Incorporation of clinical factors to improve the diagnostic accuracy of artificial intelligence-based chest X-ray analysis for detecting pulmonary tuberculosis

Coralie Geric, MScPH

PhD student, Division of Experimental Medicine, McGill University

McGill University

Centre international de TB McGill — McGill International TB Centre

A PAHO/WHO Collaborating Centre for Tuberculosis Research

Centre universitaire de santé McGill — McGill University Health Centre

Institut de recherche — Research Institute

recru

Unité d'épidémiologie respiratoire et de recherche clinique
Respiratory Epidemiology and Clinical Research Unit

Background

- For over a century, chest radiography (CXR) has been an essential tool in TB diagnosis, clinical care & follow-up
  - Interpretation is subjective and humans dichotomize interpretations to classify CXR as normal vs abnormal, with abnormalities consistent with TB or not
- Advances in artificial intelligence-based CXR analysis (or CAD) allows for objective, quantitative measurements of the degree of abnormality
- Commercial CAD output a continuous score on a 100-point scale
  - Higher scores = increased likelihood of TB
- However, humans have been applying cut-off values to interpret CXR as usual (normal vs abnormal)

Background – threshold scores

- Cut-off values are referred to as threshold scores
- Example:
  - CAD score ≥ 30 → CXR consistent with TB
  - CAD score < 30 → CXR **not** consistent with TB

Disadvantages:

- Sensitivity & specificity of a given threshold score are affected by clinical variables
  - Age, sex, HIV, and prior TB
- It is recommended that users perform accuracy studies to identify thresholds in their population

Objectives

- Can we make better use of CAD abnormality scores?

- We sought to:
  1. create a clinical model that uses continuous CAD scores and incorporates clinical data to estimate the predicted probability of pulmonary TB
  2. compare the clinical model vs using the CAD score alone for the diagnosis of culture or PCR confirmed pulmonary TB

Methods

- Developed a clinical model using logistic regression
  - Outcome = TB
  - Predictors = clinical variables (age, sex, HIV, prior TB)
- Used individual patient data from three studies in Pakistan, Zambia, and Tanzania
- CXR analyzed using two commercially available CAD
  - CAD4TB v6 (Delft Imaging, Netherlands) & qXR v2 (qure.ai, India)

Methods

- First, we asked does adding clinical data improve discrimination, compared to using CAD alone
  - Compared ROC curves of CAD alone vs CAD + clinical variables
  - Internally validated the clinical model using bootstrap validation

- Next, we compared the accuracy of differentiating between participants with & without TB when using CAD alone vs the clinical model
  - Used predication probabilities that achieved pre-specified sensitivities, and calculated the corresponding specificity, positive predictive value & negative predictive value

- Each of the above was performed separately for each software

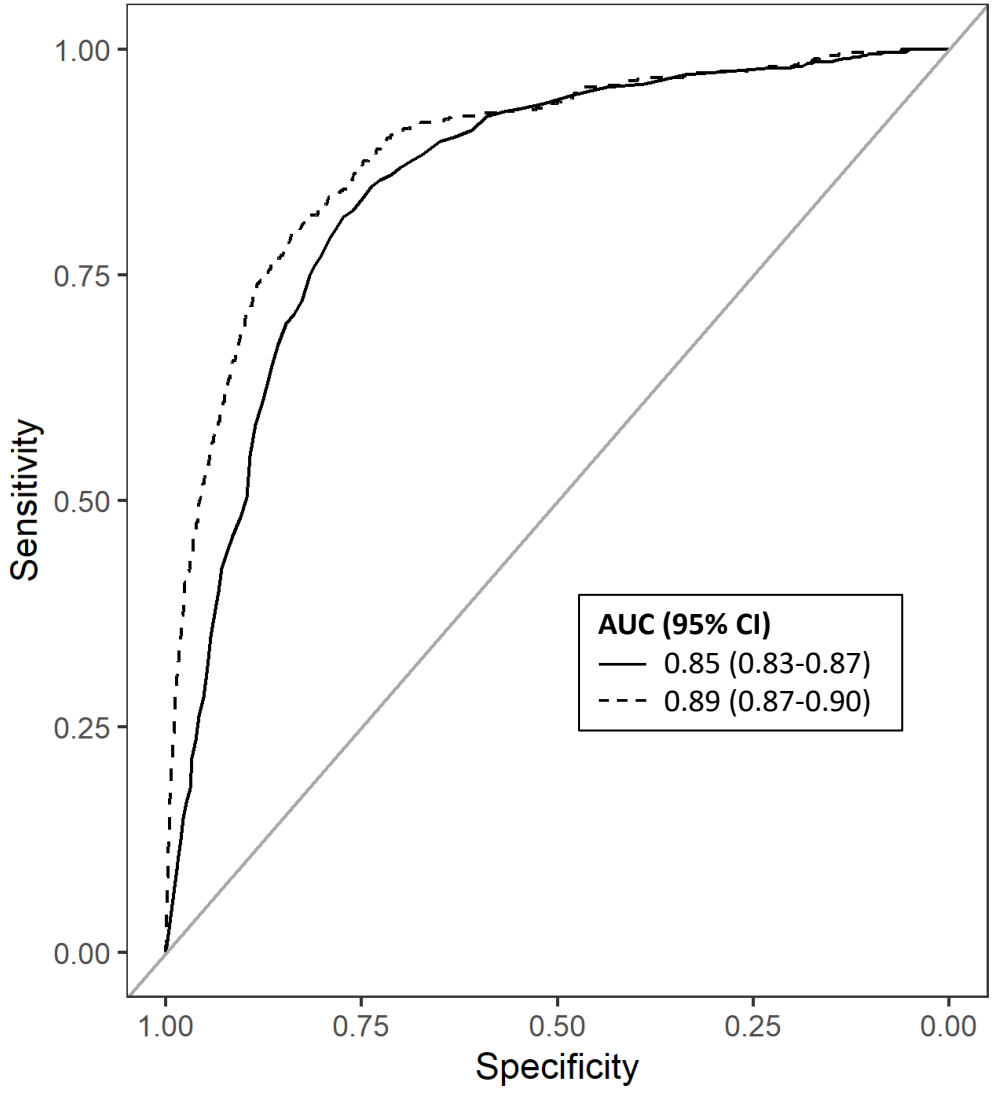
Results

Table 1. Characteristics of 3308 included participants

Characteristic	Overall (N=3308)	Pakistan (N=2283)	Tanzania (N=708)	Zambia (N=317)
Age, median (IQR)	35 (25, 48)	33 (23, 49)	38 (31, 50)	35 (28, 43)
Female, N(%)	1566 (47.3)	1091 (47.8)	352 (49.7)	123 (38.8)
HIV-positive	492 (14.9)	3 (0.1)	308 (43.5)	181 (57.1)
Previous TB, N(%)	704 (21.3)	517 (22.6)	111 (15.7)	76 (24.0)
NAAT or culture positive for *MTB*	566 (17.1)	292 (12.8)	187 (26.4)	87 (27.4)
Smear-positive	420 (12.7)	221 (9.7)	141 (19.9)	58 (18.3)

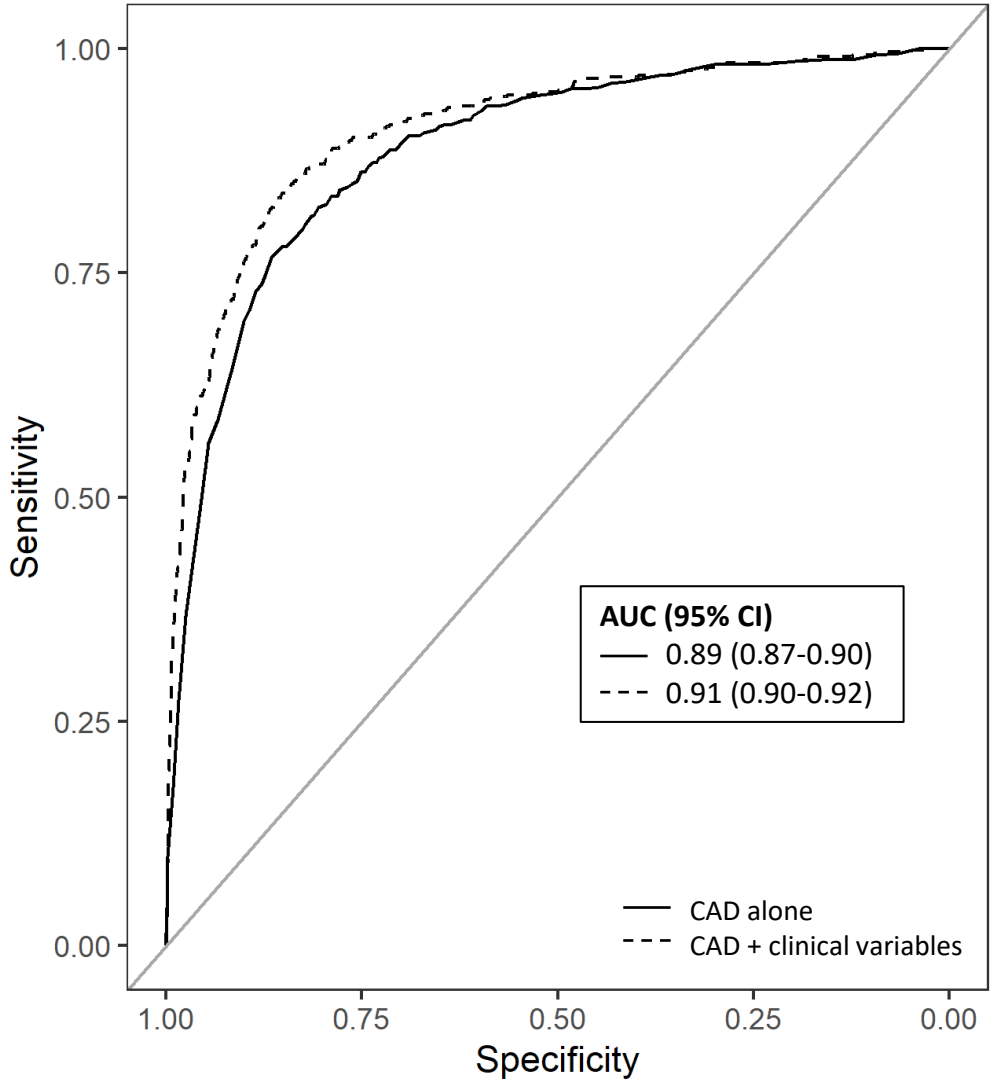
→ Those with TB were less likely to have a history of TB (13.6% vs 22.9%, p<0.001) and more likely HIV-positive (24.4% vs 12.9%, p<0.001)

A. CAD4TB



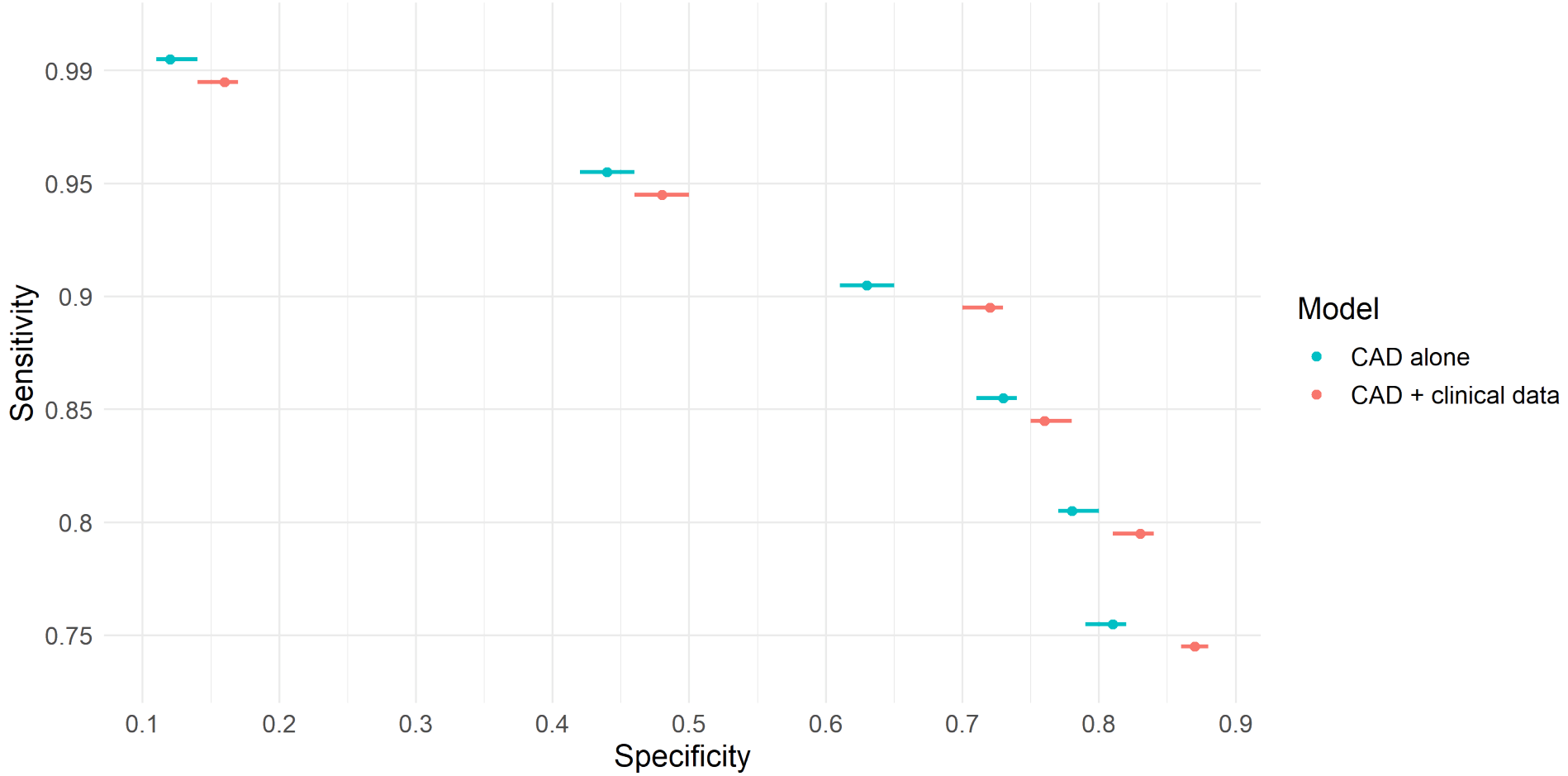
* DeLong's p-value < 0.001

B. qXR

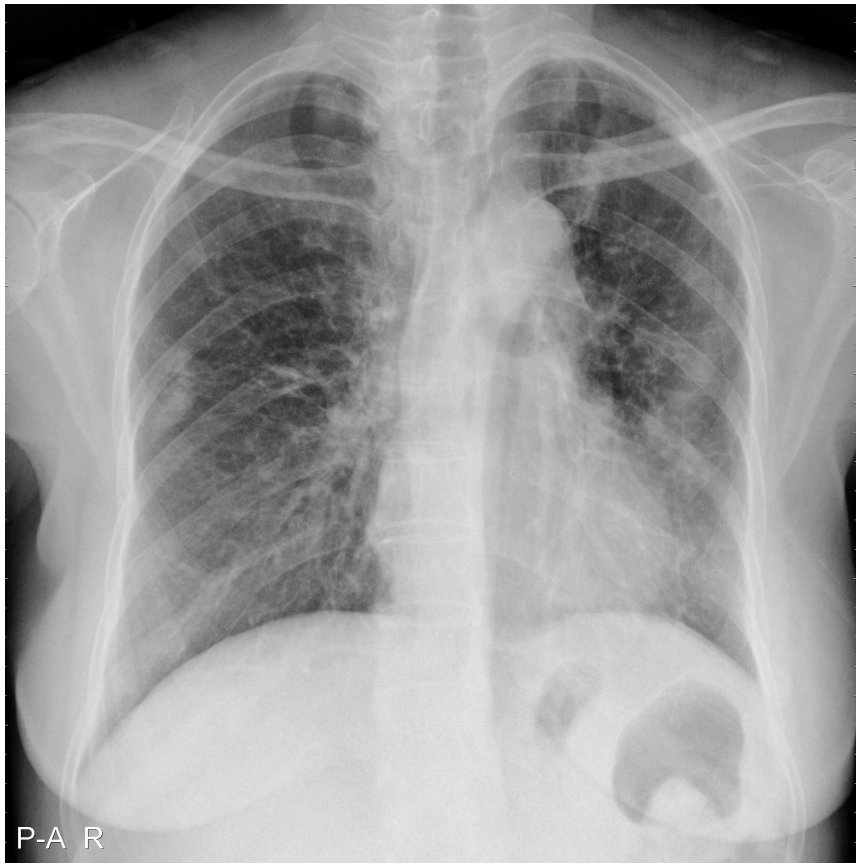


* DeLong's p-value < 0.001

Specificity of CAD alone vs CAD + clinical data



Example



- To achieve a sensitivity of 0.90, the cut-off values were:
  - CAD alone → CAD score ≥ 55
  - CAD + clinical data → predicted probability ≥ 9.5%
- Participant was 55 years old, female, HIV-negative, and had a history of TB and a CAD score of 63
- CAD alone:
  - CAD score = 63 ≥ 55 → consistent with TB
- CAD + clinical data:
  - predicted probability = 1.6% < 9.5% → **not** consistent with TB
- TB ruled out by two negative cultures & negative NAAT

Strengths & limitations

- Strengths:
  - Used individual patient data from multiple sites and countries, enhancing generalizability
  - High quality of studies & use of a microbiological reference standard likely reduced bias
  - Completed independently of CAD developers

- Limitations:
  - Internal validation only, limiting generalizability
  - Did not account for potential random effects from different sites

Conclusions & future directions

- Estimating the probability of TB using a model with continuous CAD scores **and clinical data** was more accurate at classifying individuals with TB symptoms than using the CAD score alone
- Having increased specificity, the clinical model could reduce the number of TB tests performed unnecessarily, without compromising the detection of people with TB

- Future directions:
  - External validation
  - Develop a point-based risk score system
  - Enhance model by incorporating additional clinical data

Acknowledgments

Co-investigators

- Gamuchirai Tavaziva
- Dr. Marianne Breuninger
- Dr. Keertan Dheda
- Dr. Ali Esmail
- Dr. Monde Muyoyeta
- Dr. Klaus Reither
- Dr. Aamir J Khan
- Dr. Andrea Benedetti
- Dr. Faiz Ahmad Khan

OBSERVATOIRE INTERNATIONAL SUR LES IMPACTS SOCIÉTAUX DE L'IA ET DU NUMÉRIQUE

Québec

Fonds de recherche – Nature et technologies
Fonds de recherche – Santé
Fonds de recherche – Société et culture