“TRY BEFORE YOU BUY": COMPARATIVE EVALUATION OF FDA-APPROVED AI ALGORITHMS FOR CORONARY CALCIUM SCORING ON LUNG CANCER SCREENING CTS

Syed Muhammad Awais Bukhari, MD, Sreyes Venkatesh, Annie Singh, Cody R. Johnson, MD, Robert C. Gilkeson, MD, Amit Gupta, MD

Session Title: R5B-SPCH. Chest Imaging Thursday Afternoon Poster Discussions II

Session Start: 12/4/2025 12:45:00 PM

Location: LEARNING CENTER

RSNA Poster Presentation Abstract

***Purpose:** While FDA-approved AI tools for coronary artery calcium (CAC) scoring have demonstrated utility on non-gated chest CTs, their performance on low-dose CTs (LDCTs) from lung cancer screening programs, where incidental CAC detection is critical, remains less clear. Given the cardiovascular risk overlap in this population and the lack of LDCT-specific training in these algorithms, our objective was to benchmark and compare the performance of three commercially available, FDA-cleared CAC scoring AI tools. This study underscores the need for local validation prior to clinical and commercial adoption.

***Methods and Materials:** We retrospectively identified 44 patients (mean age: 61±5.6 years; 57.8% females, 42.2% males) who underwent both ECG-gated calcium scoring CT (CSCT) and LDCT within a 16-day interval. CAC scores were calculated on LDCTs using three FDA-approved AI tools (CAC-AI_A, CAC-AI_B, and CAC-AI_C) via a vendor-neutral AI deployment platform (CARPL.AI). A semi-automated CAC score measured by a board-certified radiologist on CSCT (CAC_G) served as the reference standard. All CAC scores were stratified into Agatston-based risk categories: Very Low (0), Low (1-99), Moderate (100-399), and High (≥400). Performance metrics included ANOVA for mean comparisons, concordance correlation coefficient (CCC), Bland-Altman analysis, and weighted kappa (κ) statistics for risk classification agreement.

***Results:** Mean CAC scores were as follows: CAC_G = 314.33 ± 651.84; CAC-AI_A = 295.0 ± 533.2; CA_C-AI_B = 317.2 ± 595.9; CAC-AI_C = 313.4 ± 623.0 (p = 0.998). CCC values with the ground truth (CAC_G) were 0.96 (CAC-AI_A), 0.97 (CAC-AI_B), and 0.93 (CAC-AI_C). Mean differences from CACG were: CAC-AI_A: -18.6 ± 161.9 (95% CI: -67.9 to 30.6); CAC-AI_B: +2.9 ± 128.2 (95% CI: -36.1 to 41.8); CAC-AI_C: -1.0 ± 226.5 (95% CI: -69.8 to 67.9). Risk category agreement (κ) compared to CAC_G was highest for CAC-AI_B (κ = 0.76), followed by CAC-AI_C (κ = 0.69) and CAC-AI_A (κ = 0.47).

***Conclusions:** All three AI tools demonstrated reasonable agreement with gated CSCT for CAC scoring on LDCTs. However, in our institutional setting, vendor B's tool (CAC-AI_B) showed the most reliable performance for both quantitative scoring and risk stratification, followed by vendor C. Vendor A underperformed relative to the others.

***Clinical Relevance/Application:** As AI tools enter clinical practice, particularly in high-impact domains such as incidental CAC detection, local validation remains essential. Despite regulatory approval, differences in real-world performance across platforms can influence clinical utility. Vendor-neutral platforms now enable institutions to "try before you buy," facilitating informed, data-driven decisions that align with local imaging protocols and patient populations.