Improving the Efficacy of TI-RADS Through Artificial-Intelligence

Lev Barinov, PhD, Koios Medical

Introduction

Thyroid nodules occur in up to 68% of individuals. Most nodules (~95%) are benign and many malignant nodules would not result in symptoms or death. Still, over 600,000 fine needle aspirations (FNAs) are performed annually with a positive predictive value of only ~30%. ACR TI-RADS was developed to standardize diagnostic criteria, reduce biopsy rates, and limit the overdiagnosis of thyroid cancer. TI-RADS increases reader concordance while reducing unnecessary biopsies by 19.9-46.5%. In parallel, recent advances in artificial intelligence (AI) techniques have created opportunities to further improve the ACR TI-RADS system. In this study, we analyzed the use of an additional AI-generated nodule risk descriptor that independently assesses risk of malignancy. This predictive indicator is subsequently mapped to an integer point value ranging from -2 to +2 to be incorporated into the already-established TI-RADS point-based clinical management criteria to improve patient management decisions. The system, developed by Koios Medical, also prepopulates TI-RADS descriptors creating a putative point total that the reader can then consider and modify at their discretion.

Hypotheses

AI generated assessment of the 5 existing TI-RADS components and inclusion of an additional independent AI based malignancy risk assessment and point modifier will positively impact reader performance.

Methods

A multi-reader, multi-case study involving 15-readers (11-radiologists, 4-endocrinologists) and 650 FNA-proven nodules (130 malignant) was conducted. Readers evaluated each nodule twice across two sessions separated by a 4-week amnesia period. In each session, nodules were randomly presented and evaluated in one of two conditions:

1) Manual scoring of a TI-RADS report form (TI-RADS Only)
2) AI -prepopulated TI-RADS report form, augmented with an AI-based risk descriptor and point modifier (TI-RADS+AI).

Assuming a recommendation for FNA was a positive result, diagnostic performance in the two reading conditions were assessed via parametric ROCAUC and operating point analyses. Inter-reader variability was assessed via Pearson’s R correlation. Interpretation time was analyzed as a relative change between reading conditions.

Results

Average AUC improvement for TI-RADS+AI versus TI-RADS was 0.083 (95% CI, 0.066-0.099). TI-RADS+AI produced a significant increase in sensitivity and specificity of 8.4% (95% CI, 5.4%-11.3%) and 14% (95% CI,12.5%-15.5%), respectively. Inter-reader variability was 0.622 and 0.876 for TI-RADS and TI-RADS+AI, respectively. Interpretation time decreased by 23.6% (p<0.001) for TI-RADS+AI.

Conclusion

Automated AI-based prepopulation of the 5 TI-RADS descriptors combined with use of an additional AI-based risk descriptor and point modifier significantly improves reader diagnostic accuracy while simultaneously decreasing interpretation time and inter-reader variability.

Figure(s)

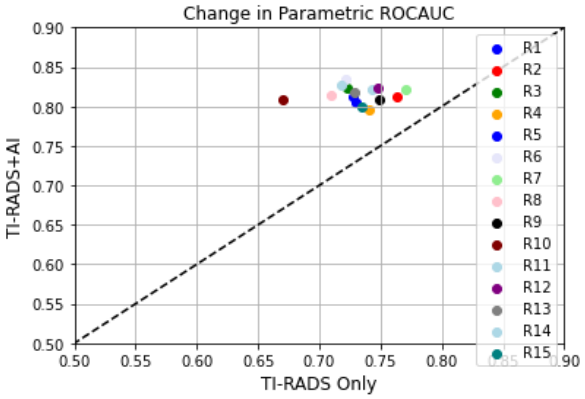

Figure 1. *Per reader parametric AUC comparing TI-RADS Only to TI-RADS+AI for all readers on all of the data. The dashed line represents equivocal results with all points above this line demonstrating an improvement for the TI-RADS+AI reading condition.*

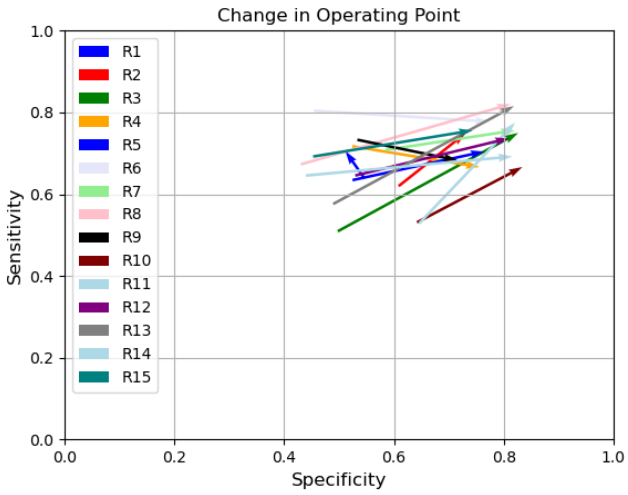

Figure 2. *Change in Sensitivity and Specificity of recommendations for all data for all readers. The base of the arrow represents the initial operating point for TI-RADS Only, while the arrow head represents the sensitivity and specificity of TI-RADS+AI*

Table 1. *Per reader and average performance with 95% confidence intervals for the parametric analysis.*

Reader	Difference in AUC	Percent Change in AUC
R1	0.076 [0.011, 0.141]	10.442 [1.541, 19.344]
R2	0.047 [-0.021, 0.114]	6.072 [-2.748, 14.893]
R3	0.100 [0.033, 0.167]	13.765 [4.446, 23.085]
R4	0.054 [-0.016, 0.125]	7.326 [-2.212, 16.864]
R5	0.087 [0.026, 0.147]	11.871 [3.541, 20.201]
R6	0.114 [0.052, 0.176]	15.735 [7.053, 24.416]
R7	0.051 [-0.014, 0.115]	6.593 [-1.765, 14.951]
R8	0.105 [0.039, 0.171]	14.742 [5.420, 24.063]
R9	0.059 [-0.004, 0.122]	7.913 [-0.514, 16.340]
R10	0.139 [0.072, 0.205]	20.685 [10.641, 30.729]
R11	0.079 [0.016, 0.142]	10.593 [2.068, 19.118]
R12	0.073 [0.008, 0.138]	9.781 [1.057, 18.505]
R13	0.088 [0.025, 0.152]	12.079 [3.327, 20.831]
R14	0.109 [0.049, 0.168]	15.128 [6.753, 23.504]
R15	0.065 [-0.003, 0.133]	8.892 [-0.387, 18.170]
Average	0.083 [0.066, 0.099]	11.386 [9.119, 13.652]

Keywords

Artificial Intelligence; Clinical Workflow and Productivity