FALSE-POSITIVE–CONSTRAINED EVALUATION OF ARTIFICIAL INTELLIGENCE LUNG NODULE DETECTION AT CLINICALLY RELEVANT OPERATING POINTS

Dalia Artal, MD · Patrick Alore, MD · Jason Sinner, MD

RadNet

deephealth

BACKGROUND

- False-positive findings remain a primary barrier to clinical adoption of AI lung nodule detection, increasing review burden and reducing radiologist trust.
- Prior studies often emphasize peak sensitivity, while fewer assess performance at operating points aligned with real-world clinical use.
- This study evaluated whether an AI lung nodule detection system maintains high sensitivity while constraining false-positive burden at clinically relevant thresholds.

METHODS

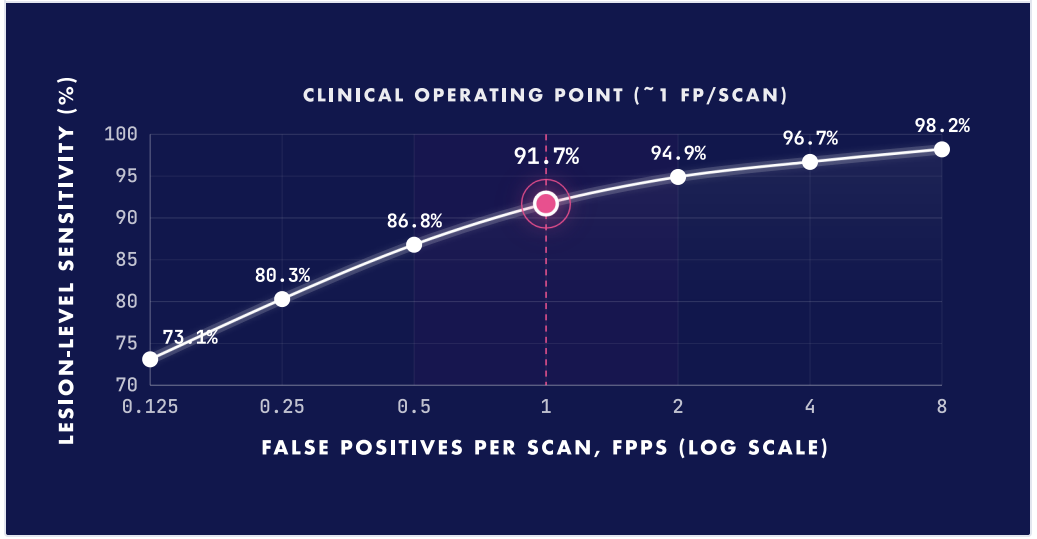
- The AI system was trained on a separate multicenter lung cancer screening cohort and evaluated on an independent multi-reader dataset reflecting interpretive variability.
- The test set included **1,009** CT examinations with **1,303** annotated nodules ≥5 mm.
- FROC analysis was used to assess lesion-level sensitivity across false-positive rates, with the primary operating point near **1 FPPS**, selected to reflect a clinically relevant false-positive constraint.
- Ninety-five percent confidence intervals were estimated using bootstrap resampling.

Key question: *Can high sensitivity be preserved without an impractical false-positive burden?*

KEY CLINICAL TAKEAWAY

AI MAINTAINS 91.7% SENSITIVITY AT APPROXIMATELY 1 FALSE POSITIVE PER SCAN.

FROC PERFORMANCE FOR NODULES 5 MM OR LARGER



RESULTS

91.7%
SENSITIVITY AT 1 FP/SCAN
95% CI 89.6–93.6
Clinically optimal operating point

94.9%
SENSITIVITY AT 2 FP/SCAN
95% CI 93.1–96.3
Higher sensitivity, moderate FP cost

86.8%
SENSITIVITY AT 0.5 FP/SCAN
Low FP burden, reduced sensitivity

98.4%
MAXIMUM SENSITIVITY (~9.6 FP/SCAN)
Near-max sensitivity, impractical FP load

DISCUSSION

- Overall FROC score was **0.889** (95% CI 0.873–0.904), reflecting robust performance across evaluated operating points.
- Sensitivity remained high across stricter false-positive constraints, supporting workflow-conscious evaluation rather than reliance on permissive peak sensitivity metrics.
- Diminishing returns beyond approximately 2 FPPS suggest limited incremental sensitivity benefit relative to added false-positive burden.
- Comparable trends observed for nodules ≥3 mm, with **90.8%** sensitivity near 1 FPPS.

CONCLUSION

- False positives are a well-documented harm of lung cancer screening and may increase downstream review burden.
- AI tools emphasizing peak sensitivity without constraining false positives may compound this burden.
- Evaluating AI at clinically relevant false-positive constraints yields **practice-relevant insight** beyond peak accuracy alone and supports workflow-conscious integration into lung cancer screening workflows.

SCAN FOR FULL STUDY AND EXTRA DATA