Supplementary Appendix

Commercial AI for Opportunistic Detection of Vertebral Compression Fractures on Routine CT: A Systematic Review of Diagnostic Performance and Clinical Yield

A. Srour, M. Omar, Y. Barash, D. Litmanovich, M. Srour-Asmar, E. Klang*, A. Gorenshtein*

** These authors jointly supervised this work and contributed equally as senior authors.*

Table of Contents

Supplementary Appendix 1. Electronic Search Strategies

The following electronic search strategies were developed using a three-block Boolean structure: (1) artificial intelligence / commercial software, (2) vertebral compression fracture, and (3) computed tomography. The opportunistic / incidental setting was applied as a screening criterion rather than a search filter, to maximise sensitivity and avoid missing relevant studies that do not explicitly use those terms. Database-specific syntax, field tags, and MeSH / thesaurus terms were applied throughout. All searches were executed on 31 May 2026. No language restrictions were applied at the search stage; language was applied as an inclusion criterion during full-text screening.

S1.1 PubMed

```
(
  /* Block 1 — Artificial intelligence / commercial software */
  "Artificial Intelligence"[MeSH] OR "Machine Learning"[MeSH] OR
  "Deep Learning"[MeSH] OR "Neural Networks, Computer"[MeSH] OR
  "Image Interpretation, Computer-Assisted"[MeSH] OR
  artificial intelligence[tiab] OR machine learning[tiab] OR
  deep learning[tiab] OR neural network*[tiab] OR
  convolutional neural network*[tiab] OR automated detection[tiab] OR
  "computer-aided detection"[tiab] OR commercial software[tiab] OR
  "commercially available"[tiab] OR proprietary[tiab] OR
  "FDA cleared"[tiab] OR "FDA approved"[tiab] OR
  "CE marked"[tiab] OR "CE certified"[tiab] OR
  "regulatory cleared"[tiab] OR algorithm*[tiab] OR radiomics[tiab] OR
  HealthVCF[tiab] OR HealthOST[tiab] OR Nanox[tiab] OR
  "Zebra Medical"[tiab] OR Coreline[tiab] OR AVIEW[tiab]
)
AND
(
  /* Block 2 — Vertebral compression fracture */
  "Spinal Fractures"[MeSH] OR "Fractures, Compression"[MeSH] OR
  "Osteoporotic Fractures"[MeSH] OR
  "vertebral fracture"[tiab] OR "vertebral fractures"[tiab] OR
  "vertebral compression fracture"[tiab] OR "vertebral compression fractures"[tiab] OR
  "spinal fracture"[tiab] OR "spinal fractures"[tiab] OR VCF[tiab] OR
  "osteoporotic vertebral fracture"[tiab] OR "osteoporotic vertebral fractures"[tiab] OR
  "vertebral deformity"[tiab] OR "vertebral deformities"[tiab] OR
  "vertebral collapse"[tiab] OR "compression fracture"[tiab] OR
  "compression fractures"[tiab] OR "vertebral height loss"[tiab] OR Genant[tiab]
)
AND
(
  /* Block 3 — Computed tomography */
  "Tomography, X-Ray Computed"[MeSH] OR "computed tomography"[tiab] OR
  "CT scan"[tiab] OR "CT scans"[tiab] OR MDCT[tiab] OR MSCT[tiab]
)
AND
(
  "2017/01/01"[Date - Publication] : "3000"[Date - Publication]
)
```

PubMed total: 638 records

S1.2 Scopus

```
TITLE-ABS-KEY (
  /* Block 1 — Artificial intelligence / commercial software */
  (
    "artificial intelligence" OR "machine learning" OR "deep learning" OR
    "neural network*" OR "convolutional neural network*" OR
    "automated detection" OR "computer-aided detection" OR
    "commercial software" OR "commercially available" OR proprietary OR
    "FDA cleared" OR "FDA approved" OR "CE marked" OR "CE certified" OR
    "regulatory cleared" OR algorithm* OR radiomics OR
    HealthVCF OR HealthOST OR Nanox OR "Zebra Medical" OR Coreline OR AVIEW
  )
  AND
  /* Block 2 — Vertebral compression fracture */
  (
    "vertebral fracture*" OR "vertebral compression fracture*" OR
    "spinal fracture*" OR VCF OR "osteoporotic vertebral fracture*" OR
    "vertebral deformit*" OR "vertebral collapse" OR
    "compression fracture*" OR "vertebral height loss" OR Genant
  )
  AND
  /* Block 3 — Computed tomography */
  (
    "computed tomography" OR CT OR MDCT OR MSCT
  )
) AND DOCTYPE ( ar OR re ) AND PUBYEAR > 2016
```

Scopus total: 309 records

S1.3 Web of Science

```
/* Field: Topic (TS) — selected via dropdown; TS= prefix omitted */
(
  (
    "artificial intelligence" OR "machine learning" OR "deep learning" OR
    "neural network*" OR "convolutional neural network*" OR
    "automated detection" OR "computer-aided detection" OR
    "commercial software" OR "commercially available" OR proprietary OR
    "FDA cleared" OR "FDA approved" OR "CE marked" OR "CE certified" OR
    "regulatory cleared" OR algorithm* OR radiomics OR
    HealthVCF OR HealthOST OR Nanox OR "Zebra Medical" OR Coreline OR AVIEW
  )
  AND
  (
    "vertebral fracture*" OR "vertebral compression fracture*" OR
    "spinal fracture*" OR VCF OR "osteoporotic vertebral fracture*" OR
    "vertebral deformity" OR "vertebral collapse" OR
    "compression fracture*" OR "vertebral height loss" OR Genant
  )
  AND
  (
    "computed tomography" OR CT OR MDCT OR MSCT
  )
) AND DT=("Article" OR "Review" OR "Early Access") AND PY=(2017-2030)
```

Web of Science total: 294 records

Database	Date restriction	Records retrieved
PubMed	2017–present	638
Scopus	2017–present	309
Web of Science	2017–present	294
Total		1,241

Table S1. *Summary of database search yields. All searches restricted to publications from 2017 onwards. The Web of Science Topic field was selected via dropdown (TS= prefix omitted). The opportunistic / incidental setting was applied as a screening criterion rather than a search filter. Total retrieved before deduplication: 1,241 records; duplicates were removed prior to title and abstract screening.*

Supplementary Appendix 2. PRISMA 2020 Checklist

Completed PRISMA 2020 checklist (Page MJ, McKenzie JE, Bossuyt PM, et al. BMJ 2021;372:n71) for the systematic review "Commercial AI for Opportunistic Detection of Vertebral Compression Fractures on Routine CT." Location references cite manuscript section headings and Supplementary Appendix sections; manuscript page numbers are deliberately omitted because the journal paginates the accepted article, and authors will map section references to final pages at the proof stage.

Section and Topic	Item	Checklist item	Location where item is reported
Title	1	Identify the report as a systematic review.	Title ("… A Systematic Review of Diagnostic Performance and Clinical Yield").
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Structured Abstract (Background, Purpose, Methods, Results, Conclusions); PROSPERO registration stated.
Introduction	3	Describe the rationale for the review in the context of existing knowledge.	Introduction, paragraphs 1–3 (detection gap; emergence of commercial AI; fragmented evidence).
	4	Provide an explicit statement of the objective(s) the review addresses.	Introduction, final paragraph (aim: evaluate diagnostic accuracy, real-world yield, and downstream outcomes of commercial AI for VCF detection).
Methods	5	Specify the inclusion and exclusion criteria and how studies were grouped for the syntheses.	Methods, Study Selection and Eligibility Criteria.
	6	Specify all databases, registers, websites, and other sources searched, with the date each was last searched.	Methods, Data Sources and Search Strategy (PubMed, Scopus, Web of Science; searched through 31 May 2026); Supplementary Appendix 1.
	7	Present the full search strategies for all databases, including any filters and limits used.	Supplementary Appendix 1 (complete database-specific Boolean strings, field tags, and limits).
	8	Specify the methods used to decide whether a study met the inclusion criteria, including number of reviewers and whether they worked independently.	Methods, Study Selection (two reviewers, A.S. and M.O., independent; E.K. adjudicator; no dedicated review-management software).
	9	Specify the methods used to collect data from reports, including number of reviewers and any processes for confirming data.	Methods, Data Extraction (two reviewers independently; pre-specified form; values not reported recorded as "not reported"; no imputation).
	10a	List and define all outcomes for which data were sought.	Methods, Data Extraction (sensitivity, specificity, PPV, NPV, AUC, fracture yield; AI-versus-radiologist comparison; downstream treatment outcomes).
	10b	List and define all other variables for which data were sought.	Methods, Data Extraction (study design and setting, country, demographics, CT region, AI product and version, regulatory status, reference standard, funding and COI).
	11	Specify the methods used to assess risk of bias, including the tool(s) used and number of reviewers.	Methods, Risk of Bias Assessment (QUADAS-2; two independent reviewers; consensus and E.K. arbitration); Supplementary Appendix 3.
	12	Specify for each outcome the effect measure(s) used in the synthesis or presentation of results.	Methods, Data Synthesis (diagnostic-accuracy estimates with 95% CIs presented per study; no pooled effect measure).
	13a	Describe the processes used to decide which studies were eligible for each synthesis.	Methods, Data Synthesis (structured narrative synthesis grouped by performance domain).
	13b	Describe any methods required to prepare the data for presentation or synthesis.	Methods, Data Extraction and Data Synthesis (no data conversion or imputation; metrics reported as in source articles).
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Methods, Data Synthesis; Table 1, Table 2; Figures 2–3.

Section and Topic	Item	Checklist item	Location where item is reported
	13d	Describe the methods used to synthesise results and the rationale. If meta-analysis was done, describe the model.	Methods, Data Synthesis (narrative synthesis; meta-analysis not performed owing to heterogeneity in design, populations, platforms, reference standards, and metrics; rationale stated).
	13e	Describe any methods used to explore possible causes of heterogeneity.	Results, Sources of Heterogeneity (population, prevalence, threshold, analysis unit, reference standard, industry involvement).
	13f	Describe any sensitivity analyses conducted.	Not applicable — no quantitative pooling was performed; robustness is addressed narratively (Methods, Data Synthesis; Results, Sources of Heterogeneity).
	14	Describe any methods used to assess risk of bias due to missing results (reporting biases).	Methods, Risk of Bias Assessment (QUADAS-2 Flow and Timing domain); Discussion, Limitations (industry involvement; grey literature not searched).
	15	Describe any methods used to assess certainty in the body of evidence.	Methods, Risk of Bias Assessment (QUADAS-2 domain judgements); no formal GRADE assessment was undertaken (stated in Limitations).
Results	16a	Describe the results of the search and selection process, ideally using a flow diagram.	Results, Study Selection; Figure 1 (PRISMA 2020 flow diagram).
	16b	Cite studies that might appear to meet the inclusion criteria but were excluded, and explain why.	Results, Study Selection (24 full-text exclusions with reasons: commercial/regulatory status unconfirmed, n=12; abstracts/non-peer-reviewed, n=6; no VCF/accuracy data, n=6).
	17	Cite each included study and present its characteristics.	Results, Study Characteristics; Table 1 (eight included studies).
	18	Present assessments of risk of bias for each included study.	Results, Risk of Bias; Supplementary Figure 1; Supplementary Appendix 3 (per-study QUADAS-2 ratings).
	19	For all outcomes, present for each study summary statistics and an effect estimate with its precision.	Results, Diagnostic Performance; Table 2 (per-study sensitivity, specificity, PPV, NPV, AUC with 95% CIs).
	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Results, Risk of Bias and Diagnostic Performance subsections.
	20b	Present results of all statistical syntheses conducted.	Results, Diagnostic Performance (narrative ranges and per-study estimates; no meta-analysis performed).
	20c	Present results of all investigations of possible causes of heterogeneity.	Results, Sources of Heterogeneity.
	20d	Present results of all sensitivity analyses conducted.	Not applicable (no quantitative synthesis).
	21	Present assessments of risk of bias due to missing results for each synthesis.	Discussion, Limitations (reporting-bias and industry-involvement caveats).
	22	Present assessments of certainty in the body of evidence for each outcome.	Results, Risk of Bias; Discussion (retrospective-only evidence base; no formal GRADE).
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Discussion, paragraphs 1–2 and Relation to Existing Literature.
	23b	Discuss any limitations of the evidence included in the review.	Discussion, Limitations (study-level: retrospective designs, no outcome endpoints, industry involvement).
	23c	Discuss any limitations of the review processes used.	Discussion, Limitations (review-level: heterogeneity precluding meta-analysis; grey literature, registries, and manufacturer websites not systematically searched).
	23d	Discuss implications of the results for practice, policy, and future research.	Discussion, Future Directions; Conclusions (supervised screening within osteoporosis care pathways; need for prospective outcome studies).

Section and Topic	Item	Checklist item	Location where item is reported
Other information	24a	Provide registration information for the review, including register name and registration number.	Methods, opening paragraph: PROSPERO CRD420261423256.
	24b	Indicate where the review protocol can be accessed, or state that one was not prepared.	Methods, opening paragraph (registered PROSPERO protocol, publicly accessible).
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	Methods, opening paragraph ("conducted in accordance with the registered protocol; no amendments were made after registration").
	25	Describe sources of financial or non-financial support and the role of funders.	Acknowledgements ("the authors received no specific funding for this work").
	26	Declare any competing interests of review authors.	Competing interests statement ("the authors declare that they have no competing interests").
	27	Report which of the following are publicly available and where: data collection forms; extracted data; analytic code; other materials.	Extracted per-study data are reported in Table 1, Table 2, and Supplementary Appendix 3; the extraction form and dataset are available from the corresponding author on reasonable request.

Note. The review used narrative synthesis rather than meta-analysis because of heterogeneity across study designs, populations, AI platforms, reference standards, and outcome metrics; items 13d, 20b, and 20d are therefore reported as per-study estimates and narrative ranges rather than pooled effect sizes. No formal GRADE assessment was performed; certainty is characterised through the QUADAS-2 domain judgements (Supplementary Appendix 3). Risk of bias due to missing results was considered qualitatively within the QUADAS-2 Flow and Timing domain.

Supplementary Appendix 3. Adapted QUADAS-2 Instrument and Per-Study Risk of Bias

Risk of bias and applicability were assessed with QUADAS-2 (Whiting PF, Rutjes AWS, Westwood ME, et al. Ann Intern Med 2011;155:529–536), the standard instrument for diagnostic-accuracy studies. Four domains were evaluated: Patient Selection, Index Test, Reference Standard, and Flow and Timing. The index test was defined as commercially available, regulatory-cleared AI software applied to CT images for opportunistic VCF detection. Each domain was rated as low risk, high risk, or some concerns, and an overall judgement was assigned per study using the worst domain-level rating. Two reviewers (A.S. and M.O.) independently applied the instrument; disagreements were resolved through consensus, with E.K. serving as arbitrator when required. Per-study domain ratings for all eight included studies are reported in Table S2 and visualised in Supplementary Figure 1.

Because all included evaluations were retrospective and several relied on non-independent or partial reference standards, the risk-of-bias profile bears directly on the safety case for clinical deployment. The index-test domain was uniformly low risk, indicating that the AI systems themselves were applied with pre-specified, regulatory-cleared thresholds; however, the concentration of concerns in the reference-standard and flow-and-timing domains marks precisely where the current evidence is weakest and where prospective, independently adjudicated validation is most needed before unsupervised use. These judgements should therefore temper inferences about real-world performance and reinforce the supervised-deployment model recommended in the main text.

A. Domain-level assessment

Domain	Concern	QUADAS-2 signaling questions	High-risk indicators
D1. Patient Selection	Selection bias and applicability of the cohort to routine opportunistic CT practice.	Was a consecutive or random sample of patients/CT examinations enrolled? Was a case-control design avoided? Did the study avoid inappropriate exclusions?	Post-hoc enrichment of the analytic cohort with confirmed-fracture scans; highly selected single-population cohort (e.g., dedicated falls or oncology) limiting generalisability; non-consecutive or convenience sampling; cohort fracture prevalence atypical of routine opportunistic CT.
D2. Index Test (Commercial AI Software)	Bias from ambiguous software specification, threshold setting, or non-independent interpretation.	Were the AI results interpreted without knowledge of the reference standard? Was the detection threshold pre-specified?	Unclear software product, version, or operating threshold; threshold tuned post-hoc on the test data; AI output not interpreted independently of the reference standard; undisclosed configuration favouring sensitivity or specificity.
D3. Reference Standard	Bias from a reference standard that may misclassify VCF or depend on the AI output.	Is the reference standard likely to correctly classify VCF (expert consensus or standardised Genant grading)? Was it interpreted without knowledge of the AI output?	Routine clinical radiology reports used as the reference standard (known to under-report VCF); single reader without adjudication; differential or partial verification (only AI-positive cases re-read); reference standard not independent of the index test.
D4. Flow and Timing	Bias from inconsistent case handling, attrition, or partial verification.	Did all patients receive the same reference standard? Were all enrolled patients included in the analysis?	Algorithm-failure or technical exclusions omitted from the denominator (may inflate sensitivity); partial verification; different cohorts for the AI and comparator arms; unplanned protocol deviations.

B. Overall rating scheme

Overall risk-of-bias rating scheme. Low risk: all four domains rated low risk. Some concerns: at least one domain rated some concerns, with no domain rated high risk. High risk: any domain rated high risk. (QUADAS-2 conventionally uses "low / high / unclear"; consistent with current diagnostic-accuracy reporting practice, "some concerns" is used here in place of "unclear" to denote partial but non-fatal limitations.)

C. Per-study risk-of-bias ratings

Study	D1 Patient selection	D2 Index test	D3 Reference standard	D4 Flow & timing	Overall	Predominant concern
Page 2023	● High	● Low	● Low	● High	● High	Post-hoc cohort enrichment; retrospective flow
Kolanu 2020	● Low	● Low	● Some concerns	● Some concerns	● Some concerns	Partial verification; pre-existing report baseline
Fernandes-Pereira 2024	● Some concerns	● Low	● Low	● Some concerns	● Some concerns	Population generalisability; retrospective flow
Bendtsen 2024	● Low	● Low	● Some concerns	● High	● High	Algorithm-failure exclusions; radiographer readers
Spångeus 2025	● Some concerns	● Low	● Low	● Some concerns	● Some concerns	Enriched geriatric cohort; scan-level analysis
Dai 2025	● Low	● Low	● Low	● Low	● Low	None
Mathew 2025	● Low	● Low	● Some concerns	● Some concerns	● Some concerns	Joint-consensus reference; vertebra-level denominator
Behanova 2026	● Some concerns	● Low	● High	● Some concerns	● High	Routine reports as reference standard (differential verification)

Table S2. *Per-study QUADAS-2 risk-of-bias and applicability judgements for the eight included studies. D1–D4 denote the four QUADAS-2 domains. A graphical summary is provided in Supplementary Figure 1.*

● Low risk ● Some concerns ● High risk

Across the eight included studies, the index-test domain was rated low risk in every study, reflecting uniform use of pre-specified, regulatory-cleared thresholds applied independently of the reference standard. One study was rated low risk overall (Dai 2025); four had some concerns (Kolanu 2020, Fernandes-Pereira 2024, Spångeus 2025, Mathew 2025); and three were rated high risk overall (Page 2023, Bendtsen 2024, Behanova 2026). The reference-standard and flow-and-timing domains accounted for most concerns: reference-standard limitations arose from partial or non-independent verification (high risk in Behanova 2026, where routine clinical reports served as the reference standard), and flow-and-timing concerns were present in seven of eight studies, largely reflecting retrospective validation designs and algorithm-failure exclusions.

D. Reviewer process

Reviewer process. Two reviewers (A.S. and M.O.) independently applied QUADAS-2 to all eight included studies; disagreements were resolved through discussion, with E.K. serving as senior adjudicator when consensus could not be reached. Formal inter-rater agreement was not calculated; all discrepancies were resolved by consensus prior to final risk-of-bias judgements. Applicability concerns relating to manufacturer co-authorship or funding in several studies are addressed in the Discussion.

Supplementary Figure 1

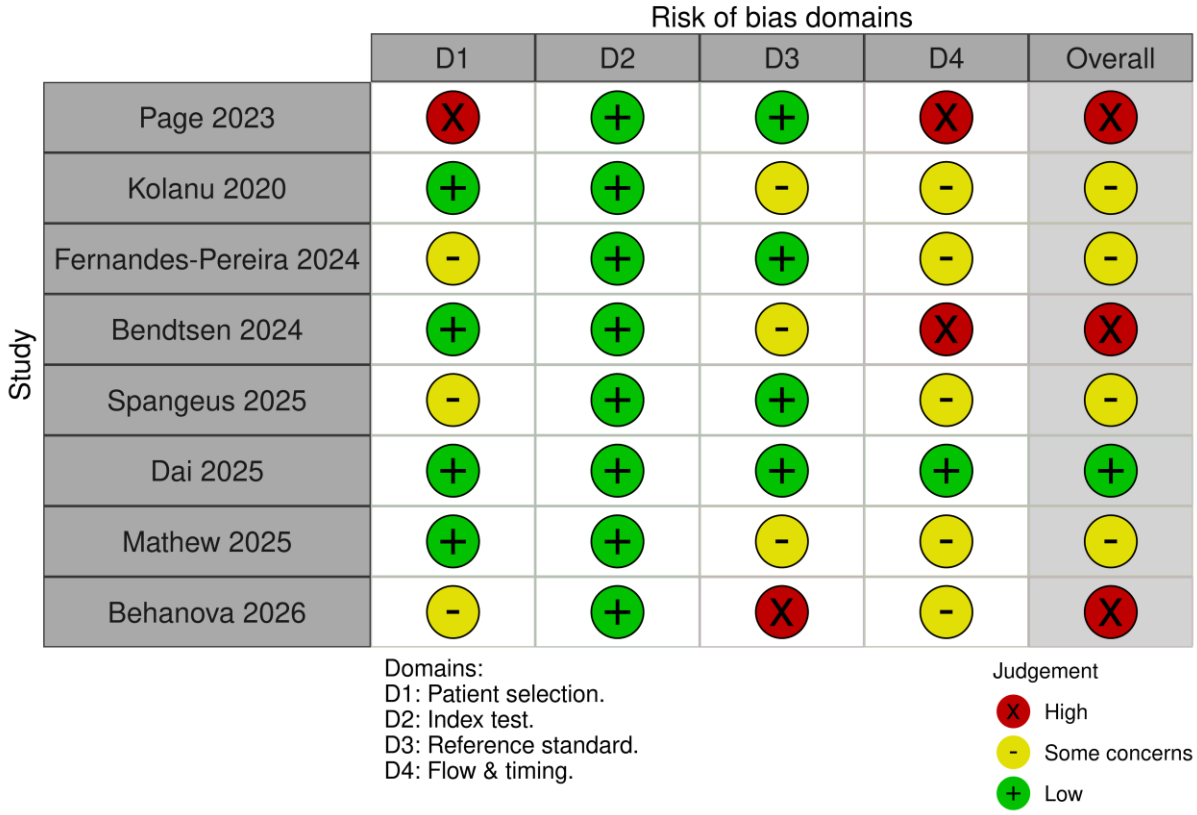

Figure S1. *Risk-of-bias assessment of the eight included studies using the QUADAS-2 framework, visualised with the robvis tool. Each study is rated across four domains (D1, patient selection; D2, index test; D3, reference standard; D4, flow and timing) and an overall judgement. Green (+) denotes low risk, yellow (−) some concerns, and red (×) high risk. The index-test domain was rated low risk for all eight studies; the reference-standard and flow-and-timing domains accounted for most concerns.*