

Context Determines Value of AI Doubt Alerts in Fracture Detection: A Multi-Site Retrospective Analysis of Clinical Indication and Anatomy

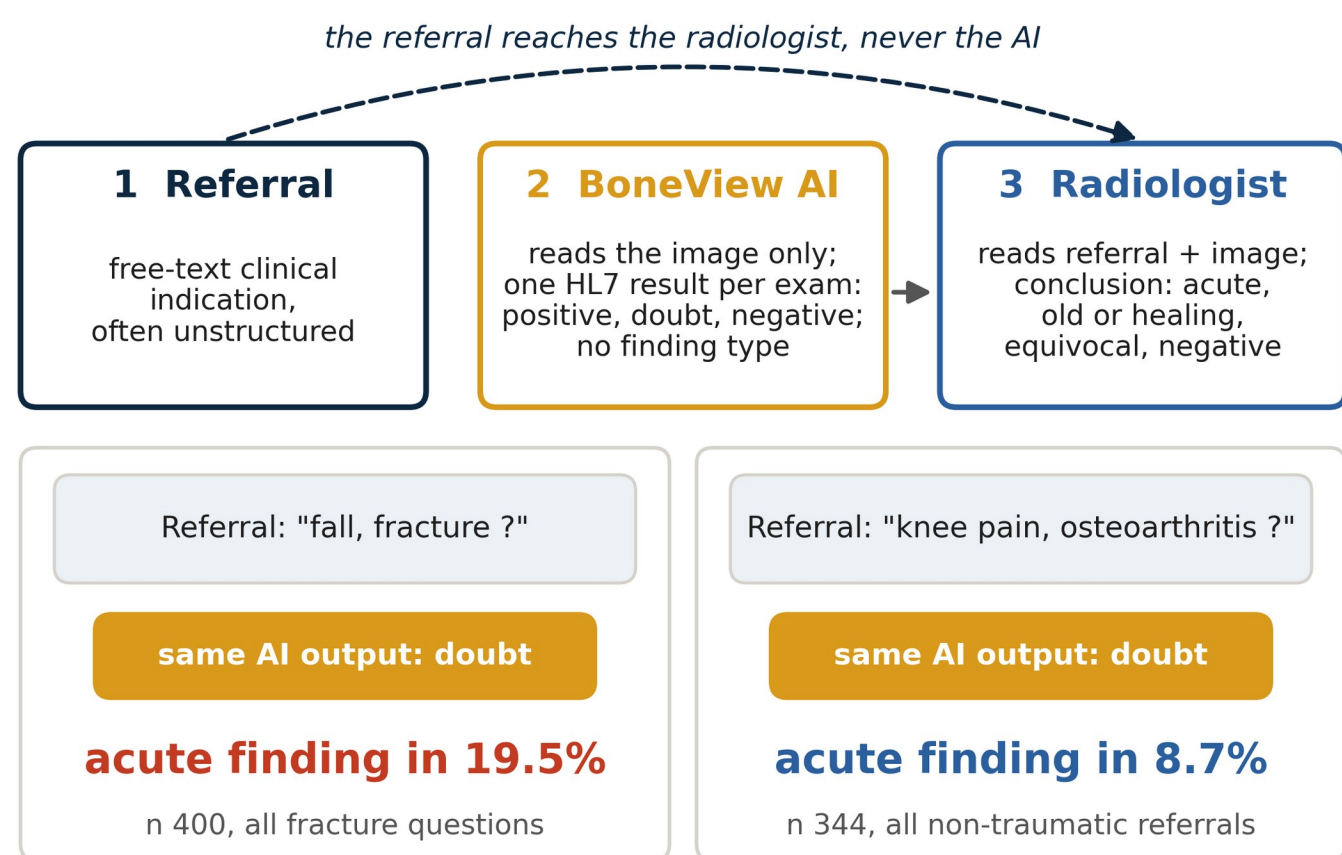
S. Morozov^{1,2}, B. Rizk¹, O. Novarina¹, N. Heracleous¹, C. Thouly¹, B. Dufour¹

¹ 3R Swiss Imaging Network, Sion, Switzerland · ² Medlogic, Brussels, Belgium



INTRODUCTION

The AI never sees the referral. Does the referral decide what a doubt output is worth?



BoneView (Gleamer) returns a positive, doubt or negative output on every radiograph, without clinical context. We asked whether the referral indication and the body region are associated with an acute finding in the radiologist's conclusion after a doubt output.

METHODS

879 doubt outputs, 22 outpatient centres, 2025.

- Doubt AI outputs from 2025 with research consent and a report, one per examination (n = 879). Analysis updated after the proceedings abstract; labels validated by two radiologists.
- Referral indication: fracture or trauma question (including follow-up of a known fracture); non-traumatic indication; no usable indication.
- Radiologist's conclusion: acute finding, old or healing fracture only, equivocal, or negative.
- Labels from a rule-based text classifier, validated against blind reading by two radiologists (100 and 50 reports): indication κ 0.62 (95% CI 0.50-0.74) and 0.76 (0.60-0.90); conclusion κ 0.72 (0.60-0.82) and 0.75 (0.56-0.90). Between radiologists: κ 0.65 (indication), 0.89 (conclusion).
- Outcome: acute finding in the conclusion (fracture, dislocation, bone lesion or joint effusion; not old or healing). Wilson 95% CI; logistic regression on fracture question and body region.

879

doubt outputs, 2025

22

outpatient centres

150

blind readings, 2 radiologists

LIMITATIONS

Reference is the issued report, not the images. AI output visible to the radiologist (incorporation bias); under-reporting acts the other way. Text-based labels (κ above). Positive and negative outputs are capped exports. Single outpatient network, 2025.

ALL THREE AI OUTPUTS

Acute finding in the report, % of reports in the cell			
positive	52.6% n 525	19.4% n 222	41.9% n 129
	Fracture question	Non-traumatic	No usable indication
doubt	19.5% n 400	8.7% n 344	5.9% n 135
negative	7.2% n 361	7.1% n 337	1.9% n 154

Fig 4. Acute finding by referral indication after each AI output: higher with a fracture question after positive and doubt outputs, no difference after negative outputs. Doubt outputs: consented 2025 reports (879). Positive and negative outputs: capped exports of 878 each (876 and 852, one per examination). Routine mix about two thirds negative, a quarter positive, 6% doubt (Jan-May 2025), so no sensitivity or NPV.

TL;DR

What. 479 of 879 doubt outputs (54.5%) followed a referral without a fracture question. The radiologist concluded an acute finding after 19.5% of doubt outputs with that question and 7.9% without (region-adjusted OR 2.97, 95% CI 1.88-4.67).

Why. The AI never sees the referral, so the same output carries a different probability for each referral; the tool answers a trauma question the referral often did not ask.

So what. Send the referral indication to the AI; no interruptive alert without a fracture question; ask vendors for indication-specific operating points.

54.5%

of doubt outputs: no fracture question in the referral (479/879)

19.5% vs 7.9%

acute finding, with vs without a fracture question

aOR 2.97

95% CI 1.88-4.67, adjusted for body region

TABLE 1. DOUBT OUTPUTS BY REFERRAL INDICATION (n = 879)

Referral indication	Doubt outputs n (%)	Acute finding	Old or healing only	Equivocal	Negative	Acute finding % (95% CI)
Fracture or trauma question, incl. follow-up	400 (45.5)	78	77	73	169	19.5 (15.9-23.7)
of which follow-up of a known fracture	66 (7.5)	4	51	6	2	6.1 (2.4-14.6)
Non-traumatic indication	344 (39.1)	30	27	22	263	8.7 (6.2-12.2)
No usable indication	135 (15.4)	8	10	13	102	5.9 (3.0-11.3)
No fracture question, combined	479 (54.5)	38	37	35	365	7.9 (5.8-10.7)
Total	879 (100)	116	114	108	534	13.2 (11.1-15.6)

Acute finding, % of all reports in the row. 7 reports without an evaluable conclusion are counted in the denominator and not shown in the columns.

ACUTE FINDING

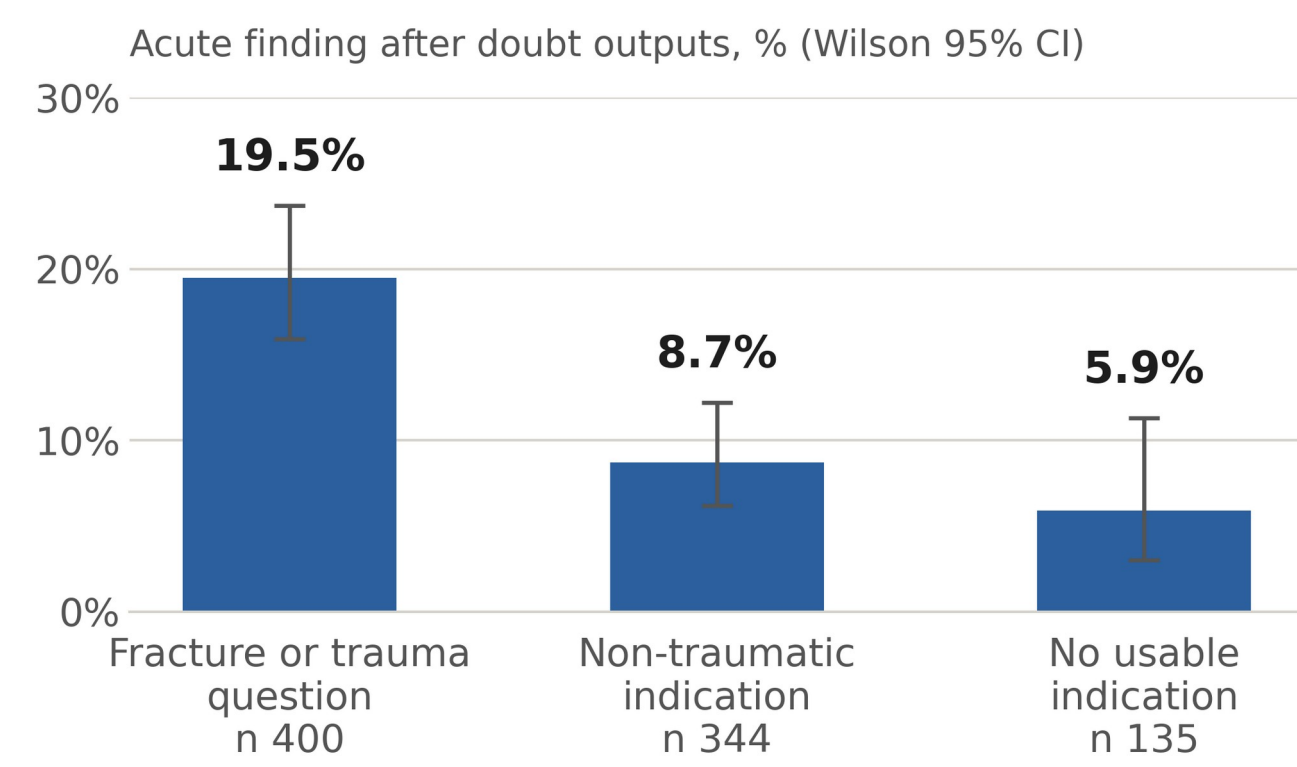


Fig 1. Acute finding after doubt outputs by referral indication, Wilson 95% CI. Similar with a non-traumatic or no usable indication.

EQUIVOCAL CONCLUSIONS

Radiologists equivocate most after a doubt output with a fracture question.

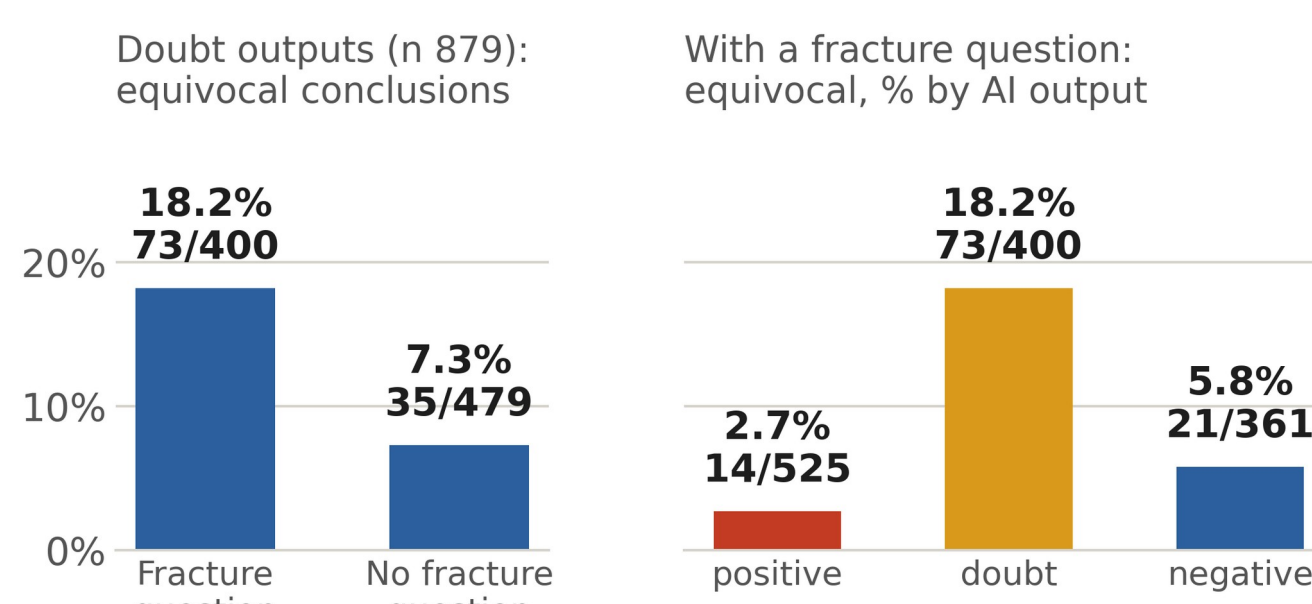


Fig 2. Left: equivocal conclusions after the 879 doubt outputs (108 in total: 73 with a fracture question, 35 without). Right: equivocal conclusions among reports with a fracture question, by AI output (positive n 525, doubt n 400, negative n 361).

BODY REGION

With a fracture question, acute findings range from 26% at the foot to 6% at the shoulder.

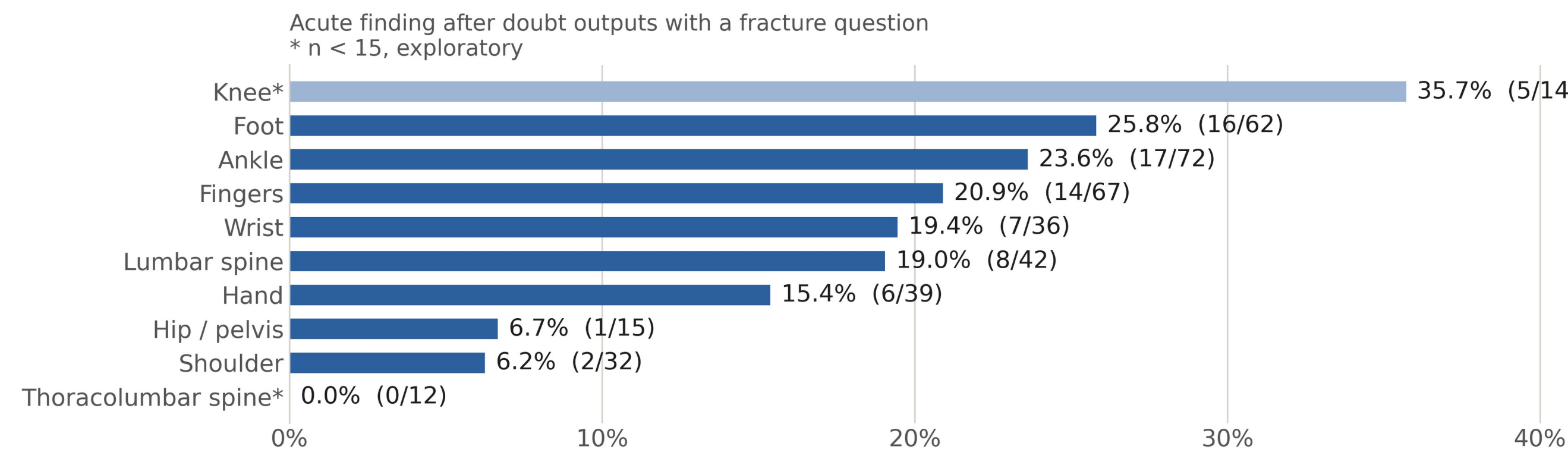


Fig 3. Acute finding after doubt outputs with a fracture question, by body region. Elbow (n 8) and ribs (n 1) not shown.

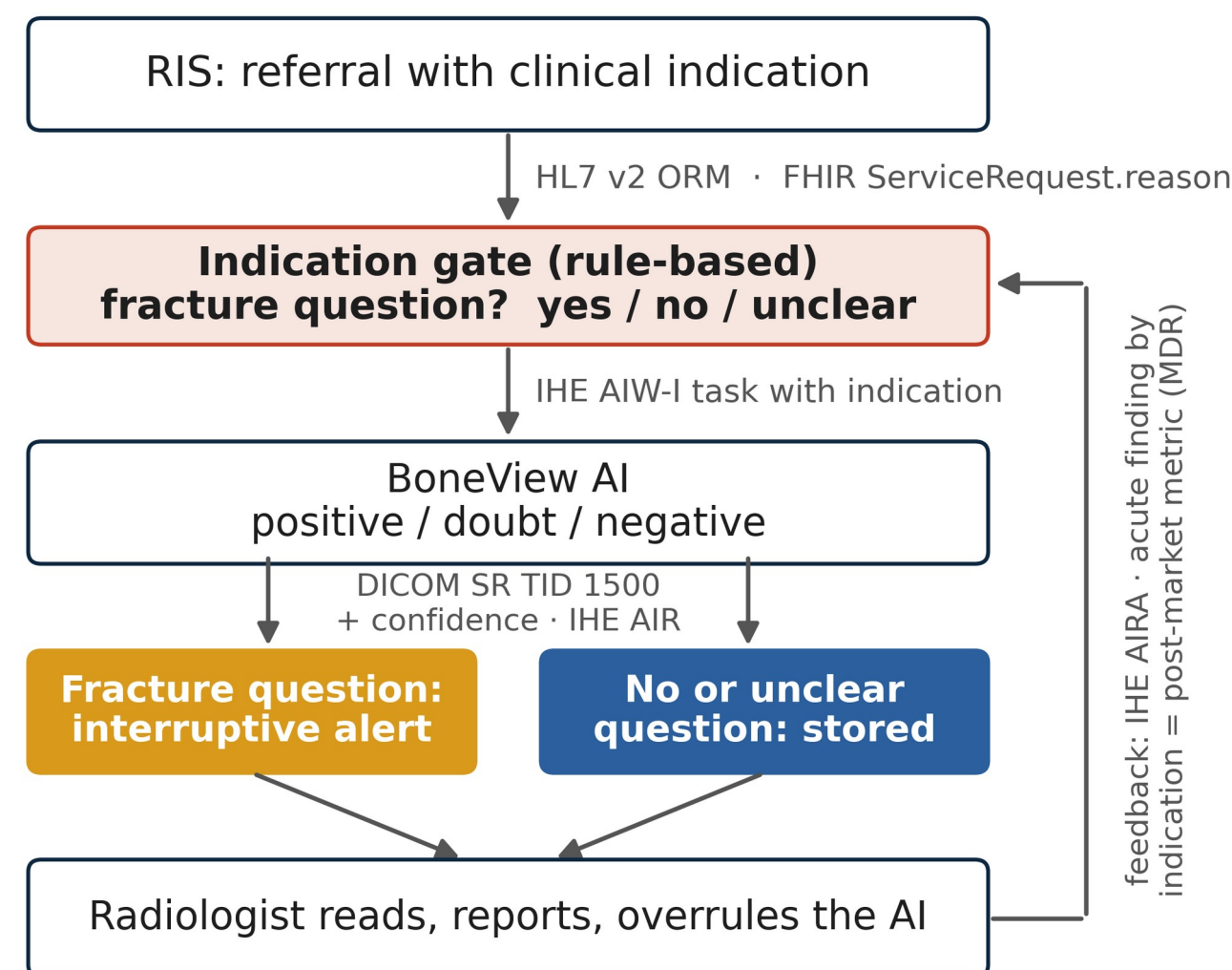
WHAT YOU CAN REUSE

- Referral indication rule: read the indication only; fracture question / no fracture question / no usable indication, never counted as a fracture question.
- Four-way conclusion: acute, old or healing only, equivocal, negative. Never fold equivocal into negative.
- Validate text labels by blind reading with κ , two radiologists; change a rule only for a classifier error, never for a radiologist's judgement.
- Report the value of AI outputs stratified by referral indication and body region. Ask vendors for it as a post-market metric.

THE INTEGRATION LAYER

Implications, not a result of this study.

Pass the referral to the AI. Suppress the interruption, keep the result.



Why not switch it off: without a fracture question the tool still reports findings, acute in 8.7% of non-traumatic referrals.

The standards exist. We know of no commercial PACS or RIS that offers the context-aware delivery layer.

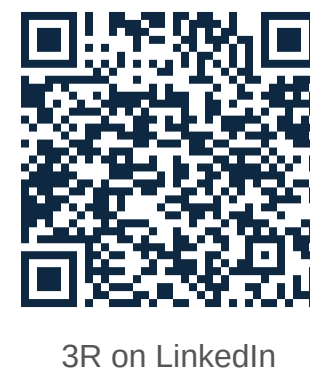
OPEN QUESTIONS

- Should non-trauma radiographs go through the AI at all? Yes, but not as an interruptive alert: an acute finding still followed 8.7% (30/344) of doubt outputs with a non-traumatic referral. Options: AI on demand (computed, shown when the radiologist asks), a silent worklist flag, and no AI result in the patient portal for these referrals.
- How many referrals carry no usable indication, and what does it cost? 15.4% of doubt outputs (135/879) followed such a referral; an acute finding was concluded in 5.9%. A mandatory structured indication field (FHIR ServiceRequest.reasonCode) is the cheapest fix.
- Equivocal images, shared uncertainty or automation bias? With a fracture question, radiologists equivocate after 18.2% of doubt outputs vs 2.7% of positive and 5.8% of negative. Only a design with the AI output hidden can separate these.

TAKE-HOME POINTS

- 1 A doubt output is a probability, not a finding: an acute finding followed 20% with a fracture question, 8% without.
- 2 Do not switch it off outside trauma (acute finding in 9% there): make it on demand or a silent flag; keep it out of the patient portal.
- 3 With a fracture question, radiologists equivocate after 18% of doubt, 3% of positive and 6% of negative outputs. Equivocal images, shared uncertainty or automation bias: still to be separated.
- 4 The near fix is integration: referral to the AI (HL7 ORM / FHIR ServiceRequest), context-aware alert policy, vendor operating points per indication. The standards exist; the layer does not.
- 5 The deeper problem is scope: the tool's four findings cover only part of what non-trauma referrals ask. AI should cover the clinical questions and pathology spectrum of real referrals.

References. 1. Guermazi A et al. Radiology 2022;302:627-636. doi 10.1148/radiol.210937. 2. Duron L et al. Radiology 2021;300:120-129, doi 10.1148/radiol.2021203886. 3. Husarek J et al. Sci Rep 2024;14:23053, doi 10.1038/s41598-024-73058-8. 4. IHE Radiology profiles AIW-I, AIR, AIRA (trial implementation), profiles.ihe.net/RAD.
Contact. sergey@medlogic.co · benoit.rizk@groupe3r.ch



3R on LinkedIn

DISCLOSURES (as published). B.R. holds equity in Gleamer, the manufacturer of the BoneView tool evaluated in this study. Data extraction and analysis were performed independently by S.M., N.H. and O.N., who have no financial relationship with Gleamer. S.M. is affiliated with 3R as a paid consultant via Medlogic. B.R. was blinded second reader for label validation.
Paper: CLINICCAI 2026 proceedings, paper 48.