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
Background

• For over a century, chest radiography (CXR) has been an essential tool in TB diagnosis, clinical care & follow-up
  o 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
  o 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:
  o CAD score ≥ 30 → CXR consistent with TB
  o CAD score < 30 → CXR not consistent with TB

Disadvantages:

• Sensitivity & specificity of a given threshold score are affected by clinical variables
  o 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
  o Outcome = TB
  o 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
  o 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
  o Compared ROC curves of CAD alone vs CAD + clinical variables
  o 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

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

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)
AUC (95% CI)

- CAD alone: 0.85 (0.83-0.87)
- CAD + clinical variables: 0.89 (0.87-0.90)

AUC (95% CI)

- qXR: 0.91 (0.90-0.92)

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

• Limitations:
  o Internal validation only, limiting generalizability
  o 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:
  o External validation
  o Develop a point-based risk score system
  o 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