest X-ray is a 2 D image and so this causes limitations in the ability to detect VFF's in comparison a CT scan can be used to create a 3D reconstruction and is therefore more likely to identify VFF's. In addition, breast shadows on a CXR can sometimes make interpretation of the lower thoracic spine difficult on an image. The lateral CXR is not a standard projection at all of acute NHS trusts, as most patients not go for CT imaging. The lateral CXR quite often demonstrates the thoracic spine to a high standard of radiographic quality.
Expert 11:	-
Expert 12:	Not necessarily. While both X-ray and CT scans are used for imaging, patients referred for CT scans often have more complex clinical presentations or suspected multi-system conditions, which may influence the prevalence of VFF in that population. X-rays are more commonly used as an initial diagnostic tool, and fractures might be underdiagnosed compared to CT, which provides higher-resolution imaging. Thus, VFF prevalence may be higher in CT-referred patients.
Expert 13:	for x ray prevalence is the same as these present as fractures on X-ray. Prevalence on CT is often not symptomatic so is a true opportunistic pickup.
Expert 14:	-
Expert 15:	Probably not. The prevalence of VFF in radiographs and CT differs, as one of the most common indications for x-ray of thoracic or lumbar spine particular in primary case is a query of a vertebral fracture. In other words, radiographs of thoracic or lumbar spine in primary case are predominantly (may be in 90%) requested by primary care clinicians for a suspected VFF. Therefore, the prevalence of VFF is probably higher than CT in the given age group. Opportunistic use CT for identification of VFF (asymptomatic) performed for other reasons would only demonstrate asymptomatic fractures.
Meeting on 14/03/2025	Not discussed
Question 2b	In terms of opportunistically detecting Vertebral Fragility Fractures (VFFs) can we assume that the sensitivity and specificity in standard care (reporting radiographer, no AI) for opportunistic detection of VFF is the same for X-ray and CT?

Expert 1:	Again, no I don't think we can. There are multiple factors that would influence the sensitivity/specificity including review protocol (i.e. local standardised reporting methods for reporting CTs/X-rays), the sensitivity of the method (i.e. CT more sensitive at detecting subtle VFFs as mentioned above), and the clinical vignette (CT requests more likely to have more information about the patient presentation which may either negatively or positively influence the reporter's approach to reviewing the image and detecting (or not) VFFs).
Expert 2:	-
Expert 3:	-
Expert 4:	I am not aware of any hard data on this, but I think this would be worth looking into before modelling. My suspicion would be that opportunistic detection may be higher on x-ray as the nature of x-ray would steer the reporter to specifically check for bony injury/pathology. Whereas CT chest, abdo, pelvis referrals are typically for soft tissue pathology and so VFF checks or commenting may be forgotten or considered irrelevant depending on the patient's situation. Checking for VFF on CT would ideally require specific bony windowing and sagittal/coronal MPRs which may not always be standard or prioritized in CT reporting for soft tissue indications.
Expert 5:	Maybe – the AI tool will have learnt from manual annotations, but those manual annotations would be done by a highly qualified expert. NHS standard reporting is likely to be less accurate than these highly qualified experts. On the other hand, AI sometimes can't run on images where there is movement artifact/a lot of degenerative change, so the reported sensitivity etc for AI will be much higher than for manual because it tends to run on the easy to interpret scans. Standard human reporting will report all scans including the difficult ones so although the sensitivity etc is likely to be lower, on a population level it may be more valid…. not certain if I'm making sense
Expert 6:	-
Expert 7:	CT has more available imaging to visualise VFFs compared to X-ray. Therefore incidental VFF findings will be flagged up more in CT than X-ray. However, you tend to get more X-rays specifically for VFFs, therefore % diagnosed will tend to be higher.
Expert 8:	Unable to comment
Expert 9:	-
Expert 10:	No in my opinion and from my experience reporting radiographers are well versed in identifying opportunistic findings on X-rays in comparison to consultant radiologists with CT scans – the reason for this is that Radiologists tend to report the CT to answer the clinical question which is quite often acute in nature and so there is some opinion that other incidental/opportunistic findings are not necessary to include in the report. Although there has been some improvement in this practice due to the RCR guidelines on reporting/identifying VFF's. I also believe that this dependent on the specialty of the radiologist, if their specialty is MSK, they are more likely to report VFF's than a neuroradiologist or GI radiologists. [redacted]
Expert 11:	-
Expert 12:	No. CT scans provide higher resolution and better visualization of vertebral structures than X-rays, meaning sensitivity and specificity are likely to be higher in CT. X-rays are more prone to false negatives due to image quality limitations and overlapping structures. Therefore, VFF detection in standard care without AI would likely have a lower sensitivity for X-rays compared to CT.

Expert 13:	as above the sense and spec will be different in standard care
Expert 14:	-
Expert 15:	Probably not. As it was eluded in the previous answer, an x-ray reporter (a radiographer or a radiologist) would specifically look for a fracture (addressing the main clinical question) while a CT reporter (usually a radiologist) would have to look for the presence of a vertebral fracture on sagittal reconstructs as a part of their desirable check list, although it is not mandatory.
Meeting on 14/03/2025	Not discussed
Question 3	Do you typically refer all the patients with detected VFF (positive test) to VFF management (e.g., further investigation, pain management, treatment)?
Expert 1:	Yes. Locally within our Rheumatology dept, most patients with VFFs (unless contraindicated such as in palliative patients) would undergo a DEXA (if not already completed) and likely be referred for IV Zolendronic Acid (IV Bisphosphonates) or other therapies in severe cases (Teriparatide, Denosumab, Romosozumab). Our threshold for treatment is low given the risk to QOL of future VFFs and the potential benefit on pain in the shorter term with therapy. In other areas with a FLS, I would expect all patients to be referred to this service.
Expert 2:	-
Expert 3:	-
Expert 4:	This is not my area of expertise, and so not talking from any personal experience. There could be cases where patient treatment could be considered not in the patient’s best interests, for example advanced dementia or palliative care situations. Or situations where the patient is unlikely to comply with medical or physiotherapy courses.
Expert 5:	No way – it’s approx. 12% of the older female population! Most can/should be managed in primary care with oral medications to reduce the risk of future fractures
Expert 6:	-
Expert 7:	-
Expert 8:	Unable to comment
Expert 9:	-
Expert 10:	See question 3a
Expert 11:	-
Expert 12:	Not always. Referral decisions depend on clinical factors such as the patient’s symptoms, overall health, fracture severity, and existing management strategies.
Expert 13:	No not typically. For x-rays some make a comment but it's not automatic. similarly on CT - there are regular audits and reminders to colleagues which increases reporting of VFFs but it is not consistent. It is not a decision to no referral but more an omission
Expert 14:	-

Expert 15:	I personally do refer the majority of newly identified VFFs to FLS, but I have the advantage of MSK Radiology sub specialization and have a special interest in spinal imaging, so I am probably not the best example in this respect. However, the following vertebral fractures I do not refer to FLS: • Traumatic (a major mechanism of trauma) • Pathological fractures (with known or obvious lesions affecting a particular vertebral body) • Those patient who had recently DXA on PACS (assuming that they are under care of FLS already) • When clinical details mention known osteoporosis or osteoporotic fractures.
Meeting on 14/03/2025	[redacted]– not identified, majority do nothing. [redacted]– young patient (would refer). 80-90% [redacted]– elderly, confirmed vertebral fragility (more than 1 fracture; 80+). Tend not to have DXA (not a viable) – have parental treatment without scan. [redacted]– not all trusts code, reporting radiographer – don't advise further management. Training issue. [redacted]– their trust immediate reporting, speaks to patient whilst still in department, alot of pain (acute) established radiographer reporting process. [redacted]– Variable standards. Consensus: agreement but with caveat of it being inconsistent across trusts
Question 3a	If the answer is no, how do you make a decision to no referral?
Expert 1:	-
Expert 2:	-
Expert 3:	-
Expert 4:	As above
Expert 5:	Symptoms, complicated, young, unusual
Expert 6:	-
Expert 7:	As long as the patient's VFF is indeed the result of fragility, and not pathological in the case of breast/prostate/myeloma cancer, then these VFFs should be referred for management and treatment. In cases the VFF is mild (Grade 1) there are no current indications that management is entirely necessary.
Expert 8:	Unable to comment
Expert 9:	-
Expert 10:	No depends if the patient's pain is managed well, there is no risk of an unstable acute fracture and they have good mobility. We refer the patient back to the care of their GP to manage the patient appropriately. We do however recommend a FRAX fracture risk assessment and a DXA if appropriate. We use a generic code in our radiological reports to communicate this. If the patient has had an acute injury, is in acute pain, cannot mobilise well or there is a risk that the VFF is unstable we will refer the patient to ED to get a trauma and orthopedic (T&O) opinion and a possibly a CT scan to assess the injury further. We are in the process of developing a direct pathway so that these patients can be referred directly to T&O without the need to attend ED.
Expert 11:	-
Expert 12:	If the VFF is an incidental finding with no symptoms. If the patient is already under treatment for osteoporosis or a known vertebral fracture. If the patient is frail with limited treatment benefit. If the fracture is old and stable, with no current impact on health.

Expert 13:	-
Expert 14:	-
Expert 15:	Please, see above.
Meeting on 14/03/2025	Not discussed
Question 3b	What are the most important factors that you consider in making decision to no referral?
Expert 1:	N/A
Expert 2:	-
Expert 3:	-
Expert 4:	Again I do not have first hand experience with this but would consider how debilitating and how much pain the fractures are, and what the likelihood of compliance is for treatment. If there is neurological symptoms is important to consider.
Expert 5:	Symptoms; or if I'm considering medications to reduce risk of further fracture and am wondering about a second line medications to reduce the risk of future fractures – a VFF means their fracture risk is higher than those with some other types of fracture so a second line agent may be justified
Expert 6:	-
Expert 7:	If the patient will benefit from treatment at all. If the patient has a VFF and is on palliative care then it would be inappropriate to refer to other specialist if the prognosis is shorter than the benefits of treatment. We do not want to burden palliative patients with another disease towards the end of their life or if a patient has a new cancer diagnosis. I refer to Ong T, Sahota 2020 (doi: 10.1007/s00198-020-05710-8 PMID:33170309) for their VFF screening pathway.
Expert 8:	Unable to comment
Expert 9:	-
Expert 10:	An unstable, moderately/severely displaced fracture, a recent acute injury, the patients pain is manageable or well managed with analgesia, the patient's mobility and possible neurological deficit – these are indications to refer the patient urgently for further assessment/action. We would definitely not refer the patient if the fractures are demonstrated on previous imaging but we would recommend a fracture risk assessment, investigate for other causes of fracture and a DXA scan if appropriate.
Expert 11:	-
Expert 12:	Symptomatic vs. asymptomatic fracture. Evidence of previous fractures and ongoing treatment. Patient comorbidities and frailty. Risk-benefit assessment of additional investigations/treatment.
Expert 13:	-
Expert 14:	-
Expert 15:	Please, see above.
Meeting on 14/03/2025	Not discussed
Question 4	When make a decision to refer the patients for extra assessment or receiving VFF care, what type of services (e.g., further diagnosis, future risk assessment, pain management) would the patient receive?

Expert 1:	In the case of a FLS being available, it's likely they would be referred to this service. For those without an FLS, most patients would be either referred to Rheumatology or Endocrinology depts (practice is variable depending on the hospital) then, the choice of further management/investigation would depend on the patient. Most would undergo investigation for Osteoporosis (through DEXA usually) unless either contraindicated or recently completed. Most would undergo a FRAX assessment with application of NOGG guidance following DEXA. As mentioned above, locally we would tend to treat patients with IV Bisphosphonates or alternatives (as described above) for VFFs given benefits over PO. In addition, they would likely be prescribed analgesia and may have further investigations performed if there were concerns over secondary causes or pathological fractures.
Expert 2:	-
Expert 3:	-
Expert 4:	If the Trust has a fracture liaison service, the patient should be identified by the FLS and offered a bone health assessment. This would constitute a DXA scan and screening questionnaire of risk factors (including FRAX assessment). Decision making would then follow NOGG 2024 guidelines (particularly section 8). If there is not a FLS, then I think care could vary. Generally speaking, it would be up to the referring clinician to consider bone health assessment based on the radiology report mentioning a VFF. In reality it is likely that this is not acted upon enough.
Expert 5:	Pain management, physio, osteoporosis services in secondary care, vertebroplasty rarely
Expert 6:	-
Expert 7:	Patient likely to receive: DEXA/QCT bone density assessment, Referral to Fracture Liaison Service or Metabolic Bone Service if a complex case Falls assessment, Blood Tests/Bone Profile tests (VitD, PTH, Calcium, AlkPhos)
Expert 8:	As a Physiotherapist we would see patients from this referral potentially in an outpatient or community setting or in a Falls prevention setting. The aims of Physiotherapy would be to prevent the list of adverse outcomes listed in point 7. We might also work closely with our Occupational Therapists in these setting to support the patient's gaining independence with their activities of daily living.
Expert 9:	-
Expert 10:	Further imaging – CT/MRI Fracture risk assessment – usually completed by GP/Nurse practitioner DXA if appropriate If they have osteoporosis on the DXA scan, they maybe referred to a specialist osteoporosis clinician for assessment for parental treatment. GP can refer the patient to MSK physio services if appropriate
Expert 11:	-
Expert 12:	Referral typically leads to: • Further diagnosis: DEXA scan for bone density assessment, MRI if needed to assess fracture age. • Future risk assessment: FRAX (Fracture Risk Assessment Tool) score, osteoporosis screening, fall risk evaluation. • Pain management: Analgesics, physiotherapy, spinal orthoses if necessary. • Treatment for bone health: Bisphosphonates, vitamin D, calcium supplements, or anabolic agents for osteoporosis. • Specialist referral: Rheumatology, endocrinology, orthopedics, or pain clinics if required.
Expert 13:	They would be go for a FLS assessment and then on to a dexa scan and then care

Expert 14:	-
Expert 15:	DXA, blood tests (including bone and hormone profile, myeloma scree etc), Pain management including an option of vertebroplasty for those whose pain is not controlled by analgesics or pain a significant impact on their lifestyle, FRAX score, spinal referral, secondary prevention of fragility fractures by administration of medicine. All these are probably better managed by FLS if available or a GP.
Meeting on 14/03/2025	[redacted] – Wouldn't refer to MRI to determine age of fracture (long waiting list); not done in Cambridge routinely. [redacted] – not MRI to date a fracture (can use 1% in sensitivity analysis) [redacted] – don't consider injection treatment for pain management (until MRI). [redacted] – VFA for more imaging Consensus: if we are careful with wording and use MRI in a scenario analysis (1%) we can include there
Question 5	What is the usual pathway following a false negative (has VFF but test negative) diagnosis of VFF? How would the process look like?
Expert 1:	The pathway here would be entirely variable depending on the referring clinician's approach. Here are some options: The patient would re-present to healthcare (either acutely to A&E or to their GP) with recurrent/persistent back pain. This may lead to further investigation (e.g. CT or MRI if X-ray performed) or referral to secondary care (usually either Rheumatology or Orthopaedics). The clinician would question the report if it doesn't align with the clinical picture (i.e. patient has clear features of a VFF clinically but nothing reported). This is of course more difficult in opportunistic detection scenarios given the scan was performed for other reasons. In this case, and we do this a fair bit in Rheumatology, we would discuss the images either within a Rheumatology-Radiology MDT, or with the reporting radiologist/radiographer, or with an MSK radiologist to gain consensus on the results. This may lead to a repeat scan or different imaging modality.
Expert 2:	-
Expert 3:	-
Expert 4:	-
Expert 5:	Nothing – if no VFF is reported then the pathway ends. If they have symptoms they may be offered pain killers, physio etc
Expert 6:	-
Expert 7:	There are no formal pathways for false negative VFFs. If a patient has a VFF missed, then it is unlikely to come to clinical attention until their next radiology appointment (providing the imaging includes the VFF). A VFF might be picked up again if the radiological imaging is reviewed a second time during audit or other investigations, whereby the suspected fracture will have to go to discrepancy review as a Radiological MDT meeting. One other possibility is that the VFF will be picked up in audits of radiological reporting. The Royal College of Radiology has an annual vertebral fracture audit which might pick up any missed VFFs. All of which will have an addendum added to the report and reported to the patient due to Duty of Candor.
Expert 8:	Unable to comment
Expert 9:	-

Expert 10:	[redacted]. These cases should be dealt with as a Radiology discrepancy – the patient has a VFF but it has not been identified on the image. This should be shared via peer review / REALM Radiology Education and learning meeting were all diagnostic discrepancies are discussed. These cases can be catergorised such as observer error, interpretative error/ equipment/ radiographic quality error. So cases would be submitted to REALM for discussion. An addendum would be added to the original report to make it accurate following REALM and the consensus of opinion. If there is a duty of candor and the discrepancy would have changed patient management/treatment we send the patient a letter and apologise for the mistake.
Expert 11:	-
Expert 12:	If a VFF is missed initially, the patient might present later with worsening pain, postural changes, or another fracture. The typical process includes: • The patient returns to their GP or primary care provider due to ongoing symptoms. • If suspicion persists, the GP may request another X-ray or CT scan, leading to delayed diagnosis. • If osteoporosis is suspected, the patient may undergo a DEXA scan even without an explicit VFF diagnosis. • If persistent pain is present, a referral to orthopedics, rheumatology, or pain management may occur.
Expert 13:	Unless there is another scan or symptom it would not be picked up
Expert 14:	-
Expert 15:	Probably it is a missed opportunity. The images might be reviewed in the future by a radiologist while reporting another study and making comparison between a current exam and a previous one. Another potential encounter with a study is when a radiologist prepares for a MDT meeting and reviews previous relevant images including a study with a false negative result. Also, if a patient remains symptomatic, a GP or another clinician asks to review a study of concern or to give a second opinion. Otherwise, another opportunity to identify a missed fracture would be when a patient has another x-ray or CT.
Meeting on 14/03/2025	Consensus: All in agreement
Question 5a	How long does it take on average for a patient to be picked up again from another mean (e.g., they go to their GP and have another CT scan)?
Expert 1:	Again, this varies depending on the patient. On average, I would guess that this could be 3-6 months due to: • Patients may struggle to gain another appointment with their GP or hospital doctor, this may take 2-4 weeks depending on locality. • Repeat scan requests can take 6-8 weeks to complete, sometimes longer depending on the modality. • Referral to secondary care (if this is performed instead of further imaging) can lead to a delay of at least 3 months before the patient is seen. In the circumstances where the clinician questions the report and re-reviews with colleagues, this would be shorter (could be as little as one week). In the circumstances where the patient does not seek medical input after the initial scan, this could be years before “healed VFFs” are detected.
Expert 2:	-
Expert 3:	-

Expert 4:	In this situation the opportunity to identify the VFF has been missed. I think it would be unlikely to be picked up from imaging by other non-reporting clinicians/referrers. It is unlikely that the VFF would be picked up unless they start to experience back pain symptoms (which are often not initially referred for imaging but rather pain relief/physio). I suspect many VFFs are not picked up at all, or not until after additional low impact fractures are experienced.
Expert 5:	I don't know, and I don't think there's anything published about this
Expert 6:	-
Expert 7:	Impossible to say. Some have back pain for a short while and never go back to their GP. Some have chronic backpain but do not want to burden GP practices as the patient feels "back pain is just something we get when we are older". Others the pain is unbearable and will attend A&E and get picked up straight away. There is not clear average amount of time for a VFF to be picked up again.
Expert 8:	Unable to comment
Expert 9:	-
Expert 10:	This is dependent on the reporting turnaround – our current reporting target for a GP report is 5 working days unless related to suspicion of cancer. Direct access means there is quicker turnaround for modalities such as X-ray / CT and so some trusts have a quicker turnaround. This is also dependent on the GP reading and auctioning the report.
Expert 11:	-
Expert 12:	This varies based on symptom severity and healthcare access. Some may be reassessed within weeks to months if pain worsens. If initially asymptomatic, diagnosis may be delayed for months or even years until another fracture occurs or symptoms develop.
Expert 13:	Following false positive - return to normal care
Expert 14:	-
Expert 15:	The timeframe of picking up a fracture again is varied, multifactorial and difficult to define depending whether a patient remains symptomatic, develops new symptoms requiring imaging or so on). Assuming that the highest risk for development another VFF is the first 24 months and if the second fracture becomes symptomatic, it deduces that 2 years might be an arbitrary figure for the second symptomatic VFF. Over the past 2 decades the Art of Medicine has undergone a dramatic alteration with the heavily weight on imaging/screening for a pathology due to a number of reasons. What comes of it is an average person has more CT scans with all collateral damages now than a few decades ago with an unintentional increased opportunity to detect silent VFFs.
Meeting on 14/03/2025	[redacted]- Even if they get an appointment, they are unlikely to have the test repeated especially if the GP has a test result that is negative, typically they would wait up 12 months if not longer [redacted]– I agree [redacted]– Agree, those who are older expect to get back to pain so often don't report. I don't think it is possible to say Consensus: Agreement
Question 6	What is the usual pathway following a false positive diagnosis of VFF?
Expert 1:	The same as with a True Positive – The patient would be referred for further investigations or review by FLS/secondary care. It may then be that, when undergoing a repeat scan or further assessment by a specialist, the VFF reported is queried or images are re-reviewed by an MSK radiologists who may alter the report. It may also be that this doesn't occur, and patients receive treatment inappropriately.
Expert 2:	-
Expert 3:	-

Expert 4:	In my view, the implications of a false positive may not be as severe as in other pathways. In an asymptomatic patient, if there is an FLS the patient would be referred for further evaluation and osteoporosis may still be detected even if the VFF was called incorrectly. If not osteoporotic then there was some burden/stress of the bone health assessment, but probably equal reassurance that they are not osteoporotic. If symptomatic, the patient likely received pain relief, which may still be helpful even if not caused by a VFF. Unnecessary physiotherapy or bracing may be indicated. More serious vertebroplasty or surgery is unlikely as these are for more advanced cases which are less likely to be false positives (unless perhaps mistaken for a serious pathological fracture, which I would still think is rare).
Expert 5:	Depends – see answer to 3 – if they're symptomatic they are likely to get symptomatic treatment. They may be started on a medication to reduce risk of fracture.
Expert 6:	-
Expert 7:	A false positive reading by a radiologist is very rare. As they are the clinical experts at reading and calling a VFF, we have to assume they are correct. If it is strongly assumed a VFF has been incorrectly called by a Radiologist, then the case will be made at a Radiological MDT where multiple radiologists will ultimately decide the case.
Expert 8:	If the aim of the AI is for the incidental findings of VVF and no VVF has been detected the image that was taken should be reviewed for the primary purpose of it being taken. My opinion would be to look back at the presenting complaint (e.g. back pain) and rule out any other pathology for causing this (e.g. Kidney Stones,) then address this presenting complaint. If the imaging is NAD on the report, and the presenting complaint requires further investigations (i.e. pathology/therapy) these referrals should then be made.
Expert 9:	-
Expert 10:	Same as above the case will be discussed at REALM/ or with an MSK Consultant Radiologist and an addendum would be added to the original report to make it accurate. If there is a duty of candor and the discrepancy would have changed patient management/treatment we send the patient a letter and apologise for the mistake.
Expert 11:	-
Expert 12:	If a patient is incorrectly diagnosed with a VFF (false positive), the usual pathway involves: • Further imaging: A follow-up X-ray, CT, or MRI may be requested to confirm or rule out the fracture. MRI is particularly useful for distinguishing old from new fractures. • Clinical reassessment: The patient's history, symptoms, and risk factors for osteoporosis are re-evaluated. • DEXA scan: If osteoporosis is suspected, a bone density scan may be ordered even if the fracture is ultimately ruled out. • Specialist referral: Some patients may be referred to rheumatology, endocrinology, or orthopedics for further assessment. • Unnecessary treatment risk: If not identified as a false positive early, the patient may be prescribed osteoporosis treatment, pain medication, or lifestyle changes they don't actually need.
Expert 13:	Following false positive - return to normal care
Expert 14:	-
Expert 15:	The patient is referred to FLS for further management +/-spinal referral.

Meeting on 14/03/2025	[redacted]– on unnecessary treatment risk I don't work in DEXA outside of reporting images, but my guess that they would go on osteoporosis treatment if they were diagnosed so I don't think it is bad thing. Trying to do a DEXA in patient groups with spinal diseases it becomes challenging to scan, we don't routinely scan those over 75 so we often just treat the patient without the scan. [redacted]– saying that if someone is in pain we would pain manage them anyway, the bisphosphonates might be the only one to be wary of. AI will lead to false positives no matter how good it is as they don't want to miss anything, but if you desensitise it then it runs the risk of missing [redacted]– Agree with conflicting guidelines, one arm says treat, but anabolic's say not to prescribe pain relief without DEXA Consensus: agree with caution on false positives
Question 7	What are the main adverse outcomes of missed or delayed diagnosis of VFF?
Expert 1:	Progression to further fractures (1 in 5 with a vertebral fracture will sustain another within a year if left untreated). Chronic pain, disability and loss of function. Spinal deformity incl height loss, kyphoscoliosis. More severe complications longer term – respiratory or GI complications. Increased mortality due to the above. Psychosocial disruption, impact on work/wellbeing and mental health. Wider system consequences – Increased healthcare access (hospital admissions, referrals) leading to increased NHS costs, higher care dependency due to clinical consequences.
Expert 2:	-
Expert 3:	-
Expert 4:	VFFs are a strong risk factor for further low impact fractures, including common neck of femur fractures, so missing opportunistic VFFs likely has a large economic effect in future fracture care. Treatment for VFF and osteoporosis is very likely to be cost effective: but the short term relatively low cost of increased treatment is often not acceptable despite the larger benefits, because these are longer term (ie future fractures avoided).
Expert 5:	Lack of starting medications to reduce risk of future VFFs/hip fractures, and all the associated adverse outcomes from these e.g. HRQoL, mortality and costs from secondary/social care
Expert 6:	-
Expert 7:	There is a massive increase in risk of secondary fracture after the first VFF. Up to 20% of women with a VFF will have a second fracture within a year. Therefore finding the first VFF is critical to start treatment as early as possible to prevent the subsequent fractures from happening. There is also a large increase in morbidity and mortality associated with VFFs and delaying treatment or delayed diagnosis will further increase these risks. https://doi.org/10.1016/j.beem.2008.06.001
Expert 8:	Acute pain, leading to Chronic Pain, Altered Posture, Muscle imbalance and affecting GAIT pattern and balance, Leading to an increased risk of falls, Reduced Mobility, Decreased Cardiovascular fitness, Reduced Quads muscle mass and strength and endurance, Reduction in Activities of Daily Living
Expert 9:	-
Expert 10:	[redacted] Increased pain and limited mobility for the patient. Risk of neurological deficit if the fracture is unstable if acute. Decrease in self impendence and reliance on others if loss of mobility High risk of sustaining another fragility fracture such as hip fracture is not assessed/ treated for osteoporosis

Expert 11:	-
Expert 12:	A missed or delayed VFF diagnosis can lead to: • Progression of fractures: An untreated VFF increases the risk of further vertebral fractures. • Chronic pain: Delayed diagnosis may lead to persistent, poorly managed back pain. • Spinal deformity (kyphosis): Multiple untreated fractures can result in height loss and a hunched posture. • Reduced mobility and quality of life: Patients may become more sedentary due to pain and discomfort. • Increased fracture risk elsewhere: Patients with untreated osteoporosis are more likely to sustain hip and other fractures. • Higher healthcare costs: Delayed diagnosis can lead to more complex and expensive treatments later. • Increased mortality: Studies suggest that untreated VFFs are associated with higher mortality due to complications like immobility, respiratory issues, and cardiovascular strain.
Expert 13:	Main outcomes are increased morbidity and mortality. I refer to some of the data in this Scottish paper https://shtg.scot/media/2511/20241115-dfls-imto-v12.pdf
Expert 14:	-
Expert 15:	Development of another VFF or other fragility fractures and a kyphotic deformity.
Meeting on 14/03/2025	[redacted] – additional 12 GP per year (with fractures). [redacted] – increases risk of falling (changes gait) Consensus: agree and consider risk of falling due to kyphosis
Question 8	Based on your experience, would a 1-hour training session be enough to familiarize the radiologists with a new AI technique in radiography?
Expert 1:	It’s difficult to comment on this given I am not aware of the complexity and implementation challenges related to the technologies being assessed. If the AI system’s user interface is simple with good integration with current systems (PACS/ICE/EHR etc) and minimal specialist input required from the user, then a 1hr training session would be sufficient.
Expert 2:	-
Expert 3:	-
Expert 4:	I think initially yes one hour is enough to explain how to use and report findings based on the software, but a lot of onward auditing and monitoring would be required.
Expert 5:	-
Expert 6:	-
Expert 7:	Yes it should be enough time. If the AI takes longer to familiarise then it is unlikely radiologists will engage with the AI as it could add significant time to their reporting.

Expert 8:	This will depend on how well the AI Technology has been embedded into the current clinical pathway. If the technology is interoperable with a PACs system and it purely a case of selecting the images to send to the AI software and the return report and highlighted images are returned to a PACs system, or even better if this process is done seamlessly in the background. This would mean an hour-long training session should be sufficient, to cover how to view the AI annotations and where to find the report. An SLA should be agreed to determine the level of support for any incidents where the AI Technology is 'offline' or support for clinicians when errors occur. However, if the technology is not interoperable with a PACS program, and therefore the users have to export, upload and send images across to the AI technology. If the reports are also returned to a separate could based platform, this will also require more time not only for training but for user access passes (usernames/ Passwords/Troubleshooting access) to be set up and for stakeholder engagement in the new ways of working to be understood (winning hearts and minds, to engage in this new way of working).
Expert 9:	-
Expert 10:	As an introduction, I would agree a 1-hour training session would be sufficient with links to other resources, so that one can have a fresher when required.
Expert 11:	-
Expert 12:	No, a single 1-hour session is unlikely to be sufficient for full competency. While it may introduce the AI system's interface and basic functionality, radiologists typically need: • Hands-on practice with real cases to understand how AI integrates into workflows. • Discussion of strengths, limitations, and biases of the AI tool. • Guidelines on when to trust AI outputs vs. relying on human interpretation. • Follow-up training to reinforce learning and address user challenges.
Expert 13:	A one hour session would be enough to train radiologists. The sessions can just be run as an on demand webinar.
Expert 14:	-
Expert 15:	Yes, I think it would be suffice to have an idea about the use AI in radiography.
Meeting on 14/03/2025	Not discussed
Question 8a	How many of these sessions need to be run to cover all the radiologists in your hospital per year?
Expert 1:	This isn't necessarily related to the number of radiologists (given you could theoretically train all radiologists at the same time/with an e-learning solution) and is more related to time required for the training/fitting this in around clinical provision. Locally we have approximately 40 consultant radiologists and many more SpRs/support staff. You'd likely need at least 10 sessions per year, although this could be delivered in-person initially and then through e-learning in subsequent years for 'refresher' training. This again depends on the user interface and complexity of the AI system integration.
Expert 2:	-
Expert 3:	-
Expert 4:	I could not comment on this based on my current role. I would generally make the point that trying to get reporting clinicians together for training would be very difficult and would require multiple staggered sessions. Some reporting staff may inevitably be reluctant with the idea and may not engage with training or use of the software.

Expert 5:	-
Expert 6:	-
Expert 7:	We a large staff of Radiologists and therefore would require many sessions to cover all staff members. An hour long training session is a lot of time off for radiologists to find as a group. We also have a large student population in radiology and therefore we would potentially require yearly training sessions for trainee radiologists to understand the AI.
Expert 8:	I work at NHS England Full time, and there are approximately 3,377 clinical radiologists working in the NHS in England. In 2020, there were 1308 applications for 311 specialty training places. I would estimate from my Physiotherapy Bank Work at a Private Hospital in Poole that 2-3 sessions per year would cover the team, this would allow for BAU and shift working staff to access training slots.
Expert 9:	-
Expert 10:	This dependent on working patterns/shifts – I would say at least 2-3 sessions to try to capture the radiologist/radiographer work force.
Expert 11:	-
Expert 12:	This depends on the size of the radiology team. A large hospital might require multiple sessions (e.g., 5–10 per year) to cover different shifts and specialties. Some hospitals may opt for small group training or online modules for better accessibility.
Expert 13:	-
Expert 14:	-
Expert 15:	As our NHS Trust is a teaching hospital and we have a group of radiologist-trainees who rotate every 3-4 months, 3 sessions a year would be ideal for them. However, for the vast majority of permanent staff a session per year sounds adequate.
Meeting on 14/03/2025	Not discussed
Question 8b	Do you usually need an update training for a new AI device?
Expert 1:	can't comment directly on this given lack of experience. However, I would suggest that the answer for this would be "yes".
Expert 2:	-
Expert 3:	-
Expert 4:	I don't think much further training on the logistics of the software will be required. But monitoring of the software performance and identifying its perceived weaknesses is vital. Some models will "evolve" as they learn from continued cases and so monitoring of performance will be ongoing and essential.
Expert 5:	-
Expert 6:	-
Expert 7:	Not unless it is a different AI technology.
Expert 8:	I don't have any direct experience of this for AI devices however for digital technologies being used to support medical services we do provide update training for new digital technologies. This has ensured effective use of the digital technologies especially when new features have been added. It has also helped to reduce errors due to incorrect use and application. I would assume this would be the same for AI Technologies, especially as they are exposed to more data sets to learn form and new features are added.
Expert 9:	-
Expert 10:	Yes
Expert 11:	-

Expert 12:	Yes, updates in AI software, algorithm changes, or new features often require refresher training.
Expert 13:	-
Expert 14:	-
Expert 15:	Yes. An update for a new AI device would require a sort of teaching/training in various formats.
Meeting on 14/03/2025	Not discussed
Question 8c	If so, how many update training do you run per year for a specific AI device?
Expert 1:	Difficult to comment again, I would guess perhaps one/year/clinician.
Expert 2:	-
Expert 3:	-
Expert 4:	Post market surveillance of diagnostic medical imaging software is very topical and there are many suggested requirements as part of best practice once a software is installed. It may be worth considering these documents: • https://www.rcr.ac.uk/our-services/all-our-publications/clinical-radiology-publications/ai-deployment-fundamentals-for-medical-imaging/ • https://www.gov.uk/government/publications/medical-devices-post-market-surveillance-requirements/the-medical-devices-post-market-surveillance-requirements-amendment-great-britain-regulations-2024-guidance-on-implementation#:~:text=The%20post%2Dmarket%20surveillance%20report,this%20regulation%20coming%20into%20force
Expert 5:	-
Expert 6:	-
Expert 7:	-
Expert 8:	For the previous digital technologies, I provided quarterly updates in its first year and then these reduced to annually, with options to request additional training from teams as per their requests, e.g. In Service Training Session for specific teams. For the Application of the AI technologies training was completed in the sandbox environment the technology was fist integrate to and then again once it went live. Ad-Hoc training sessions were delivered to a specific stakeholder group using the technology over 1 year or when the AI provider had pushed out a new update.
Expert 9:	-
Expert 10:	As above, I would suggest 2-3 sessions
Expert 11:	-
Expert 12:	Typically, 1–2 update sessions per year per AI system, depending on how frequently the software evolves. If significant changes occur, additional workshops or hands-on refreshers may be needed.
Expert 13:	-
Expert 14:	-
Expert 15:	Sorry, I am not clear with the question.
Meeting on 14/03/2025	Not discussed
Question 9	On average, how much additional time do you think will be needed for reading and interpreting of AI reports versus standard care per scan?
Expert 1:	15 minutes assuming the systems are clear, transparent and interpretable.
Expert 2:	-
Expert 3:	-

Expert 4:	I suspect this requires much more investigation and is variable in the evidence base. I suspect in VFF negative cases (or at least when AI says no VFF) reporting time is unlikely to be very different as the reporter will not spend any additional time looking for VFF then they normally would. In VFF positive cases I suspect reporting cases will take longer as the reporter will need to check not only if they agree the VFF is present, but also the severity and implications of the VFF, and if there are multiple VFFs, or additional VFFs not identified by AI. I suspect the evidence base will be mixed on this as vendor-sponsored studies will be keen to try and evidence improved efficiencies. It is a difficult question because current practice is very poor at identifying VFF. So naturally increasing VFF pick up will increase reporting time because some reporting staff were forgetting to look for VFF at all, but it is an improvement for patients. This is different to other AI, for example fracture detection software where reporting clinicians are already reporting to a high standard, so the comparison of reporting time is more in context of a marginal gain in accuracy. But with VFF it feels more “disruptive” as reporting clinicians are likely to be reporting many more VFF cases that they may have otherwise not been aware of, which would naturally take longer on balance.
Expert 5:	-
Expert 6:	-
Expert 7:	Depends entirely on the AI output. Some AI outputs will show limited imaging to highlight the fracture, some will provide many images to highlight the VFF. The downside to this is that images output from AI have a caveat that says “Not for diagnostic purposes” for each reformatted image. Therefore the radiologist will have to review original images in their PACS system to determine if the AI Positive finding is indeed a fracture. This could take an additional 2-3 minutes for each positive VF case. There will also be cases where there is unwanted time taken to rule out a VFF if the AI provides a False Positive (FP) result. FP results can be very burdensome and frustrating to Radiologists if the AI is presenting fractures that are not present. Again, there needs to be time taken to review the images on the original PACS system, and AI visualization systems are not usually approved for diagnosis
Expert 8:	I think this depends on how the AI Technology is being integrated into current workflows. If the AI is scanning PAC’s for images to read automatically for VVF, and pushing notifications/reports back to PAC’s or assigned stakeholders (users) in real time, then I would expect the reading and report of this image to be quicker vs the standard care per scan. [redacted] If, however, the image must be manually selected and sent to the AI technology, then this AI repot is generated over a matter of days, the radiologist in the Standard Care Pathway may have already drafted a report and will be waiting for the AI report to see if this aligns before finalizing it and publishing. Alternatively, if the AI turnaround time is too slow they may not a) engage with the AI technology at all, b) write two reports, one with the Radiologist only, and one with the AI interpretation. It could be that the AI report is not integrated into the PACs reporting system therefore the time it takes for the radiologist to log onto the AI platform where this report is uploaded to, to find the image/case number and then read this interpretation and write this over into the PAC system they currently use will add additional time. This time could be lengthened depending on how quick it is to log into this system and find the case they are looking for.
Expert 9:	NR

Expert 10:	According to Guermazi et al (2022) reading and interpreting times for detecting bone fractures in radiographs tended to decrease with the use of AI. However in the early implementation and until all reporters are trained and confident in using AI there will be a slight delay in the reporting turnaround. The additional time is dependent on whether the image is an X-ray vs a CT – An X-ray will take only a few seconds/minutes more when interpreting the AI output. A CT is likely to take longer due to the number of images included in a CT scan. I would suggest that additional time to interpret and report a CT scan is more likely to be more like 10-15 minutes. Guermazi, A. et al. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. Radiology 302, 627–636 (2022). Canoni-Meynet, L. et al. Added value of an artificial intelligence solution for fracture detection in the radiologist's daily trauma emergencies workflow. Diagn. Interv. Imaging https://doi.org/10.1016/j.diii.2022.06.004 (2022). In the study by Jones et al with post implementation of a CXR AI viewer, there was a slight negative effect on reporting times for the majority of radiologists which was expected as looking at the AI output will take longer. In the same study it was noted that even with the negative impact the model had on reporting time, the majority of radiologists (70%) were still satisfied with reporting time while using the CXR viewer, suggesting that the diagnostic improvements offered by the model were enough to offset the additional perceived reporting time. Jones CM, 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:e052902.doi:10.1136/bmjopen-2021-052902
Expert 11:	-
Expert 12:	Initially: AI-assisted reporting may take 5–10 minutes longer per scan, as radiologists review both AI outputs and traditional imaging. With experience: As radiologists become familiar with AI, this time may reduce to 1–3 minutes extra per scan. In some cases: If AI automates prioritization or flagging of fractures, it might even save time by directing attention to critical findings.
Expert 13:	think current and new tools can automate AI reports into regular reports eg similar to lung nodules AI software.
Expert 14:	-
Expert 15:	I would imagine that the reading an AI report on an identified fracture (a positive case) comprises 3 steps: • Reading a report • Checking images • Confirming or refuting an AI result by analysing the images. For the first 2 steps, it would probably take a few seconds (10-15sec) for a radiograph, whereas for CT it would be slightly longer, close to 1 min depending on the presence of sagittal MPRs and the speed of PACS to download images. As for the 3rd step (the most time consuming activity) it could take a few minutes to check whether a fracture identified by AI is a fracture mimic or a true fracture, or if a fracture picked up by AI is a sequel of a previous trauma (not true VFF). The latter would require correlation with previous images.
Meeting on 14/03/2025	Not discussed

Appendix D3: Questions sent to SCMs and experts 25/03/2025

#	Date responses received	Name, Affiliation
1	02/04/2025	[redacted]
2	No response	[redacted]
3	No response	[redacted]
4	26/03/2025	[redacted]
5	25/03/2025	[redacted]
6	25/03/2025	[redacted]
7	25/03/2025	[redacted]
8	26/03/2025	[redacted]
9	27/03/2025	[redacted]
10	31/03/2025	[redacted]
11	No response	[redacted]
12	No response	[redacted]
13	No response	[redacted]
14	No response	[redacted]
15	25/03/2025	[redacted]

Question 1	The EAG has identified a paper which applied an AI technology to Chest X-rays, but this required a frontal (AP or PA) and lateral (LAT) projection. Considering the generalisability of this evidence, in NHS standard care what percentage of all chest X-rays would have both AP+LAT or PA+LAT imaging available?
Expert 1:	From my experience, the majority of chest x-rays performed are AP/PA only without lateral views. I would estimate that a maximum of 20% of all chest x-rays performed would have a lateral view, although in reality this might be as low as 10%.
Expert 2:	-
Expert 3:	-
Expert 4:	No, I do not think that data from CT scan studies can be extrapolated on X-rays, as radiographs (I mean lumbar or thoracic spine radiographs) are specifically used to assess for the presence of vertebral fractures whereas in CT studies vertebral fractures are incidental findings not related to clinical questions raised. Radiologists reporting CT do not routinely report vertebral fractures for various reasons such as they may be not routinely look at sagittal MPRs, not aware of the importance of identifying VFFs, assume them as incidental findings not relevant to a clinical presentation, use different descriptors rather than “a fracture” so on. Therefore, I believe that sensitivity and specificity of spinal radiographs are significantly higher in a real world. If fact, our audit in 2021 identified that all fractures were spotted by reporters, but they used other descriptors.
Expert 5:	Almost none I would imagine, but I don’t have any data to support this
Expert 6:	Lateral radiograph of the chest will only be undertaken when there is concern over the appearances of the anterior projection. Therefore, the percentage of lateral views will be in single figures. Further, when taking images to identify lung or heart disease a fast exposure to high kilovoltage is used. The best images for bones are longer exposures with a lower kilovoltage to allow the lungs to move slightly blurring that imaging making the bones more obvious. Therefore any lateral graphs taken as part of the chest examination will not be as effective as the lateral thoracic spine radiograph.
Expert 7:	Sagittal LAT views are only done when assessing if there is bone involvement (Spine, Vert bodies or Ribs). LATs are not needed to check for dense lung nodules, which I assume are the majority of plain film x-ray cases.
Expert 8:	I would assume that the imaging view would be requested depending on the clinical indication. If the X-ray is to look for thoracic and lung structures then AP/PA imaging with a lateral view would be common for patients who are ambulatory, however there will be a subset of patients who will have x-rays taken on the ward and these are likely to be AP films. My experience from working as a physiotherapist in an acute setting with inpatients is that it would be more common to have an AP image than a lateral imaging. However, in an outpatient setting such as in clinic or outpatient appointments you tend to have the luxury of both.
Expert 9:	The majority of chest X-rays (~60%) are performed with only a PA or AP view, while a smaller proportion (<40%) include both a frontal (AP or PA) and lateral (LAT) projection. LAT views are typically requested only under specific clinical circumstances. This implies that AI models that rely on both frontal and lateral projections may have limited generalisability within NHS standard care, where the LAT projection is not routinely included.

Expert 10:	The lateral chest X-ray is not a standard radiographic projection with, many patients being referred to CT for further evaluation. The main reason a lateral CXR projection is performed is for pacemaker check. I have calculated the number of lateral chest images that we have performed at my trust 01/04/24 – 25/02/25 and the number is 457.The total number of chests examinations we performed over the same period was 94081. Percentage of patients that had a lateral projection at MYTT 0.5% AP and Lateral are not usually performed, as the gold standard is a PA chest projection. The AP projection is only performed on acutely unwell patients and therefore they cannot standard for the standard lateral CXR projection
Expert 11:	-
Expert 12:	-
Expert 13:	-
Expert 14:	-
Expert 15:	In our NHS trust all chest radiographs are performed either in PA or AP(if a patient cannot stand against cassette) position. No lateral chest images are used routinely in CXRs.
Question 2	In the economic model base case, the EAG has assumed for the standard of care arm (No AI-assisted technology) a sensitivity and specificity of opportunistic detection of VFF in CT scans of 25.3% and 89.1%, respectively. These data are taken from the expert elicitation exercise conducted by Dalal et al. (2022). Do these figures seem reasonable?
Expert 1:	Yes, these values seem reasonable given the underreported nature of VFFs
Expert 2:	-
Expert 3:	-
Expert 4:	I could not comment on this based on my experience sorry
Expert 5:	I think this is probably the best evidence available, but are probably ridiculously accurate (to 1 dp) given it was based on the views of 7 people only, and will vary depending on what has been written in the request, to which radiologist is reporting it, to how much osteoarthritis the person has in their spine.

Expert 6:	The work undertaken in the paper you describe was University Manchester where there has for many years been a drive to detect vertebral compression. In my opinion many centres will be lower than these numbers. The following paper from 2023 is written by a large consensus authors but the principal authors are the same. D. C. Howlett et al Radiology reporting of incidental osteoporotic vertebral fragility fractures present on CT studies: results of uk national re-audit. Clinical radiology, 78(12):e1041–e1047, 2023. ISSN 0009-9260. doi: 10.1016/j.crad.2023.09.004. ObjectType-Article-1 You might also look at these papers:- D. L. Md Kendler et al . Vertebral fractures: Clinical importance and management. Am J Med, 129(2):221.e1–221.e10, 2016. ISSN 0002-9343. doi: 10.1016/j.amjmed.2015.09.020. URL https://www.amjmed.com/article/S0002-9343(15)01012-8/pdf. L. Ferrar, G. Jiang, J. Adams, and R. Eastell. Identification of vertebral fractures: An update. Osteoporos Int, 16(7):717–728, 2005. ISSN 0937-941X. doi: 10.1007/s00198-005-1880-x. Leon Lenchik, Lee F. Rogers, Pierre D. Delmas, and Harry K. Genant. Diagnosis of osteoporotic vertebral fractures: Importance of recognition and description by radiologists. AJR Am J Roentgenol, 183(4):949–958, 2004. ISSN 0361-803X. doi:10.2214/ajr.183.4.1830949. Mitchell R, Jewell P, Javaid M, Mckean D, and Ostlere S. Reporting of vertebral fragility fractures: can radiologists help reduce the number of hip fractures? Arch Osteoporos, 12:1–6, 2017. doi: 10.1007/s11657-017-0363-y.
Expert 7:	It is generally believed that the average sensitivity of radiologists lies between 25-40%. While not unreasonable to assume this figure, an audit performed by myself at 4 NHS trusts showed radiologists were 28%, 43%, 47% and 78%. This falls well outside the 25.8% mentioned above. The Royal College of Radiologists usually perform yearly audits on this matter, and their latest figure is that 26.2% of CT reports commented on the severity of fracture. Although fractures can be reported without the need to comment on severity, therefore the percentage of reported VFFs might be higher than the 26.2% in the paper. **10.1007/s00330-020-06845-2**
Expert 8:	Unable to comment
Expert 9:	The sensitivity of 25.3% and specificity of 89.1% for opportunistic detection of vertebral fragility fractures (VFF) in CT scans, as reported in Dalal et al. (2022), reflect the challenges in routine clinical practice where the spine is often not the primary focus of CT scans. Given that these figures are based on expert elicitation, they provide insight into the limitations of traditional methods. However, using the median values from the same study, which were 18.3% for sensitivity and 92.2% for specificity, further emphasizes the difficulty in detecting VFFs without AI assistance. The low sensitivity indicates that a substantial number of fractures are missed, while the higher specificity suggests that when fractures are detected, they are more likely to be correctly identified. These figures highlight the need for AI-driven solutions to improve detection rates, particularly in terms of identifying fractures that might otherwise be overlooked.

Expert 10:	The RCR audit 2019, involved 6,357 patients with an overall radiology departmental response rate of 63%;1,362 (21.4%) of patients had a VFF on auditor review. The sensitivity 25.3% is too low in my expert opinion, considering the type of CT scanning that is performed that demonstrates the spine, CTPA, HRCT, acute abdominal CT scanning and CT scanning for cancer staging, I would expect more opportunistic detection. In the study Dalal et al 2022 – there was considerable uncertainty expressed by the experts that took part in the study, on being the differences in specialism that the experts had and their professional backgrounds. 89.1% - This seems like a reasonable figure in determining the diagnosis of the actual cause of the VFF's. Often other imaging/ investigations are required to provide a definitive diagnosis such as MRI.
Expert 11:	-
Expert 12:	-
Expert 13:	-
Expert 14:	-
Expert 15:	I agree that about only 25-30% CT scans with vertebral fractures performed for other reasons are reported as fractures by radiologists (comparing with our local audit's results). I am not sure about the specificity and how it was defined in the study.
Question 3	Given the above for CT scans, in the base case for X-rays, would it be reasonable to use the same figures as for CT scans in the base case for X-rays? If not, as a percentage, how much higher or lower would it be reasonable to assume?
Expert 1:	I don't think that it would be reasonable to apply the same figures to x-ray, given CT is more sensitive for detecting vertebral fractures compared with x-ray. Therefore, I would assume a reduced sensitivity with x-ray by around 30% (taking the sensitivity to 18%) but would assume a similar specificity.
Expert 2:	-
Expert 3:	-
Expert 4:	As above, I would not have direct experience on this. My suspicion would be that sensitivity is higher with x-ray reporting purely because assessment of bony anatomy is more likely to be routinely reviewed as part of x-ray reporting practice, and less likely to be overlooked than with CT reporting. But I do not report personally myself.
Expert 5:	I think it's reasonable to use the same figures
Expert 6:	What does "base case mean". If you are trying to say that CT has the same sensitivity for compression fractures as radiographs this is not true. CT is a much more accurate way of detecting vertebral compression. The relative strengths are not something that can be guessed. You will need to find some articles that cover this for example: Anderson, Mark W. "Imaging of Thoracic and Lumbar Spine Fractures." Seminars in spine surgery 22.1 (2010): 8–19. Web.
Expert 7:	We cannot assume that reporting standards will be the same for x-rays as CTs. X-rays on the spine are usually performed in the assumption that a VFF is present, whereas CT they are found opportunistically. Therefore, reporting of VFFs would be different between these two modalities, and I would assume x-rays will have a higher reporting percentage than CT for the reasons above.
Expert 8:	Unable to comment

Expert 9:	It would not be reasonable to assume the same figures for X-rays as for CT scans in the base case, due to the differences in radiation exposure and imaging quality. CT scans provide more detailed information and have higher sensitivity for detecting vertebral fragility fractures (VFFs). Without the lateral view, X-ray sensitivity would likely be lower than that of CT scans due to the risk of missed fractures from overlapping structures. Without lateral view (AP or PA only): • Sensitivity: ~15% (lower sensitivity due to missed fractures from overlapping structures). • Specificity: ~86% (slightly lower due to potential misinterpretations). With lateral view (AP+LAT or PA+LAT): • Sensitivity: ~50% (significantly improved sensitivity due to better vertebral visualization). • Specificity: ~91% (comparable or slightly better than CT, with fewer false positives). The addition of the lateral view (AP+LAT or PA+LAT) notably improves the X-ray's sensitivity, bringing it closer to that of CT scans. Without the lateral view, X-ray sensitivity is much lower, but with both frontal and lateral projections, X-ray sensitivity can be considerably enhanced, making it more effective in detecting VFFs.
Expert 10:	In a local audit, approximately 10% of all lateral chest projections had evidence of VFF's on imaging. Using this data I would suggest that the opportunistic detection of VFF's on an X-ray will be much lower than a CT scan. The specificity of VFF's will also be lower on an X-ray as, X-ray alone cannot determine cause of the fracture.
Expert 11:	-
Expert 12:	-
Expert 13:	-
Expert 14:	-
Expert 15:	No, I do not think that data from CT scan studies can be extrapolated on X-rays, as radiographs (I mean lumbar or thoracic spine radiographs) are specifically used to assess for the presence of vertebral fractures whereas in CT studies vertebral fractures are incidental findings not related to clinical questions raised. Radiologists reporting CT do not routinely report vertebral fractures for various reasons such as they may be not routinely look at sagittal MPRs, not aware of the importance of identifying VFFs, assume them as incidental findings not relevant to a clinical presentation, use different descriptors rather than "a fracture" so on. Therefore, I believe that sensitivity and specificity of spinal radiographs are significantly higher in a real world. If fact, our audit in 2021 identified that all fractures were spotted by reporters, but they used other descriptors.

Appendix D4: Additional regulatory guidelines information

Regulation	Training data set requirements	Updating algorithms
UK	Under the Medical Device Regulations 2002 (SI 2002 No 618, as amended), AI systems classified as medical devices must provide detailed documentation about their training data and workflows	MHRA Guidance on Software as a Medical Device addresses iterative updates to algorithms, stating that: • Changes that significantly affect the intended use, performance, or risk profile of the AI system require a new regulatory review.

	as part of the technical file submitted for regulatory approval. MHRA stated that manufacturers must: • Describe the source and quality of training data. • Explain how the data ensures representativeness and avoids bias. • Outline the workflow for data preprocessing, model training, and validation.	• Manufacturers must document the rationale for updates and conduct additional validation to ensure continued compliance. • For deep learning algorithms that use real-world data to improve performance, manufacturers must ensure transparency and maintain clinical safety.
EU	Key requirements include: • A description of the data sets used, including demographic diversity and clinical relevance. • Justification for the sufficiency and quality of the data. • Details on the algorithm's development process, including preprocessing, training, and testing workflows.	• Under the EU MDR 2017/745, significant modifications to AI systems, including updates to algorithms based on real-world data, must be reported to the competent authority. • Annex I and Annex II of the MDR require ongoing clinical evaluation and post-market surveillance to ensure that iterative updates do not compromise safety or performance. • Manufacturers must maintain detailed records of all algorithm changes and their impact on clinical outcomes.

Abbreviations: EU, European Union; MDR, Medical Device Regulations; MHRA, Medicines & Healthcare products Regulatory Agency

NICE National Institute for Health and Care Excellence

Early Value Assessment

Artificial intelligence technologies to aid the opportunistic detection of vertebral fragility fractures

Assessment report overview

This overview summarises key information from the assessment and sets out points for discussion in the committee meeting. It should be read together with the final scope and the external assessment report. A list of abbreviations used in this overview is in Appendix A.

1. The technologies

This assessment is on artificial intelligence (AI) technologies to assist the opportunistic detection of vertebral fragility fractures (VFFs) on X-ray and CT images involving the spine. They are intended to be used as decision aids for healthcare professionals interpreting the X-ray or CT scan. Some companies provide the software directly, whereas others provide it through multivendor platforms. The technologies use X-ray or CT scans in digital imaging and communications in medicine (DICOM) format, which are stored on the hospital's picture archiving and communications system (PACS). Images are then interpreted using proprietary AI-derived algorithms.

Different technologies report and display results in different ways including as annotated images within PACS Viewers, DICOM Secondary Captures or through standalone applications. Some also have notifications or summary reports.

The technologies included in this assessment are shown in table 1.

Table 1 Interventions

Technology (company)	CE mark	Population	Image type	Compatible imaging
Annalise Enterprise CXR and Annalise Container CXR (Annalise.AI)	Class IIb	People ≥16 years of age	X-ray	Chest (AP, PA or LAT orientation)
BoneView (Gleamer)	Class IIa	People >2 years of age	X-ray	Appendicular skeleton, ribs and thoracic-lumbar spine
TechCare Spine (Milvue)	Not yet certified	NR	Lateral X-ray	Thoracic or lumbar spine lateral views
BriefCase-Triage (Aidoc Medical)	Class IIa	People ≥18 years of age	CT	Chest, abdominal
CINA-VCF Quantix (Avicenna.AI)	Not yet certified	People ≥50 years of age	CT	Chest, abdominal
HealthVCF (Nanox AI)	Class IIa	People ≥50 years of age	CT	Chest, abdominal pelvic showing T1-L5
HealthOST (Nanox AI)	Not yet certified	People ≥50 years of age	CT	Chest, abdominal pelvic showing T1-L5
IB Lab FLAMINGO (IB Lab)	Class IIa	People ≥50 years of age	CT	Thoracic or lumbar spine

Abbreviations: AP, Anterior-posterior; LAT, Lateral; NR, Not reported; PA, Posterior-anterior

See section 2 and table 1 in the external assessment report (EAR) for additional details about the included technologies.

2. The condition

A VFF is a break in the spine that occurs when bones are weaker than normal. It is defined as a reduction in vertebral height or vertebral deformity as a result of structural failure (when there is a height reduction or 20% or more, or an endplate deformation) after a fall from standing height or less (low energy trauma). But they can also occur spontaneously from day-to-day activities involving very little trauma or stress. VFFs are the most common type of fragility fracture caused by osteoporosis (a result of bone weakness) which reduces bone density and strength. Osteoporotic VFFs are common in the elderly and particularly in postmenopausal women, but they can also be associated with other conditions or factors, such as chronic or long-term corticosteroid or glucocorticoid usage or malignancy in the vertebrae. Other risk factors include history of falls, family history of hip fracture, low body mass index, smoking, alcohol intake and secondary causes of osteoporosis such as rheumatoid arthritis, inflammatory bowel disease or malabsorption.

There is considerable variation in the terminology used to describe the condition. The external assessment group (EAG) has preferred the term “vertebral fractures”, which can include fractures from moderate and high impact trauma.

3. Current practice

VFFs can be identified when a person presents to a healthcare setting with symptoms suggestive of a VFF or they can be detected incidentally on radiographic images that include the spine (taken for reasons other than a suspected VFF). Yet, up to 70% are missed. A radiographic image is reviewed by a radiologist, reporting radiographer or other trained reporter, usually within 24 hours of the image being taken. Positive findings identified by a radiographer may be reviewed by a radiologist for confirmation.

4. Unmet need

People with a VFF often experience deformity, height loss, immobility and pain, which leads to reduced quality of life. The risk of death is also substantially higher. Furthermore, VFFs are a strong predictor of further osteoporotic fractures, especially of the hip. The economic cost of fractures to the NHS was £5.25 billion in 2017 with a predicted increase to £6.83 billion by 2030. There are effective pharmacological and non-pharmacological treatment options for the management of symptomatic VFFs. Treatment can also reduce the risk of further fractures.

VFFs may be difficult to diagnose clinically. Often, they are detected incidentally on radiographic images, but up to 70% remain undiagnosed. Improving detection and treatment offer a significant opportunity to reduce the burden of VFFs and reduce the risk of further fractures.

5. Innovative aspects

The technologies use AI to detect vertebral fractures. This could improve detection rates for VFF, leading to more people receiving care when necessary. All of the identified technologies have algorithms that are fixed. Four companies (Aidoc Medical, Annalise.AI, Nanox AI and Avicenna.AI) have said that their technologies have settings to control the AI software's sensitivity and specificity, which are configured at set up or during ongoing use. This can help tailor the performance based on the hospital or centre's needs. Some of the technologies include additional features, for example for triage or prioritisation. Further details, including descriptions of the interventions, comparator, care pathway and outcomes, are in the final scope.

6. Diagnostic accuracy and clinical effectiveness

The EAG did a literature search to identify relevant published clinical evidence. The search and selection methods are in section 4 and Appendix A of the EAR.

6.1 Overview of key studies

A total of 22 clinical studies (5 abstracts,1 poster) were selected as key evidence for diagnostic accuracy and clinical effectiveness. The technologies were not equally represented: 5 studies were on Annalise CXR, 1 was on BoneView, 8 were on HealthVCF, 3 were on CINA-VCF, 4 were on IB Lab and 1 was on BriefCase-Triage. There were no studies on TechCare Spine and HealthOST. The technology name and version were poorly reported across the included studies.

Most studies were retrospective diagnostic accuracy studies, with only a few prospective designs reported. However, the EAG judged this to be appropriate for assessing diagnostic accuracy in this case due to the risk of participation bias with a prospective study (involvement in a study aiming to opportunistically detect VFFs may influence reporting of VFFs and thus not reflect current practice). Most studies compared AI output to a reference standard, but the reference standards varied across studies. Twelve studies compared vertebral fracture detection by an AI technology with standard care, which was most commonly the original radiology report. Two academic-in-confidence reports supplied by companies reported [redacted].

The majority of studies were conducted in Australia, Europe and the United States. Only one poster and two abstracts reported on studies conducted in a UK setting. The EAG noted that there was only 1 study that described the use of an AI technology that reflected how it may be used in the NHS (annotations of moderate or severe VFFs applied by the AI before images were reviewed by a radiologist/specialised radiographer) which was conducted in Denmark (Bendtsen and Hitz, 2024). Sample sizes varied widely, from a few hundred scans to over 170,000 in large retrospective datasets. Population characteristics were often poorly reported, with details on age and sex sometimes available but ethnicity frequently missing. Where reported, study populations typically included older adults (mean or median ages generally between 65 to 80 years). In terms of outcomes, diagnostic accuracy was assessed in nearly all studies. Failure rates or inconclusive AI reports were also commonly included outcome measures. A smaller number of studies evaluated impact on clinical management and time to run the software, while healthcare professional acceptability was reported in only one study. None reported on health-related quality of life associated with VFF.

Diagnostic accuracy

Sensitivity, specificity and diagnostic accuracy outcomes for all technologies are summarised in table 2. The EAG highlighted that the sensitivity and specificity (as well as the failure rate and vertebral fracture prevalence) will depend on the eligible population as defined in the technology's instructions for use. Most of the studies reported diagnostic accuracy outcomes per patient, except for one study on Annalise CXR, which reported outcomes [redacted] (Annalise.AI report) and one study on BoneView, which reported outcomes per vertebrae (Oppenheimer et al. 2024). The studies on IB Lab FLAMINGO reported outcomes both per patient and per vertebrae.

Table 2 Summary of diagnostic accuracy outcomes

Study (technology) [location]	Study characteristics	Reference standard/ Comparator	VFF prevalence	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy
Annalise CXR						
Annalise.AI report (Annalise CXR) [redacted]	[redacted]	Reference standard: [redacted] [Comparator: [redacted]]	[redacted]	[redacted]	[redacted]	[redacted]
Ghatak et al. 2024 (Annalise CXR) [US]	Retrospective diagnostic accuracy study (n=596 chest X-ray scans)	Reference standard: Agreement of 2 thoracic radiologists (arbitration of 3rd) [Comparator: Original clinical report]	45.7% (272/595)	0.89 (0.86 to 0.93) - Among all true positive cases (confirmed by reference standard) 36.4% of people had a diagnosis of VCF in the original report	0.89 (0.85 to 0.92)	-
BoneView						
Oppenheimer et al. 2024* (BoneView) [Germany]	Retrospective diagnostic accuracy study (n=512 thoracic and lumbar spine X-ray scans)	Reference standard: Two radiologists (one with MSK experience) [Comparator: NR]	**Lumbar** 12.9% (323/2504 vertebrae) **Thoracic** 10.7% (172/1610 vertebrae) NR per grade	**Lumbar** Lateral 0.63 (±0.05) Lateral and AP 0.72 (±0.04) Grade 1 0.53 **Thoracic** Lateral 0.51 (±0.07) - Lateral and AP 0.61 (±0.07) Grade 1 0.42	**Lumbar** Lateral 0.97 (±0.008) Lateral and AP 0.94 (±0.01) **Thoracic** Lateral 0.98 (±0.007) Lateral and AP 0.94% (±0.01)	-

Study (technology) [location]	Study characteristics	Reference standard/ Comparator	VFF prevalence	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy
				Grade 2 0.72 Grade 3 0.70 Grade 2 0.60 Grade 3 0.60		
BriefCase-Triage						
Wiklund et al. 2024 (BriefCase-Triage) [Sweden]	Retrospective diagnostic accuracy study (n=1,112 abdominal CT scans)	Reference standard: General radiologist with VCF detected also reviewed by MSK radiologist [Comparator: Original radiologist report]	16.8% (187/1112 patients)	Moderate or severe: 0.85 [Original radiologist report: - 0.30 (all grades) - 0.43 (grade 2/3 only)]	Moderate or severe: 0.92	-
CINA-VCF						
Dai et al. 2025 (CINA-VCF) [US and France]	Retrospective diagnostic accuracy study (n=474 chest or abdominal CT scans)	Reference standard: Two radiologists (arbitration by a third) [Comparator: Original radiology report (available for a subset only)]	35.0% (166/474 cases)	0.95 (0.91 to 0.98) [Original radiology report: - 88.2% of fractures missed in the original report were detected by AI]	0.93 (0.89 to 0.97)	0.94 (0.91 to 0.96)
Guenoun et al. 2025 (CINA-VCF) [France]	Retrospective diagnostic accuracy study (n=100 people) with chest, abdominal or pelvis CT scans	Reference standard: Two radiologists [Comparator: None]	52.0% (52/100 patients)	0.92 (0.82 to 0.98)	0.92 (0.80 to 0.98)	0.92 (0.85 to 0.97)

Study (technology) [location]	Study characteristics	Reference standard/ Comparator	VFF prevalence	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy
Avicenna.AI report (CINA-VCF) [[redacted]]	[redacted]	Reference standard: [redacted] [Comparator: [redacted]]	[redacted]	[redacted]	-	-
HealthVCF						
Chappell et al. 2024 (HealthVCF) [UK]	Retrospective diagnostic accuracy study (n=2,000 CT scans involving the spine)	Reference standard: Clinician with local radiologist adjudication [Comparator: Original radiology report]	12.8% (255/2,000)	High specificity configuration: 0.48 Balanced configuration: 0.79 [Original radiologist report: 0.51]	High specificity configuration: 0.99 Balanced configuration: 0.81 [Original radiologist report: 1.00]	-
Kolanu et al. 2020 (Zebra/ HealthVCF) [Australia]	Retrospective diagnostic accuracy study (n=2,357 abdomen or thorax CT scans)	Reference standard: 1 radiologist [Comparator: Original radiology report]	Any VCF: 23.9% (406/1,696) Moderate to severe: 17.8% (280/1,570)	Any VCF: 0.54 (0.50 to 0.59) Moderate to severe: 0.65 (0.60 to 0.71) [Original report: - 72 additional fractures were detected by AI]	Any VCF: 0.92 (0.91 to 0.94) Moderate to severe: 0.92 (0.91 to 0.94)	Any VCF: 0.83 (0.81 to 0.85) Moderate to severe: 0.88 (0.86 to 0.89)

Study (technology) [location]	Study characteristics	Reference standard/ Comparator	VFF prevalence	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy
Page et al. 2023 (Zebra/ HealthVCF) [US]	Retrospective diagnostic accuracy study (n=1,200 chest, abdominal or pelvis CT scans)	Reference standard: 2 neuroradiologists [Comparator: None]	Any: 20.9% (227/1087) Moderate to severe: 12.6% (137/1087)	Any: 0.66 (0.59 to 0.72) Moderate to severe: 0.78 (0.70 to 0.85)	Any: 0.90 (0.88 to 0.92) Moderate to severe: 0.87 (0.85 to 0.89)	-
Bendtsen and Hitz 2024 (HealthVCF) [Denmark]	Prospective diagnostic accuracy study (n=10,012 thorax or abdomen CT scans)	Reference standard: Specialised radiographer/senior radiologist if inconclusive [Comparator: Original radiology report]	9.5% (95/1000 patients)	Moderate to severe: 0.68 (0.58 to 0.78) [Original report: - 9% of VCFs detected by AI were not in the original report]	Moderate to severe: 0.91 (0.89 to 0.93)	0.89
Pereira et al. 2024 (HealthVCF) [NR]	Retrospective diagnostic accuracy study (n=964 chest or abdominal CT scans)	Reference standard: 2 radiologists specialised in musculoskeletal imaging [Comparator: General radiologist (original radiology report)]	16.1% (145/899 scans)	Moderate to severe: 0.74 (0.66 to 0.81) [Original report: - 61.2% of VCFs not in the original report were detected by AI]	Moderate to severe: 0.93 (0.91 to 0.94)	0.90 (0.87 to 0.92)
Roux et al. 2022 (HealthVCF) [France]	Retrospective diagnostic accuracy study (n=173,720 people; n=500 for accuracy) with an abdominal or lumbar CT scan	Reference standard: 2 experts (undefined) [Comparator: None]	50% (250/500)	Moderate to severe: 0.94 (0.89 to 0.98)	Moderate to severe: 0.65 (0.60 to 0.70)	-

Study (technology) [location]	Study characteristics	Reference standard/ Comparator	VFF prevalence	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy
IB Lab FLAMINGO						
Nicolaes et al. 2024 (IB Lab FLAMINGO) [China]	Retrospective diagnostic accuracy study; external validation (n=5,195 abdominal, chest or thoracolumbar spine CT scans)	Reference standard: Sub-specialist radiologist [Comparator: None]	**Per patient:** - Any grade 33.7% (1,622/4,810) - Grade 2/3 only 13.1% (628/4,810) **Per vertebrae:** - Any grade 5.4% (2,623/48,584) - Grade 2/3 only 1.9% (899/48,584)	**Per patient:** - Any grade 0.63 (0.60 to 0.65) - Grade 2/3 only 0.94 (0.92 to 0.96) **Per vertebrae:** - Any grade 0.58 (0.56 to 0.60) - Grade 2/3 only 0.87 (0.85 to 0.90)	**Per patient:** - Any grade 0.94 (0.93 to 0.94) - Grade 2/3 only 0.93 (0.93 to 0.94) **Per vertebrae:** - Any grade 0.99 (0.99 to 0.99) - Grade 2/3 only 0.99 (0.99 to 0.99)	**Per patient:** - Any grade 0.83 (0.82 to 0.84) - Grade 2/3 only 0.93 (0.93 to 0.94) **Per vertebrae:** - Any grade 0.97 (0.97 to 0.97) - Grade 2/3 only 0.99 (0.99 to 0.99)
Nicolaes et al. 2023 (IB Lab FLAMINGO) [Denmark]	Retrospective diagnostic accuracy; summary of development (n=666 abdominal or chest CT scans) and external validation (n=2,000 abdominal or chest CT scans)	Reference standard: Initial triage by medical doctor then reference standard produced by experienced radiologists [Comparator: None]	**Per patient:** - Any grade 20.9% (407/1943) - Grade 2/3 only 15.3% (297/1,943) **Per vertebrae:** - Any grade 2.7% (1,066/24,930) - Grade 2/3 only 4.3% (663/24,930)	**Per patient:** - Any grade 0.76 (0.71 to 0.80) - Grade 2/3 only 0.81 (0.76 to 0.85) **Per vertebrae:** - Any grade 0.47 (0.44 to 0.50) - Grade 2/3 only 0.53 (0.49 to 0.57)	**Per patient:** - Any grade 0.87 (0.86 to 0.89) - Grade 2/3 only 0.95 (0.93 to 0.96) **Per vertebrae:** - Any grade 0.99 (0.99 to 0.99) - Grade 2/3 only 0.99 (0.99 to 0.99)	**Per patient:** - Any grade 0.85 (0.83 to 0.87) - Grade 2/3 only 0.92 (0.91 to 0.93) **Per vertebrae:** - Any grade 0.96 (0.96 to 0.97) - Grade 2/3 only 0.98 (0.98 to 0.98)

Study (technology) [location]	Study characteristics	Reference standard/ Comparator	VFF prevalence	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy
IB Lab Report (IB Lab FLAMINGO) [[redacted]]	[redacted]	Reference standard: [redacted] [Comparator: [redacted]]	**Per patient:** [redacted] **Per vertebrae:** [redacted]	**Per patient:** [redacted] **Per vertebrae:** [redacted]	**Per patient:** [redacted] **Per vertebrae:** [redacted]	[redacted]
Spangeus et al. 2025 (IB Lab FLAMINGO) [NR]	Retrospective diagnostic accuracy study (n=101,246 CT scans for thoracic pathologies, aortic assessment, spinal imaging or abdominal pathologies)	Reference standard: 2 experienced radiologists. [Comparator: None]	45.1% (111/246 scans; NR per patient) NR per vertebrae	**Per patient:** - Grade 2/3 only 0.86 (0.78 to 0.92) **Per vertebrae:** 0.71 (0.64 to 0.78)	**Per patient:** - Grade 2/3 only 0.99 (0.96 to 1.00) **Per vertebrae:** 0.99 (0.98 to 0.99)	**Per patient:** - Grade 2/3 only 0.93 (0.89 to 0.96) **Per vertebrae:** 0.96 (0.95 to 0.97)

Abbreviations: MSK, Musculoskeletal; NR, Not reported; VCF, Vertebral compression fracture; VFF, Vertebral fragility fracture

*See table 5 in the EAR for lumbar and thoracic sensitivity values for BoneView for anterior (wedge) and middle (crush) fractures specifically.

Failure rate

The failure rate for Annalise Enterprise CXR ranged from 0% to 0.2%. One study reported that 1.4% of lumbar spine and 5.1% of thoracic spine images were unable to be processed by BoneView due to incorrectly classifying the image. One study reported a failure rate of 0.6% for BriefCase-Triage. The failure rate for HealthVCF was between 5.6% and 9.4%. For IB Lab FLAMINGO it ranged from 1.0% to [redacted]%.

Changes to clinical management

One diagnostic accuracy study on BriefCase-Triage reported that only 10% (3/30) of people with a reported finding had started pharmacologic osteoporosis treatment within 1 year follow-up (Wiklund et al. 2024). The study by Bendtsen and Hitz (2024) evaluating HealthVCF found that among 223 people with VFFs but no prior osteoporosis diagnosis, 16.6% were referred for a DEXA scan, 11.2% received a new osteoporosis diagnosis and 10.3% began treatment. Among 218 people with pre-existing osteoporosis, 16 (7.3%) had medication changes, and of the 64 not on treatment at baseline, 41 were referred for scans and 11 started treatment, though overlaps were not specified. The EAG noted that referral and treatment rates were similar between people with and without prior osteoporosis diagnosis, raising uncertainty about the added value of early detection. Connacher et al. (2019) reported that among a subset of 50 people in the UK with a VFF identified by HealthVCF and who had completed the FLS assessment, 57% had no further assessment after this, 25% had a GP referral, 6% bone clinic referral and 12% had died.

User acceptability

One study reported on staff perception following implementation of Annalise CXR across a radiology network in Australia. During the 6-week pilot phase 90% of users reported subjective improved accuracy, with 90% also reporting satisfaction with the accuracy of the AI model findings (n=10 users). At the end of the pilot phase, 90% reported a more positive attitude towards Annalise CXR and 90% reported a more positive attitude towards AI in general.

See tables 3 to 13 in the EAR for details about the available outcomes.

Data on subgroups

Thirteen studies were conducted in a population aged 50 years or older, which is a relevant subgroup. There was limited evidence about the clinical effectiveness of AI technologies in other subgroups. Several studies reported subgroup analysis (for example by age or sex), but none of them were statistically powered to determine differences in diagnostic performance by subgroup. However, two studies on IB Lab FLAMINGO found higher sensitivity in females: Nicolaes and colleagues (2023) reported a per-patient sensitivity of the technology of 0.87 (0.81 to 0.91) for females and 0.72 (0.64 to 0.80) for males. Spangeus and colleagues (2025) reported a per-patient sensitivity of 0.95 (0.89 to 1.00) for females and 0.75 (0.63 to 0.87) for males. A company report by IB Lab noted that the software [redacted]. The software version used was [redacted]. One study on CINA-VCF in abstract form included only people [redacted] (Avicenna AI report). The positive predictive value of the technology in this study was [redacted].

6.2 Ongoing studies

The EAG searched ClinicalTrials.gov on 21 March 2025, but found no ongoing studies relevant to the decision problem. Companies provided the following information about ongoing studies for each technology:

- BriefCase Triage (Aidoc Medical): 2 retrospective single-centre studies (locations and sample sizes not reported).
- Annalise CXR (Annalise.AI): [redacted].
- CINA-VCF Quantix (Avicenna.AI): 1 single-centre RCT (France), 1 multicentre RCT (4 hospitals in France), and 1 RCT in Austria for people aged 50 years or older.
- BoneView (Gleamer): No direct information provided; the EAG identified 5 relevant studies among 30 ongoing studies.
- IB Lab FLAMINGO (IB Lab): [redacted].

- Health VCF (Nanox AI): 1 UK multicentre study (ADOPT) comparing outcomes from scans taken in 2017 with those taken in 2022 across a range of metrics; one poster from ADOPT was included in the evidence by the EAG (Chappell et al. 2024).

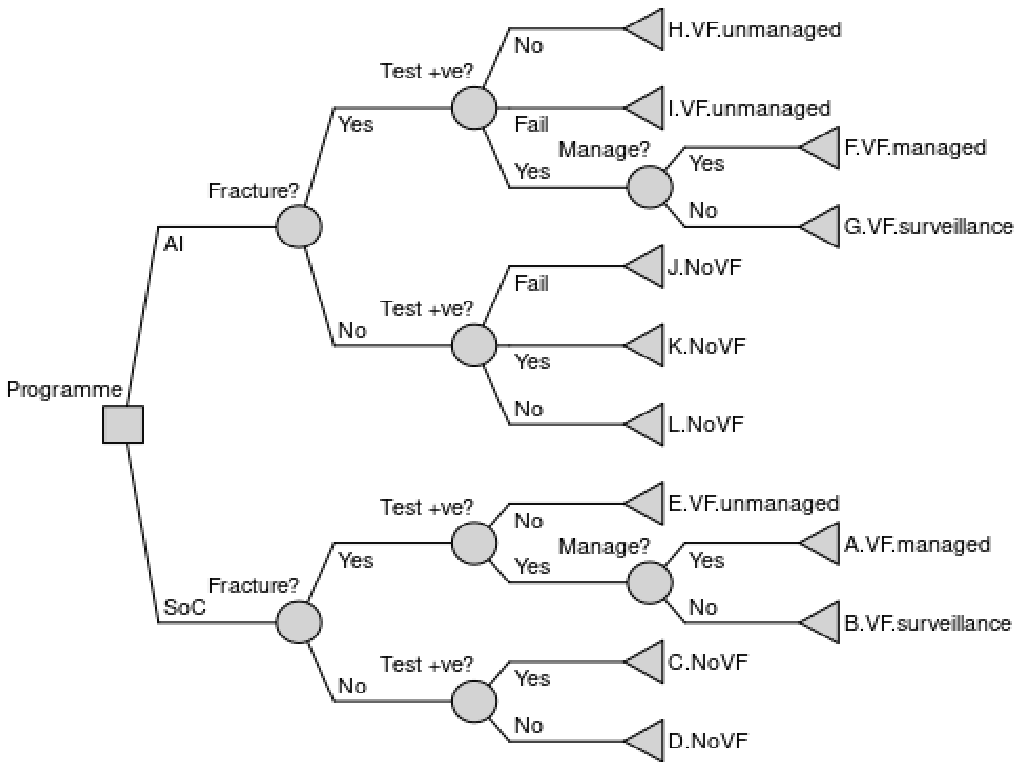
7. Health economic evidence

The EAG did a review to identify existing health economic evidence, including relevant economic models. It found no economic evaluations relevant to the decision problem, but it identified 5 publications and 2 NICE technology appraisals which were used to develop the EAG’s economic model or to inform parameters within it. An overview of these is in section 6.1 of the EAR.

7.1 Health economic model

The EAG developed an early economic model that investigated the potential cost-effectiveness of opportunistic detection of VFFs with assistance from AI technologies compared with current standard of care (reporting radiographer without AI assistance) in a hospital setting. The model was a decision tree with a 1-year time horizon. The model structure is depicted in figure 1.

Figure 1 Model structure



The EAG modelled each technology separately, as differing eligibility criteria meant that starting population characteristics, including vertebral fracture prevalence, could vary. To ensure transparency, the EAG also modelled a generic AI technology scenario.

Key assumptions by the EAG included:

- If a vertebral fracture is opportunistically detected (test positive), then a radiologist specialising in musculoskeletal imaging will *always* review the image to confirm its presence.
- When a vertebral fracture is opportunistically detected (test positive), only 14.6% of cases will receive treatment (see section on Costs). Those people will experience a quality-of-life gain over the one-year time horizon. No longer-term benefits (such as reduced risk of future fractures) were modelled.
- If a vertebral fracture is present but not detected (false negative, interpreted by the EAG as a missed opportunity of detecting vertebral fracture) no additional costs are incurred and there is no impact on outcomes.
- In the standard of care arm, the sensitivity and specificity were the same regardless of the imaging modality (CT or X-ray).
- VFF detection occurs once per-patient within the time horizon. The EAG noted that in practice a person can receive more than one scan or episode of imaging and each may be used for opportunistic detection, which could reduce the missed opportunities, but will increase the false positive rate and would incur the cost of the technology.
- The underlying prevalence of VFFs is the same in the modelling of all technologies; the prevalence differed significantly in the clinical evidence (see table 2) and may vary by technology due to the different indications and contraindications.

The impact of some of the assumptions was tested in one-way sensitivity analyses (see section on Sensitivity analyses). Further details of the economic modelling are in section 6.2 of the EAR.

7.2 Model inputs

Prevalence

The EAG used a prevalence of 29.7% in the base case. This was the mid-point of the range of values identified in the clinical evidence. The EAG noted that there was no consensus on the definition of prevalence used in the literature, and that the wide range led to considerable uncertainty. A range of values was explored in sensitivity analyses.

Sensitivity and specificity

Standard of care arm

The sensitivity and specificity values used in the standard of care arm were 0.25 and 0.89, respectively. The values were sourced from a UK based expert elicitation study (Dalal et al. 2022) and were for interpretation of CT scans. The EAG assumed that the sensitivity and specificity of interpreting an X-ray scan would be the same due to a lack of data.

Intervention arm

The sensitivity and specificity values used in the intervention arm are presented in table 3. Where more than one value was available from the clinical evidence review the EAG prioritised studies reporting accuracy per patient (rather than per vertebrae), those defining a VFF as grade 2 or 3 and those comparing with radiologist specialising in musculoskeletal imaging as the reference standard.

Table 3 Sensitivity and specificity values in the intervention arm

Technology	Sensitivity	Specificity	Source/EAG comment
Non-confidential base case	0.598	0.999	Curl et al. 2024
Annalise CXR	0.89	0.89	Ghatak et al. 2024
BoneView	0.72	0.94	Oppenheimer et al. 2024
TechCare Spine	Not reported	Not reported	No evidence identified, not included in economic model.
BriefCase-Triage	0.85	0.92	Wiklund et al. 2024
CINA-VCF Quantix	0.95	0.93	Dai et al. 2025
HealthVCF /HealthOST	0.74	0.93	Pereira et al. 2024
IB Lab FLAMINGO	0.94	0.93	Nicolaes et al. 2024

A range of sensitivity and specificity values were explored in the sensitivity analyses.

Failure rate

Where possible, the EAG used failure rate values from the same study as the sensitivity and specificity values, or, if unavailable, used data from the largest relevant study. Failure rates ranged between 0.2% and 5.6% (see table 16 in the EAR). Alternative failure rates were explored in the sensitivity analyses.

Costs

Technology costs

The EAG calculated the cost per scan for each technology which included product subscription, implementation, integration, training and maintenance costs (see section 6.2.4.2 of the EAR for specific methods and assumptions made by the EAG). The EAG assumed 65,000 scans per site in the base case for technologies which process X-ray images and 6,500 scans per site for those that process CT scans based on Diagnostic Imaging Dataset statistics (2023 to 2024). Alternative cost per scan values based on alternative scan volumes were explored in the sensitivity analyses. The EAG also assumed that the cost of AI is applied to all diagnostic images (that is, that the hospital cannot specify which images the AI is applied to). Total costs per technology per scan are presented in table 4.

Table 4 Technology costs used in the economic model

Technology	**Generic AI**	**Annalise (X-ray)**	**Bone View (X-ray)**	**Tech Care Spine (X-ray)**	**Brief Case-Triage (CT)**	**CINA-VCF Quantix (CT)**	**HealthVCF/HealthOST (CT)**	**IB Lab FLAMINGO (CT)**
Total cost per scan in base case (no. of scans) **Range used in sensitivity analyses**	£7.36 (Curl et al. 2024)	[redacted] [redacted]	£1.00 (Notional cost)	[redacted] [redacted]	[redacted] [redacted]	[redacted] [redacted]	[redacted] [redacted]	[redacted] [redacted]

Other treatment and follow-up costs

Other costs applied in the model are shown in table 5. See section 6.2.4 in the EAR for full details of costs included in the model.

A cost was only applied in the model in the case of a positive result to reflect an additional review by a radiologist to confirm the result, and an additional spine X-ray for 10% of positive cases. A cost was also applied to the 14.6% of true positive cases that were assumed to have their VFF managed based on Dalal et al. 2022. This cost was based on a NICE technology appraisal for the treatment for osteoporosis (TA464) and comprised hospitalisation, A&E, GP, referral and prescribing.

Table 5 Other costs used in the economic model

Parameter	**Cost**	**Source**
Cost applied to a positive result	£29.55	Additional review by radiologist plus 10% assumed to have additional spine X-ray
Cost applied to a negative result	£0	Assumption
Cost applied to a failed result	£0	Assumption
Cost of treatment for VFF	£5,302	Based on TA464.

Health-related quality of life

Only people with a true positive test result who had their VFF managed (14.6% of true positive cases) had a utility gain applied in the model. The utility value applied was 0.27 based on Svedbom et al. 2018. The study reported a value of 0.53, but the EAG judged this to be an overestimate because people who have had their VFF opportunistically detected may not be experiencing symptoms of VFF. Further reductions in the utility gain were explored in sensitivity analyses.

7.3 Model results

Base case (generic AI)

The generic AI technology cost £86 more per person (£149 in the intervention arm, compared to £63 for standard of care). But, there was a QALY gain of 0.004 (AI arm: 0.007, SoC: 0.003). This resulted in an incremental cost-effectiveness ratio (ICER) of £22,085 per QALY. For each 1000 people scanned, AI detected 101 additional vertebral fractures, 15 of which would be managed, with fewer missed opportunities (AI arm: 118, SoC: 222). Up to 25 additional scans would require radiologist review.

Base case (named technologies)

The incremental cost and QALYs, ICERs and additional VFFs identified per 1000 scans and additional reviews required per 1000 scans for all technologies are presented in table 6. For all technologies, the ICERs were lower than the generic AI case, but slightly higher than the £20,000/QALY threshold.

Table 6 Base case results (ICERs) for all technologies

Technology	Incremental cost per patient (£)	Incremental QALYs per patient	ICER (£/QALY)	Additional VFFs identified with AI per 1000 scans	Additional reviews required per 1000 scans
Generic AI	86	0.0039	22,085	101	25
Annalise CXR	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

BoneView	108	0.0052	20,755	135	98
BriefCase-Triage	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
CINA-VCF Quantix	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
HealthVCF	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
IB Lab FLAMINGO	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Abbreviations: AI, Artificial intelligence; ICER, Incremental cost-effectiveness ratio; QALY, Quality-adjusted life year; VFF, Vertebral fragility fracture

All technologies were more expensive than standard of care and all resulted in QALY gains and in additional VFFs detected (and so, additional scans would need to be reviewed by a radiologist).

The EAG noted that the results for the technologies which interpret X-ray scans may be an underestimate. This is because it used the same standard of care sensitivity and specificity values as those for interpreting CT images (due to a lack of data). However, the sensitivity of single radiograph is likely lower than that of multi-slice CT. So, the economic model may have overestimated the number of VFFs detected in the standard of care X-ray arm.

Of note, these results depict the cost-effectiveness within a 1-year time horizon. The EAG noted that in the long-term the technologies may be more cost-effective. This is because important long-term effects, such as reduced future fracture risk for those that have their VFF managed, or reduced quality of life and additional healthcare costs for people who have their VFF missed, are not captured by the model.

See section 6.3 of the EAR for additional detail on the results of the economic modelling.

Sensitivity analyses

Univariate changes in the utility gain associated with vertebral fracture management had the largest impact on the cost-effectiveness results for all technologies. In the generic AI case reducing the EAG's base case utility gain by 50% (from 0.27 to 0.13) increased the ICER above £40,000.

The cost-effectiveness results were not sensitive to changes in VFF prevalence, changes in the intervention failure rate or changes affecting the costs of the technologies such as volume of scans or management costs. The economic model was also not sensitive to changes in the sensitivity and specificity (both intervention and comparator arms), except in the generic base case.

Sensitivity analysis results are in tables 22 to 28 of the EAR.

8. Evidence gaps

The EAG's evidence gap analysis for selected outcomes for which some evidence was identified is presented in table 7. No evidence was available for accuracy by healthcare professional, staff costs, and downstream consequences of detection or missing fractures (such as further imaging requirements, fracture related injury, hospitalisations, referral or treatment). No evidence on the impact of improved opportunistic VFF detection on a person's health-related quality of life was available. See table 29 in the EAR for the EAG's evidence gap analysis for all outcomes.

Table 7 Evidence gaps for selected outcomes

Outcome	Annalise Enterprise CXR*	Annalise Container CXR*	Bone View	Brief Case-Triage	CINA-VCF Quantix	Health VCF	Health OST	IB Lab FLAMINGO	Tech Care Spine
Diagnostic accuracy	GREEN	RED	AMBER	GREEN	GREEN	GREEN	RED	GREEN	RED
Failure/inconclusive report rate	GREEN	RED	AMBER	GREEN	RED	GREEN	RED	GREEN	RED
Number of missed fractures	GREEN	RED	RED	RED	GREEN	GREEN	RED	AMBER	RED
Accepta-bility	GREEN	RED	RED	RED	RED	RED	RED	RED	RED
Changes to clinical practice	RED	RED	RED	RED	RED	GREEN	RED	RED	RED
Time to produce a report	RED	RED	RED	RED	RED	AMBER	RED	AMBER	RED

*The EAG noted that it did not find any published evidence directly related to Annalise Container CXR. But, the company said that Annalise Container CXR uses the same AI model as Annalise Enterprise CXR, so the evidence is generalisable between the two technologies.

The EAG highlighted that the limited information about ongoing studies means that it is unclear whether any of the evidence gaps will be addressed by any of the ongoing studies.

The EAG detailed some evidence generation recommendations in table 30 of the EAR and highlighted the following key evidence gaps:

- Poor reporting of patient demographics, indication for diagnostic imaging, software name, version number and configuration settings.
- Lack of evidence on subgroups, some technologies, diagnostic accuracy of standard care, longer term impact or changes to clinical management as a result of opportunistic detection.
- Lack of evidence in a UK setting.
- Variations in the reference standards across studies and the reference standards not always reflecting the standard of care in the NHS.

9. Equality considerations

The final scope and the scoping equality impact assessment describe equality considerations for this assessment:

- Radiology access is geographically uneven; AI for VFF detection may improve radiology service in areas lacking specialised radiologists.
- AI performance varies with bone disorders, age, and ethnic representation in training data.
- VFF risk rises with age, female sex, lower socioeconomic status, certain conditions, and medications.
- Bone density differences and lack of race-specific standards may affect diagnostic accuracy.

The EAG identified 3 additional equality considerations related to contraindications for use of some of the technologies:

- The patient age as listed in the instructions for use varies across technologies. Five of the technologies are only indicated for people over 50 years of age.
- Some of the technologies advise that image sites must not contain cement, surgical hardware or spinal metalwork.
- AI technology may misidentify scans when large fields of view are needed, this could be more prevalent in people with obesity.

10. Key points, limitations and considerations

10.1 Diagnostic accuracy and clinical effectiveness

Key points

- The evidence demonstrated that the AI technologies were able to detect additional moderate to severe vertebral fractures (as confirmed by a reference standard) which were not reported in the original radiology report, with generally high specificity across all technologies (where evidence was available).
- Evidence across technologies is varied and some technologies had little (BoneView, BriefCase-Triage) or no (TechCare Spine) published evidence identified.

Limitations

- There was a general lack of evidence from a UK setting.
- Reference standards varied across studies and did not always reflect the UK reference standard.
- There was a lack of prospective evidence and so a lack of evidence on downstream consequences of increased detection of VFF and the impact of false positives on workflow and need for further imaging.
- The configuration details and version of software used in the studies were generally poorly reported and these could impact on diagnostic accuracy estimates.
- Evidence for outcomes in subgroups was limited.

Considerations for committee

- Is evidence from other settings generalisable to the UK NHS?
- What can the studies tell us about the diagnostic accuracy of the technologies?
- What can the studies tell us about the likely impact of the technologies on downstream consequences such as patient outcomes or system impact?
- Are there specific patient populations (e.g., older adults, known osteoporosis risk) where AI detection would be especially valuable or problematic?

10.2 Health economic evidence

Key points

- Economic modelling suggests the AI technologies could be plausibly cost-effective.
- The cost-effectiveness results are subject to uncertainty.

Limitations:

- The cost-effectiveness model only estimated the short term (1-year) costs and consequences of implementing the technologies. Including longer-term benefits and costs would likely improve the cost-effectiveness of the technologies.
- The model was very sensitive to the utility gain associated with finding and treating a VFF and this parameter was uncertain, as the data available did not directly relate to the decision problem (utility gain was for symptomatic fractures).

Considerations for committee:

- Are the economic model structure and assumptions appropriate to assess the potential cost-effectiveness of the technologies?
- Are the clinical and cost parameters appropriate to assess the potential cost-effectiveness of the technologies?
    - Is the EAG's base case utility gain for detection and treatment of VFF plausible and appropriate?

- What can the economic evidence tell us about the potential for the technologies to be cost-effective?

10.3 Technical considerations

Key points:

- There is variation between technologies in the:
  - Type of image processed (CT or X-ray) and image requirements (view and anatomical area)
  - Eligibility criteria (e.g. patient age)
  - How findings are reported and displayed (see section 2 in the EAR).
- Information on training and validation of the algorithm was only available for some technologies.
- The impact of configuring the sensitivity and specificity of the technologies is uncertain.

Considerations for committee:

- Are all technologies applicable to the NHS? Are any more appropriate than others?
- Should specificity be prioritised where technologies allow for configuration (to minimise false positives)?

Appendix A. Abbreviations

A&E	Accidents and emergency
AI	Artificial intelligence
DICOM	Digital imaging and communications in medicine
EAG	External assessment group
EAR	External assessment report
EVA	Early value assessment
ICER	Incremental cost-effectiveness ratio
PACS	Picture archiving and communications system
TA	Technology appraisal
VFF	Vertebral fragility fracture
QALY	Quality-adjusted life year

NICE National Institute for Health and Care Excellence

NICE Health Tech Programme

GID-HTE10059 Artificial intelligence technologies to aid the opportunistic detection of vertebral fragility fractures: Early Value Assessment

External Assessment Report (EAR)

Collated comments table

Any confidential sections of the 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

Redacted External Assessment Report – Collated comments table:

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
1	Chuck Lam, University of Birmingham (Expert)	3	Acknowledgements	Please change name to Chuck Lam, and affiliation to “University of Birmingham”	Thank you for your comment. This change has been made to the report.
2	Chuck Lam, University of Birmingham (Expert)	21	3. Clinical Context	It may be inaccurate to state that “The prevalence of vertebral fractures is currently unknown”, instead prevalence is uncertain, as previous guidelines have stated that “**12 % of women aged 50–79** have vertebral fractures (≥ 20 % by 80 y).” (ROS “Clinical Guidance for the Effective Identification of Vertebral Fractures”, 2017.)	Thank you for your comments. We have changed “unknown” to “uncertain”.
3	Annalisa Occhipinti, Teesside University (Expert)	72	5.3	When reporting the comment of the Clinical Expert mentioning that “a better description of what the AI is doing is required”, a short sentence on explainability should be integrated. Does any of the analysed AI methods embed any explainable approach?	Thank you for your comment. We have replaced “better” with “transparent” as we have not identified any explainable approach described for any of the AI technologies.
4	Annalisa Occhipinti,	80	6.2	Typo “management 9noting”	Thank you for your comment. We have edited this typo.

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
	Teesside University (Expert)				
5	Annalisa Occhipinti, Teesside University (Expert)	99	6.3	When presenting the generic AI model (base case), the current report does not specify which parameters were used for the model or if any hyperparameter tuning was performed to optimise the model.	Thank you for your comment. Values for the base case are present in Table 16. To clarify we have tried to use consistent terminology and used "generic AI" to make this clearer (was previous described as "non-confidential" inconsistently in some sections). We assume the term hyperparameter tuning is used in the context of the economic evaluation model. If so, we can confirm that this was not performed as these are typically used when developing a machine learning model rather than evaluating them. Hyperparameters can also refer to the parameters used to define distributions of the main parameters for PSA; however this is not applicable as the EAG did not conduct PSA.
6	Tracy O'Regan, The Society and College of Radiographers	10	Background	'or associated with malignancy' – also primary tumour?	Thank you for your comment. We have edited the sentence to read: "They may be caused by osteoporosis (due to bone weakness), primary tumour or associated with malignancy."
7	Tracy O'Regan, The Society and College of Radiographers	11		'The EAG considered the waiting lists for dual energy X-ray absorptiometry (DEXA) (bone density) scans and general radiologist staff shortages, and therefore the impact of false positives on healthcare resources, workforce and waiting times needs careful consideration.' It is not clear why the impact of lack of equipment and staffing equates to 'therefore the impact of false positives on healthcare resources'. What are the current figures related to false positives, is this information available?	Thank you for your comment. We have edited the sentence to clarify that we mean false positives from AI: "The EAG considered the waiting lists for dual energy X-ray absorptiometry (DEXA) (bone density) scans and general radiologist staff shortages, and therefore the impact of false positives when implementing AI technologies on healthcare resources, workforce and waiting times needs careful consideration."

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
					False positives are described in the results section when reported in the published studies (see section 5.1.1 to 5.1.6).
8	Tracy O'Regan, The Society and College of Radiographers	12		'The Clinical Experts have advised that diagnosis of vertebral fracture would be conducted by a radiologist with musculoskeletal special interest as standard in the NHS'. That is likely the case for opportunistic detection using CT Scan. However, the assessment also refers to 'range of modalities' in which case the reporting of X-rays (also referred to as plain film) at some organisations in the UK is predominantly undertaken by reporting radiographers.	Thank you for your comment. We have added a sentence to the Executive Summary (and section 3.1) to acknowledge the role of reporting radiographers in diagnosis of vertebral fractures, referencing the RCR guidance on recognition of VFFs.
9	Tracy O'Regan, The Society and College of Radiographers	13		'The consent process for patients' diagnostic images to be processed (by an AI technology) for reasons other than their direct care requires consideration.' – this is regulated via Information Commissioners Office which the EAG may wish to reference.	Thank you for this comment. We have added the text in red, "The consent process for patients' diagnostic images to be processed (by an AI technology) for reasons other than their direct care, including transparently reporting how their data will be used and by who, is regulated by the Information Commissioners Office." Similar changes have been made in section 8.2 for consistency.
10	Tracy O'Regan, The Society and College of Radiographers	13		'The EAG notes that there is evidence from other fields' – what other evidence?	Thank you for this comment. This was related to cancer detection (Davenport et al. 2023). The EAG has updated the executive summary and section 8.2 of the EAG report to make it clear that this evidence was not specific to vertebral fracture detection.
11	Tracy O'Regan, The Society and College of Radiographers	13		'The sustainability of AI requires consideration where new versions require training and validation' it's not immediately clear what the definition of sustainability is in this context – could relate to green/sustainable AI for example.	Thank you for this comment. We have replaced the word "sustainability" with "implementation" to clarify meaning.
12	Tracy O'Regan, The Society and	21		'vertebral fractures identified by AI will need additional radiologist review,' or reporting radiographer, for example at	Thank you for this comment. We have added the additional words in red to this sentence:

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
	College of Radiographers			organisations where reporting radiographers routinely provide reports, most commonly for X-ray.	"vertebral fractures identified by AI will need additional radiologist or reporting radiographer review
13	Tracy O'Regan, The Society and College of Radiographers	21		'imaging (and therefore increased radiation dose)' – not necessarily if the reviewer does not require additional imaging.	Thank you for this comment. We have edited this sentence to highlight that increased radiation dose may only occur if additional imaging is required.
14	Tracy O'Regan, The Society and College of Radiographers	21		'AI systems are currently limited in differentiating vertebral fractures from non-fracture deformities (which may include cupid's bow or limbus vertebra which are developmental variants, Schmorl's nodes or Scheuermann's disease, H-shaped vertebrae associated with sickle cell disease of Gaucher's disease (Lenchik et al., 2004)' rather the issue is that there is no current evidence of AI systems' abilities to identify those non-fracture deformities.	Thank you for your comment. The text has been amended to clarify the point from Lenchik et al. 2024 (that spinal deformities can mimic vertebral fractures) and the views of several experts (that there is uncertainty whether the AI can differentiate deformities from fractures, and this could potentially increase false positives). Only a handful of studies explored reasons for false positives (data was extracted where available in sections 5.1.1-5.1.6).
15	Tracy O'Regan, The Society and College of Radiographers	22		'and the referrer must ensure arrangement of appropriate assessment for osteoporosis and fragility fracture and subsequent treatment'. The use of fracture liaison services and availability of those services across the UK are also an issue that may be relevant at this point.	Thank you for this comment. The EAG have added the additional text in red to clarify this point: "and the referrer must ensure arrangement of appropriate assessment for osteoporosis and fragility fracture and subsequent treatment which has consequences for downstream services."
16	Tracy O'Regan, The Society and College of Radiographers	25	3.1	'The proportion of CTs' more specifically, the proportion of CT reports (not CT images).	Thank you for your comment. We have amended the text: "The proportion of CT reports which included the terminology of "vertebral fracture"…"
17	Tracy O'Regan, The Society and	71		'The EAG notes that the reference standard described in the included evidence did not always reflect the recognised reference as used in the NHS (that is, review by a radiologist	Thank you for your comment. We have altered this sentence to focus on the main point which is that the reference standard

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
	College of Radiographers			specialised in musculoskeletal imaging)' or reporting radiographer in the case of organisations where reporting radiographers routinely report the majority of cases (X-ray).	varied across the evidence. For acknowledgement of reporting radiographers – see response to comment #8 and #12.
18	Tracy O'Regan, The Society and College of Radiographers	72		'The Fracture Liaison Service database may be well placed to expand its data collection to better quantify the use of imaging and AI, and routes to diagnosis of vertebral fracture in the future.' Excellent point.	Thank you for this comment. No change to EAG report required.
19	Tracy O'Regan, The Society and College of Radiographers	123	Table 30	'What staff roles/bands were used in the initial reporting of eligible scans (and opportunistic detection of vertebral fracture)' presumably that is to determine cost. It may be difficult to determine given the large range of role descriptions used to describe reporting radiographers across healthcare organisations. The reference to 'band' presumably references agenda for change, which uses the term band, but not applicable to radiologists. Depending on the data that is required, a more specific requirement will be useful.	Thank you for this comment. We have edited the research question to the following: "What staff roles and levels of experience/qualification were used in the initial reporting of eligible scans (and opportunistic detection of vertebral fracture)"
20	Tracy O'Regan, The Society and College of Radiographers		General	To note, the term radiographer is used in the assessment. The titles diagnostic radiographer and therapeutic radiographer are each regulated in the UK. The reader is left to assume that the document uses the term radiographer to refer to diagnostic radiographer. It may be useful to note and use the term diagnostic radiographer in the guideline document that results thanks. This was an clear, accurate and logical overview that informs in a structured manner. Thankyou for the opportunity to comment and congratulations to the EAG on a timely assessment.	Thank you for your comment regarding specifying the appropriate title. We have added a sentence to the definitions section early in the report to set context: "Mention of radiographer throughout this report should be interpreted as diagnostic radiographer". Thank you for highlighting this. Thank you for your support and time in responding to the work.
21	Alicja Raginis-Zborowska, Annalise-AI	17	2	We refer to the statement below: "Two local AI experts advised the EAG of the following in relation to regulations of AI products for medical or clinical use" The second bullet underneath this header refers to DTAC – which includes regulatory aspects for digital health technologies that are medical devices - but it is not correct to call DTAC a regulation.	Thank you for your comment. This change is reflected in the report

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
				The original DTAC pages from NHSX (now part of the NHSE Transformation Directorate) states *DTAC does not seek to introduce new requirements for the procurement of technology. It brings together legislation and recognised good practice into one place covering clinical safety, data protection, technical assurance, interoperability and usability and accessibility. [source https://transform.england.nhs.uk/key-tools-and-info/digital-technologyassessment-criteria-dtac/how-to-use-the-dtac/]* Suggested change: *"Two local AI experts advised the EAG of the following in relation to regulations and best practice concerning AI products for medical or clinical use:"*	
22	Alicja Raginis-Zborowska, Annalise-AI	17	2	We refer to the statement below: "AI systems that require internet access and that are integrated into hospital systems are required to meet DTAC." We do not believe this statement is factually correct. A digital health technology should have a completed DTAC assessment even if it is only deployed with a single organisation and has no communication beyond the organisation's network. This could be confusing DTAC with a Data Protection Impact Assessment (DPIA), which is one of the five dimensions of DTAC, and which would be required if data left an NHS organisation's network. We suggest changing this bullet to: *"AI systems will be required to demonstrate they meet the DTAC, and the supplier should expect to provide a completed DTAC assessment"*	Thank you for your comment. This has been updated in the report

<table>
<tr><th>Comment no.</th><th>Stakeholder</th><th>Page no.</th><th>Section no.</th><th>Comment</th><th>EAG Response</th></tr>
<tr><td></td><td></td><td></td><td></td><td>For an independent view, see above comment for link to NICE-hosted DTAC pages and also https://transform.england.nhs.uk/key-tools-and-info/digital-technology-assessment-criteria-dtac/</td><td></td></tr>
<tr><td>23</td><td>Alicja Raginis-Zborowska, Annalise-AI</td><td>18</td><td>2</td><td>We refer to the statement below:
“ Each technology reports findings and images in a different manner as summarised below:
• Annalise Enterprise CXR and Annalise Container CXR (cloud-based version of Enterprise) report results in a desktop application, which synchronises with PACS, that displays findings and localisation of the fracture."

We are seeking corrections to the product description and method in which AI findings are presented:
• Annalise Enterprise CXR reports results in a desktop viewer application which synchronises with PACS, and displays findings and localisation of the fracture. AI outputs from Annalise Enterprise CXR are also available as triage notifications, or DICOM secondary capture images inserted to the site PACS, which displays findings and localisation. Annalise Container CXR (a version of Enterprise which is packaged to optimise it for running on selected third-party marketplaces / platforms) output</td><td>Thank you for this comment. The EAG has applied these changes.

[The EAG note there appeared to be missing parts of the response, queried with NICE 06/05/2025. Clarification response received 09/05/2025 “Annalise Enterprise CXR reports results in a desktop viewer application which synchronises with PACS, and displays findings and localisation of the fracture. AI outputs from Annalise Enterprise CXR are also available as triage notifications, or DICOM secondary capture images inserted to the site PACS, which displays findings and localisation. Annalise Container CXR (a version of Enterprise which is packaged to optimise it for running on selected third-party marketplaces / platforms) output. Container outputs notifications and secondary capture images.”]</td></tr>
<tr><td>24</td><td>Alicja Raginis-Zborowska, Annalise-AI</td><td>20</td><td>2, Table 1</td><td>We refer to the table below:
<table><tr><th>Device (Company) [Previous Name]</th><th>Type of image</th><th>Compatible imaging</th><th>Exclusions (obtained from indications or contraindications in RFI or IFU)</th><th>Deployment Method</th><th>How are finding displayed</th><th>Where are the findings displayed</th><th>Additional features (as claimed by company)</th><th>Used in the NHS</th></tr><tr><td>Annalise Enterprise CXR and Annalise Container CXR, (Annalise.AI)</td><td>X-ray</td><td>CXR, anterior-posterior (AP) or posterior-anterior (PA) and optionally lateral (LAT) orientations</td><td>Is not to be used on patients under the age of 16 years for CXR.</td><td>Local or Cloud
Enterprise supports hybrid local-cloud deployment
Container hosted on partner platform</td><td>Annotated images with a notification</td><td>PACS Viewer or RIS and standalone application</td><td>Worklist Triage</td><td>Yes (48 Trusts)</td></tr></table>Suggested corrections:
• Deployment method: Change “Container hosted on partner platform” to “Container hosted on partner platform – local or cloud”
• How are findings displayed: Remove current text, replace with “Annotated images with optional notification”</td><td>Thank you for the corrections and clarifications. The table has been amended in the report.</td></tr>
</table>

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
				• Where are the findings displayed: Remove current text, replace with *"Annalise Enterprise: Synchronised Desktop app and/or DICOM Secondary Captures. Annalise Container: DICOM Secondary Captures"* • Additional features: Replace current text with *"Triage / prioritisation notifications to PACS or RIS reporting worklist. Device can also detect 123 additional findings on CXR, including spine related findings such as diffuse spinal osteophytes, kyphosis, osteopaenia, scoliosis, spinal fixation, spinal arthritis, spine lesion, and technical factors which may indicate poor image quality."*	
25	Alicja Raginis-Zborowska, Annalise-AI	33	Table 2	Population used to train and validate AI for Annalise.AI studies (Talwar, Annalise study AiC, Frias, Karusena, Ghatak) are reported in Seah *et al*, 2021(1).	Thank you for this comment. Unfortunately, none of the studies have explicitly reported the training or validation. However, we have added a footnote to state that the company have confirmed the training/validation of the AI with reference to the Seah et al. 2021 paper. The EAG note however that the study by Seah et al. is focused on the accuracy of chest x-ray interpretation (and was therefore excluded by the EAG, listed in Appendix A5). Spine wedge fracture is stated in the supplementary material of Seah et al. 2021 (AUC of unassisted radiologists: 0.719, and radiologists assisted by the AI: 0.857) however this is only 1 of 127 additional clinical findings and no other information on grade or definition of spine wedge fracture reported.
26	Alicja Raginis-Zborowska, Annalise-AI	33	Table 2, as well as section 5.1.1	• Annalise.ai abstract, Karusena *et al*, has been published in a peer-reviewed article, Jones *et al* 2021(2). [redacted] • Annalise.ai also recommends the addition of Seah *et al*, 2021 (1) to the evidence base used by the EAG. While the focus of this paper is for detecting abnormalities on chest-ray, the study data presents AI model (0.953), aided (0.857) and	Thank you for your comment. • Thank you for highlighting this full peer-reviewed publication. The EAG have replaced the abstract by Karunasena (2022) with the full paper by Jones (2021) throughout the report. The EAG note that the

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
				unaided (0.719) radiologist performance (AUC) for detecting spine wedge fracture (see supplementary materials available online), and which indicates the possible impact of the device as a detection support tool. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00106-0/fulltext	focus of the study is on chest findings. [redacted] • The scope of the paper by Seah et al. 2021 is focused on interpretation of chest findings. The EAG has re-reviewed the paper and has considered its exclusion is appropriate, as there is limited detail in relation to the decision problem of this EVA. Limited information was reported related to spines (definition of spine fracture not reported, only AUC with no reporting of sensitivity and specificity). Spine wedge fracture is bundled with 126 other clinical findings in supplementary material.
27	Alicja Raginis-Zborowska, Annalise-AI	37	5.1	We refer to the statement below: • "Of these 21 were diagnostic accuracy studies … and one was a survey of radiologists (Karunasena *et al*., 2022). Three studies were prospective (Bendtsen and Hitz, 2024; Connacher *et al*., 2019; Connor *et al*., 2024); of which 2 were available in abstract only." Suggestions • We recommend the EAG review / refer to the full published paper for further information (Jones *et al* 2021(2)). • This study was an observational real-world study, and thus should be considered prospective use of the device (recommend adjustment of number for prospective studies).	See response to comment 26. We have included Jones et al. 2021 (replacing Karunasena et al. 2022 throughout our report). Only 1 vertebral wedge fracture was identified as missed by the AI. However, the study acknowledges that no adjudication of discrepancies between radiologist and AI differences were conducted. Therefore, the EAG has focused on data extraction from the survey (user acceptability outcomes).

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
28	Alicja Raginis-Zborowska, Annalise-AI	44	Healthcare professional user acceptability	We recommend the EAG review / refer to the full published paper for further information (Jones *et al* 2021(2)). The survey results 9/10 radiologists believed they accuracy was improved with the model and were more positive towards following the deployment of AI.	Thank you for this comment. We have replaced the abstract by Karunasena et al. 2022 with the full paper by Jones et al. 2021 throughout the report. The EAG note that the focus of this study is on chest findings. Therefore, the EAG focused on outcomes reported in the full paper publication where it was clearer the relevance to the decision problem.
29	Alicja Raginis-Zborowska, Annalise-AI	44	Changes to clinical management	We recommend the EAG review / refer to the full published paper for further information (Jones *et al* 2021(2)): *43 cases (1.4%) had changed patient management and 29 cases (1.0%) had further imaging recommendations.*	Thank you for your comment. Changes to clinical management were not specifically related to identification of vertebral fractures (for which the AI tool missed 1 case, but this was not confirmed as classed as non-critical). Therefore, we have included the paper by Jones et al. 2021 in our report, but not extracted the outcomes specified in the comment.
30	Alicja Raginis-Zborowska, Annalise-AI	70	5.3	We refer to the statement below: “No evidence was specifically related to Annalise Container CXR; however, the EAG acknowledges that this is a cloudbased version of Enterprise CXR and the company stated that evidence should be considered generalisable between technologies.” Suggested corrections to product description: No evidence was specifically related to Annalise Container CXR; however, the EAG acknowledges that this is a *version of Enterprise which is packaged to optimise it for running on selected third-party marketplaces / platforms. Annalise Container utilises the same CXR AI model as Annalise Enterprise*, and the company *confirms* that evidence should be considered generalisable between technologies.	Thank for this comment. This change has been made in the report.
31	Alicja Raginis-Zborowska, Annalise-AI	101	6.3.2	We refer to the statement below:	Thank you for this comment. The EAG has considered a generic AI base case and cost per scan was increased in the sensitivity

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
				In the base case analysis, the use of Annalise.AI arm was £ ****** more expensive than standard of care (AI: £******, SoC: £*****) and resulted in ******** additional QALYs (AI: , *SoC:*) giving an ICER of £**** close to the willingness to pay threshold of £20,000.1 [redacted] [redacted]	analysis (see section 6.3.1 of the EAG report).
32	Alicja Raginis-Zborowska, Annalise-AI	117	8.1	We refer to the lack of health economics evidence [redacted] [redacted]	Thank you for your comment. Additional evidence can be submitted to NICE during public consultation.

<table>
<tr><th>Comment no.</th><th>Stakeholder</th><th>Page no.</th><th>Section no.</th><th>Comment</th><th>EAG Response</th></tr>
<tr>
<td>33</td>
<td>Alicja Raginis-Zborowska, Annalise-AI</td>
<td>118</td>
<td>8.2 Table 29</td>
<td>
<table>
<tr><th>Outcomes</th><th>Annalise Enterprise (CXR) (Annalise.AI)</th><th>Annalise Container (CXR) (Annalise.AI)</th><th>BoneView (Gleamer)</th><th>BriefCase-Triage (Aidoc Medical)</th><th>CINA-VCF Quantix (Avicenna.AI)</th><th>HealthVCF (Nanox AI) (previously Zebra medical)</th><th>HealthOST (Nanox AI) (previously Zebra medical)</th><th>IB Lab FLAMINGO (Powered by UCB's, BoneBot AI model)</th><th>TechCare Spine (Milvue)</th></tr>
<tr><td>Diagnostic accuracy in detection of vertebral fracture</td><td>GREEN</td><td>RED</td><td>AMBER (per vertebrae only)</td><td>GREEN</td><td>GREEN</td><td>GREEN</td><td>RED</td><td>GREEN</td><td>RED</td></tr>
<tr><td>Accuracy by HCP profession</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Failure rate or rate of inconclusive AI reports</td><td>GREEN</td><td>RED</td><td>AMBER</td><td>GREEN</td><td>RED</td><td>GREEN</td><td>RED</td><td>GREEN</td><td>RED</td></tr>
<tr><td>Number of missed fractures</td><td>GREEN</td><td>RED</td><td>RED</td><td>RED</td><td>GREEN</td><td>GREEN</td><td>RED</td><td>AMBER (1 CiC study only)</td><td>RED</td></tr>
<tr><td>Rate of missed fracture-related further injury</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Proportion of people that need further imaging</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Intervention related adverse events</td><td>GREEN (no events identified)</td><td>GREEN (no events identified)</td><td>GREEN (no events identified)</td><td>GREEN (no events identified)</td><td>GREEN (no events identified)</td><td>GREEN (no events identified)</td><td>GREEN (no events identified)</td><td>GREEN (no events identified)</td><td>GREEN (no events identified)</td></tr>
<tr><td>HCP user acceptability of AI tools</td><td>GREEN</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Changes to clinical management</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>GREEN</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Health-related QoL</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Cost of the AI software</td><td>GREEN</td><td>GREEN</td><td>RED</td><td>GREEN</td><td>GREEN</td><td>GREEN</td><td>GREEN</td><td>GREEN</td><td>GREEN</td></tr>
<tr><td>Staff costs</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Training and implementation</td><td>GREEN</td><td>GREEN</td><td>GREEN</td><td>GREEN</td><td>GREEN</td><td>GREEN</td><td>GREEN</td><td>GREEN</td><td>GREEN</td></tr>
<tr><td>Other downstream costs for diagnosis or treatment</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Time to produce a radiography report</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>AMBER (time to analyse results from software)</td><td>RED</td><td>AMBER (time to run software)</td><td>RED</td></tr>
<tr><td>Time to diagnosis or time to definitive radiology report</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Time to further referral or treatment</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Number of treatments and extent of treatments</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Number of hospital appointment/visits</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Number of hospital admissions</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
<tr><td>Type of healthcare professional interpreting the radiograph</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td><td>RED</td></tr>
</table>
Abbreviations: AI, Artificial intelligence; HCP, Healthcare professional; QoL, quality of life.

We refer to the below:
Lack of evidence (RED) for Annalise Container CXR for sections
- Diagnostic accuracy in detection of vertebral fracture
- Failure rate or rate of inconclusive AI reports
</td>
<td>The EAG note the a priori evidence search criteria were described in the published protocol and the EAG report.

See response to comment 30. Table 29 is an evidence gap analysis – the EAG has not identified any published evidence explicitly using Annalise Container CXR.

The EAG has considered the additional evidence stated in the consultation comment:
1. Seah et al. 2021 is focused on chest X-ray interpretation and findings. No accuracy by HCP was reported for vertebral fractures. Therefore, the EAG has kept “Accuracy by HCP” as relevant to the decision problem as RED.
[redacted]
2. Jones et al. 2021 refers to clinical management changes relating to clinical chest diagnoses. Only one vertebral wedge fracture was mentioned in this report (was missed by the AI model), and the authors confirmed that none of the differences between AI and radiologist were adjudicated. No change to clinical management specific to this 1 vertebral fracture cases were noted. Therefore, the EAG has kept “changes to clinical management” relevant to the decision problem as RED.</td>
</tr>
</table>

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
				- HCP user acceptability of AI tools Suggestion: No evidence was specifically related to Annalise Container CXR; however, the EAG acknowledges that this is a *version of Enterprise which is packaged to optimise it for running on selected third-party marketplaces / platforms. Annalise Container utilises the same CXR AI model as Annalise Enterprise*, and the company *confirms* that evidence should be considered generalisable between technologies. Area of evidence that should be considered and flagged as **GREEN:** 1. **Accuracy by HCP**: The CXR model validation study (Seah *et al*, 2021(1)) provides evidence of change in radiologist performance in the detection of spinal wedge fracture. [redacted] 2. **Changes to clinical management**: While not specific to the VFF finding, Jones *et al* 2021(2) (which is the full publication of the abstract considered by the EAG, Karunasera), indicated of 2972 cases reviewed with the model, 92 cases (3.1%) had significant report changes, 43 cases (1.4%) had changed patient management and 29 cases (1.0%) had further imaging recommendations. [redacted] Deployment of Annalise Enterprise addresses the evidence gap by providing a structured, clinically validated AI implementation framework that supports staff	[redacted] Shelmerdine et al. 2024. Describes a narrative review of selecting an appropriate AI tool, and briefly mentioned a case study of AI implementation in South West London. Whilst this outlines key considerations when selecting an AI tool (from the many available) it is not specific to vertebral fractures, does not report any results, and is a narrative review to guide product selection and therefore is excluded from clinical and economic outcome sections due to lack of direct relevance to the decision problem of the EVA. The EAG has added this paper to the Appendix A5 of exclusions (as lacks direct relevance to the decision problem), however has added to section 7 where it discusses integration in the NHS.

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
				training and standardizes adoption, enabling clinicians to make faster, more accurate diagnoses and reducing variability in patient care across NHS settings as per Shelmerdine *et al* 20241 (3) paper.	
34	Alicja Raginis-Zborowska, Annalise-AI	20	8.2	We refer to the statement below: Published evidence is not available for all technologies (no evidence for Annalise Container CXR (Annalise.AI). There is a general lack of transparent reporting of the software name, version number, and configuration settings which may influence results. Suggestion: No evidence was specifically related to Annalise Container CXR; however, the EAG acknowledges that this is a *version of Enterprise which is packaged to optimise it for running on selected third-party marketplaces / platforms. Annalise Container utilises the same CXR AI model as Annalise Enterprise*, and the company *confirms* that evidence should be considered generalisable between technologies. As a part of our EU MDR and MHRA approvals, Annalise Container is considered equivalent to Annalise Enterprise and the evidence for the Enterprise device is considered generalisable to the Container. The same CXR model is deployed on both devices. Therefore, the Annalise Enterprise column should be applied to the Annalise Container column too.	See response to comment 30. Table 29 is an evidence gap analysis – the EAG has not identified any published evidence explicitly using Annalise Container CXR. The EAG has also added to section 4.2 that the company have confirmed generalisability between Enterprise and Container (similar approach taken for other technologies).
35	Alicja Raginis-Zborowska, Annalise-AI	180		We refer to the statement below: Yes - DICOM PatientAGE: ≥ 016Y, Modality Header: CX, DX, DR BodyPartExamined, StudyDescription, SeriesDescription, ProtocolName: CHESPORT_CHEST or is missing and does not contain terms not supported SeriesDescription: CHEST, AP, PA,LAT,CXR,Thorax or is missing Xxxxxxxxxxxxxxxxxxxxxxx Annalise recommended change:	Thank you for the recommended change. This has been reflected in the report

<table>
<tr><th>Comment no.</th><th>Stakeholder</th><th>Page no.</th><th>Section no.</th><th>Comment</th><th>EAG Response</th></tr>
<tr><td></td><td></td><td></td><td></td><td>Yes – DICOM. Automatic fetch or forward from PACS, filtered by Study Description, Body Part Examined or Exam/Procedure Code, or a combination of these.
Patient Age: ≥ 016Y, Modality Header: CR or DX DICOM SOP Class only
Study must contain at least one frontal projection (AP or PA) and may optionally contain two AP/PA and lateral projections.</td><td></td></tr>
<tr><td>36</td><td>Alicja Raginis-Zborowska, Annalise-AI</td><td>162</td><td></td><td><table><tr><td>12.</td><td>Karunasena (J Med Image Radia Oncol, 2022; 95)
Abstract

Funding: NR

Declaration of interests: NR</td><td>Survey (n=11 radiologists in pilot for 6 weeks, and then 3-months after network-wide implementation feedback was sought from n=63 radiologists)

Intervention: Annalise Enterprise CXR
GREEN</td><td>Inclusion: NR

Exclusion: NR

Image acquisition: NR

Setting: Australia, hospitals and clinicians (N=250)</td><td>Radiologist reporting of subjective improved accuracy, changes to reporting, ease of use, how well integrated, reporting quality and time.
AMBER</td><td>Abstract only, lack of detailed methodology.

<u>Training</u>: NR

<u>Validation</u>: NR</td></tr></table>
We refer to the statement:
Abstract only, lack of detailed methodology. Training: NR; Validation NR.

Annalise comment:

The model was trained and validated on the population outlined in the Seah et al. Study, as referenced in Jones et al, 2021 which is the full published paper relevant to the Karunasena abstract. For your reference, the full publication is available here: https://bmjopen.bmj.com/content/11/12/e052902. We recommend updating the report to reflect the availability of this peer-reviewed evidence.</td><td>Thank you for your comment. The abstract by Karunasena et al. 2022 has been replaced with Jones et al. 2021 where training and validation was explicitly reported. The EAG have also added to the study characteristics table (Appendix A4) that: “At stakeholder consultation the company stated that the population used to train and validate the AI for this study was reported in Seah et al. 2021” to the studies by Talwar et al. 2023, Annalise.AI [AiC], Frias et al. 2023, Jones et al. 2021. This was already stated for Ghatak et al. 2024.</td></tr>
<tr><td>37</td><td>Alicja Raginis-Zborowska, Annalise-AI</td><td>170</td><td></td><td><table><tr><td>21.</td><td>Talwar (RSNA, 2023; W5B-SPCH-2)
Abstract

Funding: NR

Declaration of interests: Author reported nothing to disclose.</td><td>Retrospective diagnostic accuracy study (n=1,559 chest x-rays)

Intervention: Annalise Enterprise CXR
GREEN

Comparator: original radiology reporter (n=NR)</td><td>Inclusion: consecutive adults (aged 18 years and older)
Exclusion: NR
GREEN

Image acquisition: Chest X-rays acquired in 2016 (de-identified).

Setting: Australia (N=1)</td><td>Successful AI processing of image. 60 clinical findings (including pulmonary nodules, pleural effusions, spinal compression fractures, airspace opacities, acute rib fractures).
AMBER</td><td>Abstract only, lack of detailed methodology.

<u>Training</u>: NR

<u>Validation</u>: NR</td></tr><tr><td></td><td></td><td>Reference standard: a radiologist with 10 years or more experience reviewed discrepancies between intervention and comparator
AMBER</td><td></td><td></td><td></td></tr></table></td><td>See response to comment #36.</td></tr>
</table>

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
				We refer to the statement: Abstract only, lack of detailed methodology. Training: NR; Validation NR. Annalise comment: The model was trained and validated on the population outlined in the *Seah et al*. study. For your reference, the full publication(1) is available here: https://bmjopen.bmj.com/content/11/12/e052902. We recommend updating the report to reflect the availability of this peer-reviewed evidence.	
38	Alicja Raginis-Zborowska, Annalise-AI	177	Appendix A5	search | J, 2023; 3452) 55 | Annalise.AI | Annalise.AI | Seah (Lancet Digital Health, 2021; e496-e506) | Full text publication | Outcome: focus on chest findings We refer to the statement below: Outcome: focus on chest findings. Suggestion: We respectfully recommend reconsidering the exclusion of the Seah *et al*.(1) paper describing the validation of the Annalise CXR device. While we understand the rationale , it is important to note that the Annalise CXR model was explicitly developed and validated to detect a wide range of findings visible on chest X-rays—including spine-related findings visible on CXR, including diffuse spinal osteophytes, kyphosis, osteopaenia, scoliosis, spinal fixation, spinal arthritis, and spine lesions, as well as technical factors indicative of poor image quality. The validation study in question, a large-scale multi-reader, multi-case trial, assessed the model's performance across all findings included in the algorithm, not just those traditionally considered chest-specific. The model was trained on a dataset of over 782,000 studies, and the validation set included a 12.83% prevalence of vertebral fracture findings	See response to comment #26. Validation and training was queried with companies and reported in Appendix C1; we have added additional context and reference to Seah et al. 2021 to Appendix C1.

Comment no.	Stakeholder	Page no.	Section no.	Comment	EAG Response
				across 2,565 cases. This comprised 329 cases of spinal wedge fractures and 292 cases of osteopenia (see Appendix C1, page 180), demonstrating the inclusion of non-chest pathologies in the model's scope. As such, excluding this paper on the basis that it focuses only on chest findings does not fully represent the device's validated capabilities.	
39	Alicja Raginis-Zborowska, Annalise-AI		General notes after their comments table	1. 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 Aug;3(8):e496–506. 2. 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 Dec 20;11(12):e052902. 3. Shelmerdine SC, Togher D, Rickaby S, Dean G. Artificial intelligence (AI) implementation within the National Health Service (NHS): the South West London AI Working Group experience. Clin Radiol. 2024 Sep;79(9):665–72.	See response to comment #26 and #33. Thank you for your support and time in responding to the work.

NICE National Institute for Health and Care Excellence

Health Tech Programme

GID-HTE10059 Artificial intelligence technologies to aid the opportunistic detection of vertebral fragility fractures: Early Value Assessment

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**	[redacted]
2. Name of organisation	The Society and College 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).	The Society of Radiographers (SoR) is funded by membership fees – it is a professional body and trade union that represents radiographers, sonographers, and other non-medical imaging professionals including assistant practitioners. The SoR focuses on promoting the science and practice of radiography, setting professional standards, and advocating for its members' interests in the workplace and educationally. The College of Radiographers (CoR) is a charitable subsidiary, dedicated to education, research, and public benefit. The College is funded through various means including grants and donations. The College is also financed from legacy funds, such as the Valerie Carr Award, which supports pre-registration learners. The Society and College operate as a nonprofit organisation.
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 20 February 2025. **If so, please state the name of company, amount, and purpose of funding.**	None.

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.)	Incidental diagnosis of previously undetected/suspected vertebral fracture/s.
7. In your view, is there an unmet need for patients and healthcare professionals in this condition?	Yes, in terms of current levels of diagnosis and treatment of vertebral fragility fractures. That may not necessarily be met by the introduction of additional technology.

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

8. How is the condition currently treated or detected in the NHS? Is there a gold standard for diagnosis of VFFs?	NICE guidance and Royal Osteoporosis Society (noted in NICE topic guidance).
9a. Are any relevant clinical guidelines we should be aware of, and if so, which?	Not in addition to those already noted in 8.
9b. Is the pathway of care well defined? Does it vary or are there differences of opinion between professionals across the NHS?	The pathway of care does vary but not necessarily due to differences of opinion between professionals – this is related to regional differences in the provision of services for fracture liaison services, ongoing surveillance / further imaging, and treatment.

(Please state if your experience is from outside England.)	
9c. What impact would the technology have on the current pathway of care?	Likely to increase referrals to onward services. We acknowledge that there is a potential for overdiagnosis/treatment in the population although that may be balanced in contrast with the underdiagnosis of vertebral fracture in individual people.
10a. How does healthcare resource use differ between the technology and current care?	This would triage additional patients to services.
10b. In what clinical setting should the technology be used? (For example, should it be restricted to or prioritised in specialist clinics.)	Assessment for fracture should be routine across all populations i.e. assessment of whole radiograph / scan images with view to checking for vertebral fractures. We recognise that this is particularly crucial for people at high risk of fracture including for example women with early or surgical menopause, people taking medication associated with osteopenia, people with genetic disposition to osteopenia, and all people who have suffered minimal trauma fractures - albeit that information is not always available to the reporter – therefore it is sensible to suggest that all images should be reviewed to include assessment of the spine when this is demonstrated.
10c. What investment is needed to introduce the technology? (For example, for facilities, equipment, or training.)	Adequate infrastructure to implement the technology. Funding to procure, and also sustain payment for the service over a sustained period / ongoing costs including payment for the technology/software, algorithmic audit /surveillance of performance. Establishment of pathways for referral – not universal across the NHS.
11. Do you expect the technology to provide clinically meaningful benefits compared with current care?	Please see An-evaluation-of-Fracture-Liaison-Services-in-the-.pdf ‘Radiological under-reporting and non-standardised assessment of fragility fractures still persist, with low numbers of patients undergoing a risk assessment and treatment for secondary prevention of fracture. In order to improve care for these patients, the reporting of vertebral fractures must be improved and standardised in order to identify patients at increased risk of secondary fragility fractures’. This does not necessarily need to be provided by technology and could potentially be achieved by adequate training, governance and requirements of reporters to include review of anatomy for evidence of vertebral fragility fracture in reports which include imaging of the spine.
11a. Do you expect the technology to increase length of life?	No, to increase quality of life.
11b. Do you expect the technology to increase health-related quality of life?	Yes.

12. Are there any groups of people for whom the technology would be more or less effective (or appropriate) than the general population?	Yes, please see point 10b, although this distinction should be made with caution; 'Presumptions about patients being at 'low-risk', for example due to lifestyle factors such as eating healthily and undertaking exercise are often disproven with a diagnosis of osteoporosis because our genes determine so much of our osteoporosis risk.' Investigation of Unique and Shared Gene Effects on Speed of Sound and Bone Density Using Axial Transmission Quantitative Ultrasound and DXA in Twins* | Journal of Bone and Mineral Research | Oxford Academic. Also Fragile--Please-handle-with-care_2018_yradi.pdf Also see A-retrospective-service-evaluation-of-patient-awar.pdf 'Further research into the increased number of men with opportunistically identified vertebral fragility fracture is recommended to ensure the pathway is efficient and to review potential barriers to diagnosis.'

The use of the technology

13. Will the technology be easier or more difficult to use for patients or 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.)	Additional clinical requirements are the need for onward referral, treatment and pain relief, specialist advice and assessment, for example with the universal provision of fracture liaison services. This would raise the issue of thorough assessment which may include evaluation of imaging modalities and staff developement: 'SCT BMD measurement has the potential to be developed as a screening tool for osteoporosis within the fracture liaison service (FLS). This could aid in the identification of patients with osteoporosis and address the current treatment gap. Nonetheless, many factors must be considered for this application including staff training, radiation protection and patient engagement with the screening programme.' The-accuracy-and-clinical-utility-of-spectral-CT-b.pdf 'Diffusion Weighted MRI (DWI) revolutionizes vertebral compression fracture diagnosis, distinguishing between benign and malignant cases. This precision guides treatment decisions, minimizing the necessity for invasive procedures like biopsy. As a safe and reliable imaging method, DWI elevates patient care, ensuring accurate diagnostics and improved outcomes'. The-role-of-diffusion-weighted-MRI-in-the-accurate.pdf

	‘The role of the radiographer working within DXA and osteoporosis services is evolving and is an exciting area of advanced practice. Promoting this specialty within radiography may help to improve job satisfaction as well as recruitment and retention rates. As radiographers scope of practice in osteoporosis changes and evolves, it is hoped that current legislation may change to allow independent prescribing for diagnostic radiographers, which can in turn streamline patient pathways and reduce the burden on primary and secondary care.’ The-role-of-the-Radiographer-in-osteoporosis-and-f.pdf
15. 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-adjusted life year (QALY) calculation?	No
16. 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?	This is an alternative to reporters focusing on vertebral fractures. There is a potential risk that reporters further deskill in ability to diagnose this type of fracture.
17. Does the use of the technology address any particular unmet need of the patient population?	As per topic overview provided by NICE, yes.

Sources of evidence

18a. Are there any recent or ongoing clinical trials that concern the technology? Do these reflect current UK clinical practice?	---
18b. What, in your view, are the most important outcomes, and were they measured in trials?	----

18c. If surrogate outcome measures were used, do they adequately predict long-term clinical outcomes?	----
19. Are you aware of any relevant evidence that might not be found by a systematic review of the trial evidence?	Not in addition to those listed in this return.
20. How do data on real-world experience compare with the available data?	----

Equality

22a. Are there any potential equality issues that should be taken into account when considering this treatment?	Access to services that use the technology in contrast to patients who live in areas that are not using the software. 'Thought and care needs to be taken when integrating predictive software into practice. Focus on empowering patients, providing information on processes and results are key.' Understanding-patient-views-and-acceptability-of-p.pdf
22b. Consider whether these issues are different from issues with current care and why.	Similar state across UK in terms of patchy evidence, implementation, validation etc.

Topic-specific questions

23. What is the prevalence of steroid-induced osteoporosis?	---
24. Should radiographic images taken through all imaging methods (X-ray, CT, MRI) be considered for use with the technology?	That may be advantageous to patients who are unable to tolerate imaging, for example, unable to lay recumbent for CT/MR cross-sectional imaging or claustrophobic people for whom MR Scan is distressing.
25. Are the radiographic images taken for a specific condition likely to be particularly useful? For example, if they contain more vertebrae or the vertebrae where fragility fractures occur more commonly.	Yes, that does make sense but would not cover all instances of vertebral fracture for each and every individual. The assessment should also consider person-centred care and shared decision-making in line with Montgomery Principles.

Thank you for your time.

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

Your privacy

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 or NO

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