



Long-term Real-world Performance of a Pulmonary Embolism Detection and Re-prioritization Algorithm

Yin Xi, PhD, UT Southwestern Medical Center; Michael Long, MS; Kiran Batra, MD; Xinhui Duan, PhD; Ron Peshock, MD

Introduction/Background

A pulmonary embolism triage algorithm (AI-PE) for dedicated CTPA exams demonstrates 80% sensitivity and 99% specificity and reduces exam wait times. However, its long-term stability and performance in different patient populations remains unclear. The goal of this study was to assess temporal trends and subgroup variations in the diagnostic performance of an AI-PE algorithm using extended data from two institutions.

Methods/Intervention

All adult CTPA exams between 2022-04-01 and 2025-05-15 from a safety-net hospital (SNH) and a university hospital (UH) were used in this analysis. CTPA exams were performed using computed tomography (CT) scanners with 64-256 slices. Exams performed on a Photon counting CT were excluded. Structured reports and vendor provided NLP served as reference standard. Patient characteristics (age, gender, race/ethnicity and BMI), exam characteristics (site, class, priority, scanner) were extracted. Logistic regression with AIC-based backward elimination was used to compare sensitivities and specificities between subgroups. The full model included institution, class, exam priority, age, sex, BMI, race/ethnicity, year, scanner, and all two-way interactions among (1) institution, class, and priority; (2) institution, age, sex, BMI, and race/ethnicity; along with an interaction between institution and year. Wait-time, read-time, and timeliness of AI were also analyzed.

Results/Outcome

16,581 exams from SNH and 13,767 from UH were retrieved. PE was detected in 3,565 reports. Sensitivity was associated with age, sex, race/ethnicity, and the interaction between institution and class (Figure 1). Specificity was high overall. AI results were received before dictation began in 87% and before dictation completion 95% of the time. Median wait-time was shorter for PE positive exams in all priorities. Median read time was shortest for TN and longest for FN.

Conclusion

Outpatient exams at SNH, younger, lower BMI in males, and non-Hispanic Black were associated with lower sensitivities. Across the board lower sensitivity was seen compared to literature indicating the AI algorithm may be under calling PE.

Statement of Impact

Continuous monitoring ensures efficacy and identifies areas for further improvement.

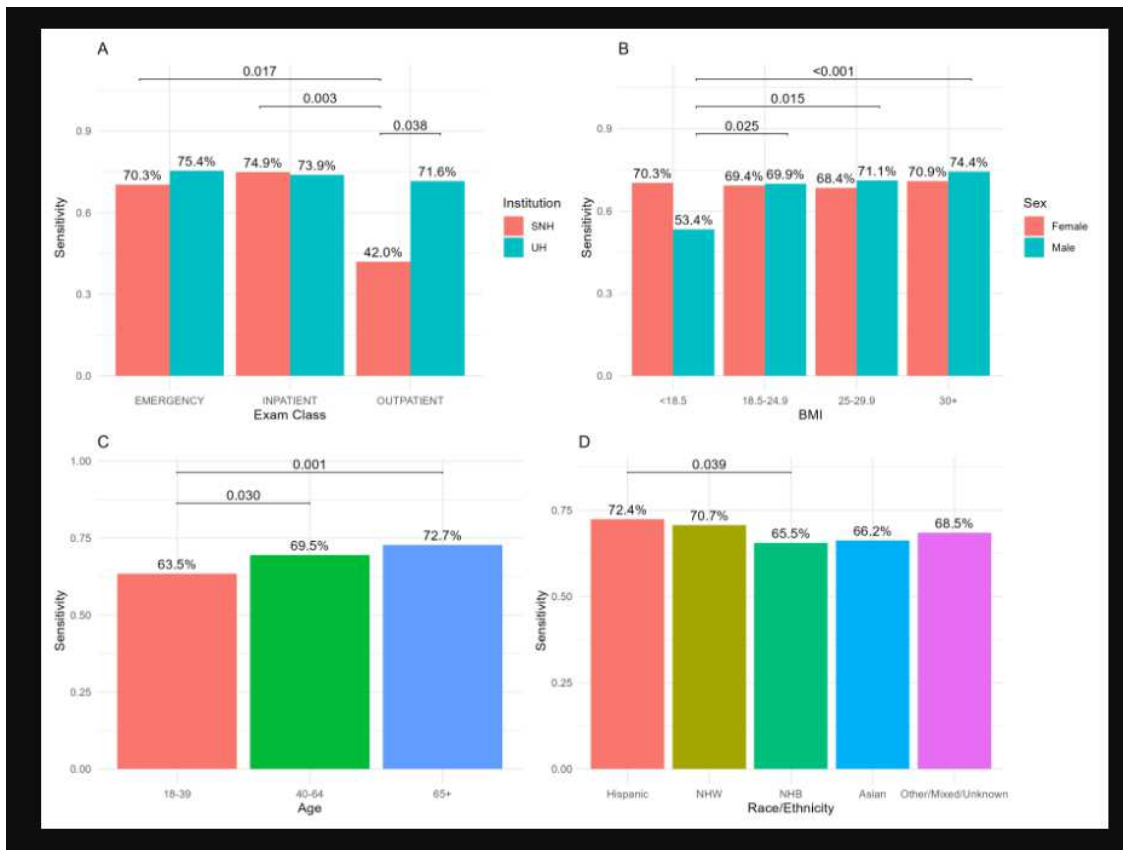


Figure 1. Factors associated with changes in sensitivity. After hierarchical backward elimination, the final model for sensitivity included institution, class, age, sex, race/ethnicity, and interactions between institution and class, and between sex and BMI. A. Sensitivity of class by institution. B. Sensitivity by BMI by Sex. C. Sensitivity by age. D. Sensitivity by race/ethnicity. BMI: body mass index; NHB: non-Hispanic Black; NHW: non-Hispanic White.

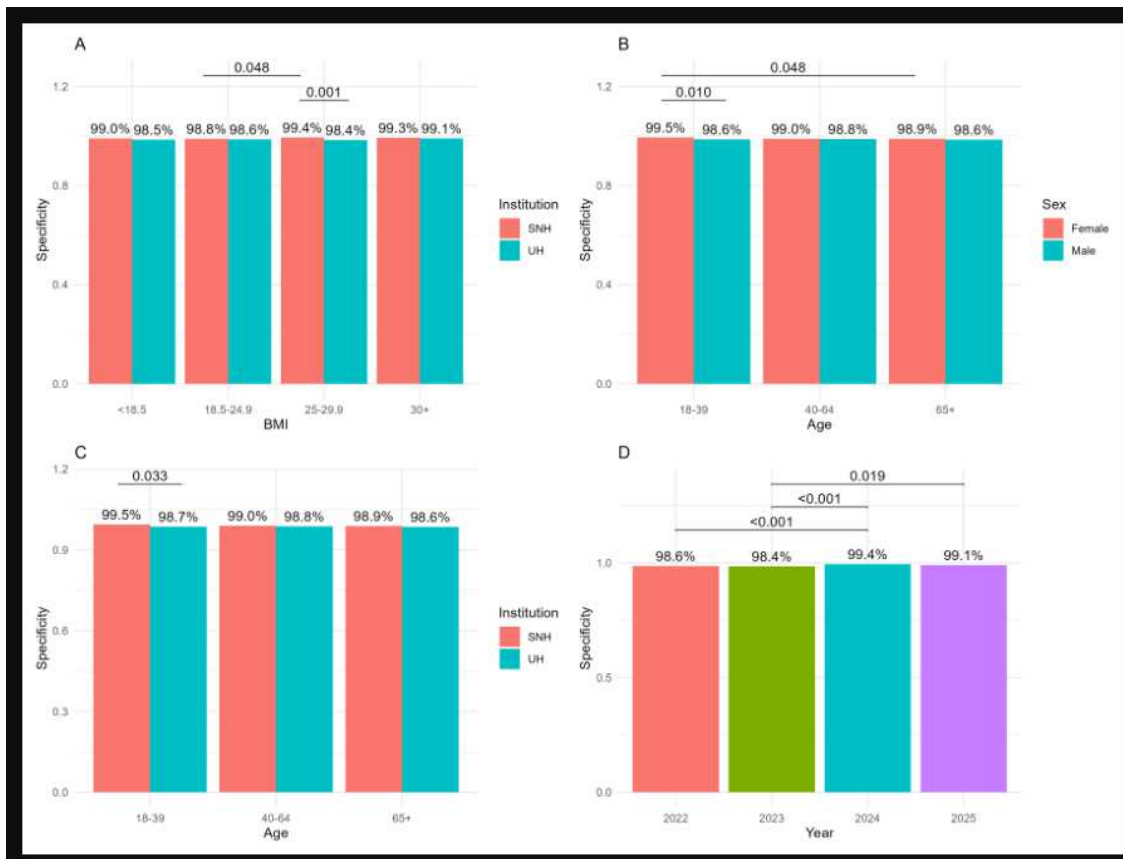


Figure 2. Factors associated with changes in specificity. After hierarchical backward elimination, the final model for specificity included institution, age, sex, BMI, year, and interactions between institution and BMI group, between age group and sex and between age and sex. A. Specificity by BMI by institution. B. Specificity by age by sex. C. Specificity by age by institution. D. Specificity by year. BMI: body mass index; NHB: non-Hispanic Black; NHW: non-Hispanic White.

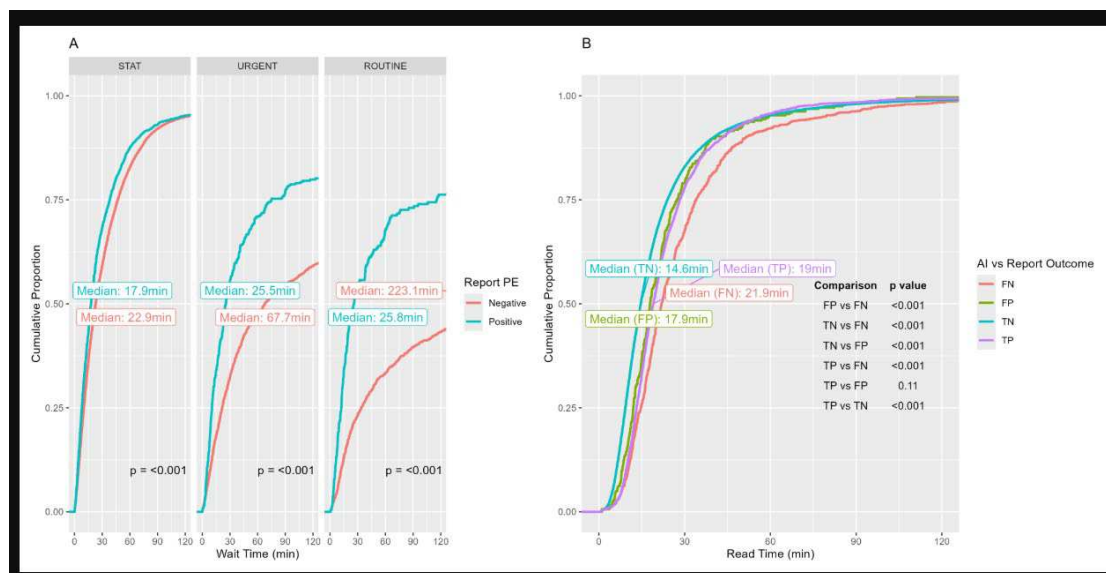


Figure 3. Cumulative proportion of wait-time and read-time. Wait-time was defined as the time between exam end and dictation start. Read time was defined as the time between dictation start and first report submitted. A. Wait-time stratified by exam priority and report PE positive vs negative. B. Read time stratified by comparison between AI and report. FN: false negative (i.e. AI negative and report positive); FP: false positive; TN: true negative; TP: true positive.

Keywords

Artificial intelligence; pulmonary embolism; real-world experience; diagnostic performance