Introduction to Threshold Selection

MODULE 4

Stop TB Partnership

Course Outline

- Introduction to threshold scores
- Threshold score selection for local context
- Threshold score in this programme
- Summary

INTRODUCTION

This module introduces the key concepts of threshold score selection when using CAD and the topic of threshold selection. It concludes by introducing the threshold score for the project.

Learning Objectives

By the end of this module, participants should be able to:

- Understand what a threshold score is and how to set it.
- Know the effect of changing threshold on key screening targets.
- Describe why a threshold score needs to be chosen based on the local context.
- Know some of the strategies for threshold selection and how the project threshold score was selected.

INTRODUCTION TO THRESHOLD SCORE SELECTION

What is a “threshold score”?

It is a numerical value between 0 and 100 (or 0 and 1).

It translates the continuous output of CAD (abnormality score) into a binary output (a classification).

The first classification: Any chest X-ray with a score **above** the threshold value is automatically classified as “**TB**” (or similar) by CAD.

The second classification: All X-rays with a score **lower** than the threshold value are automatically assigned **“No TB”** (or similar) by CAD.

All images classified as “TB” by CAD should receive further confirmatory diagnostic testing.

Where CAD classification alone informs the triage decision, **the threshold score will determine key outcomes for an intervention**, such as the number of confirmatory diagnostic tests needed.

Basic Concepts in Threshold Selection

When using CAD classification alone to determine triage decisions, a threshold score can be chosen **based on programmatic goals.**

Some important factors to consider when identifying programmatic goals include:

Impact of Threshold Selection

There is a clear trade-off between key considerations for programs, so a threshold score needs to be adjusted in an informed way.

In general, a **low** threshold score results in:

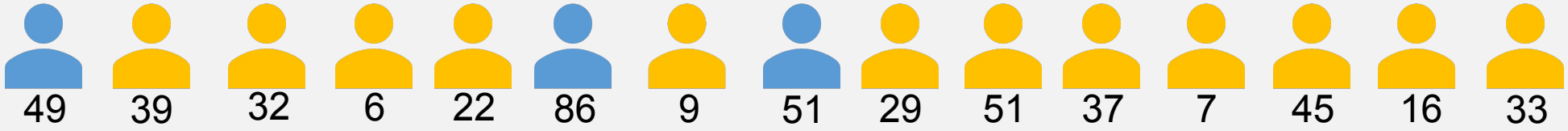
- **High** sensitivity but **low** specificity
  - More X-rays will have scores above the threshold, but a smaller proportion of these will have TB based on a diagnostic test.
- Needing to test **more** people to find a positive case, and therefore needing more diagnostic tests
- Increasing likelihood of **over-diagnosis** of TB

In general, a **high** threshold score results in:

- **Low** sensitivity but **high** specificity
  - More X-rays will be below the threshold, but a larger proportion of those above the threshold will have TB based on a diagnostic test.
- Needing to test **fewer** people to find a positive case, and therefore needing fewer diagnostic tests
- Increasing likelihood of **under-diagnosis/ missed cases** of TB

Threshold Score Trade-offs in Action

CAD score of population



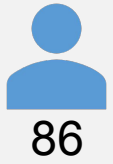
Person without TB

Person with TB

Population with "Possibility of TB" according to CAD

Situation A:
Saving on diagnostic tests

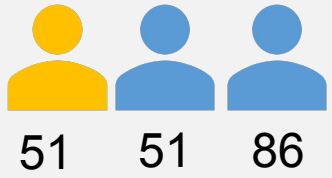
Threshold score is 75.



Sensitivity = 33%
Specificity = 100%
Number of confirmatory tests needed = 1
Number of missed/ undiagnosed cases = 2

Situation B:
Optimizing sensitivity with resource constraint

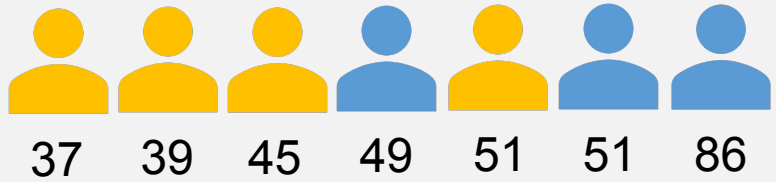
Threshold score is 50.



Sensitivity = 67%
Specificity = 92%
Number of confirmatory tests needed = 3
Number of missed/ undiagnosed cases = 1

Situation C:
No limit on testing resources

Threshold score is 35.



Sensitivity = 100%
Specificity = 67%
Number of confirmatory tests needed = 7
Number of missed/ undiagnosed cases = 0

Stop TB Partnership

Factors that Influence CAD Performance

- Underlying TB prevalence
- Presentation of TB in individuals with
  - Prior TB history
  - Co-morbidities (HIV, diabetes)
- Prevalence and proportion of other lung diseases
  - Silicosis, COVID-19
- Prevalence of risk factors for TB in specific populations

Factors that Influence CAD Performance

CAD's performance is shown to vary in different demographics and use populations.

The performance of CAD in a given population is therefore **impossible to predict precisely,** because it will depend on a combination of factors.

Individual variations in CAD performance may also occur.

The best way to choose a threshold score that will lead to a desired programmatic outcome is to collect local operational data.

How to Choose a Threshold Score

Selecting an appropriate threshold score is often described as challenging.

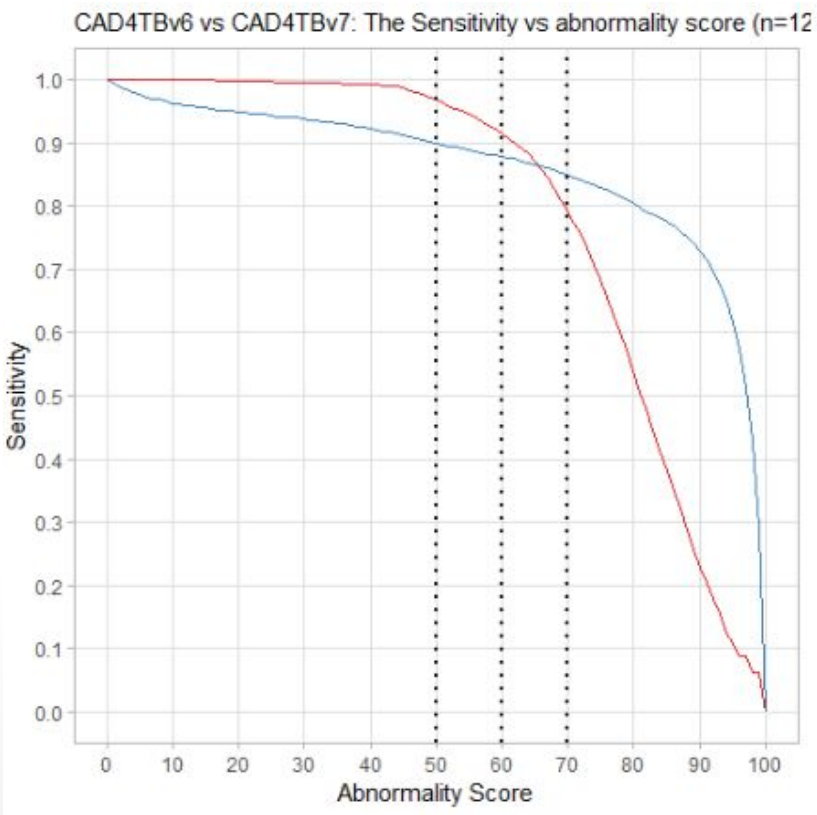
It is not possible to select one threshold score that applies between all CAD products, different software versions of the same CAD product, and different use cases and achieves the same results.

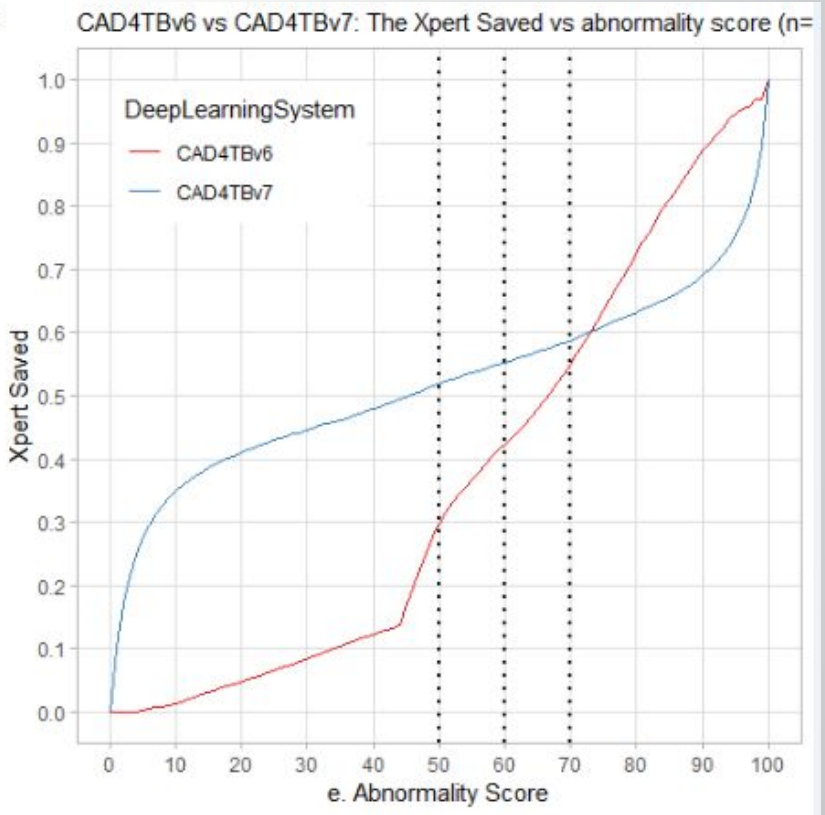
- Every CAD product is developed differently—an X-ray assigned 30 (or 0.3) by one CAD is not equally likely to have TB as an X-ray assigned 30 from another.
- Every CAD product performs differently in different sub-populations (for example older ages, HIV+), depending on the data used to develop it.
- Different versions of the same product may even be developed differently and perform differently in different sub-populations.

Performance Change between Versions

Preliminary results from a study comparing version 6 and version 7 of CAD4TB shows that **version 7 significantly outperformed version 6** when compared to the Xpert reference standard.

	Abnormality score	Sensitivity	Xpert saved
V6	50	0.97	0.3
	60	0.92	0.43
	70	0.8	0.55
V7	50	0.9	0.52
	60	0.88	0.55
	70	0.85	0.59





How to Choose a Threshold Score

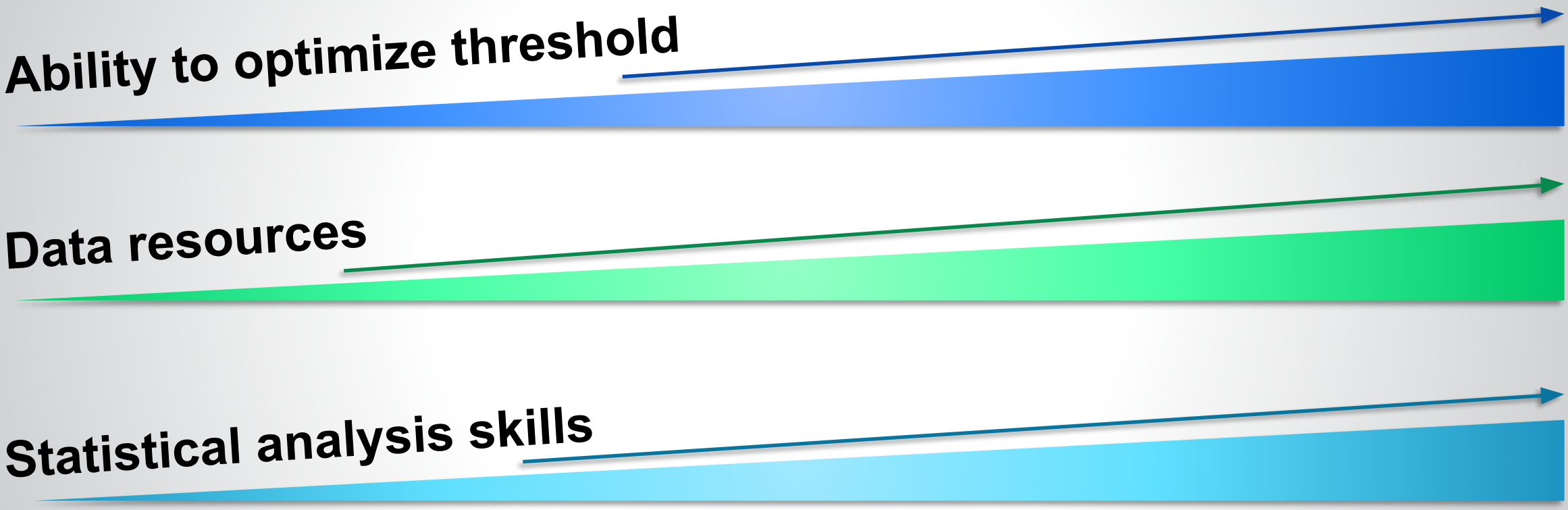
There are four main strategies for selecting a threshold score:

1. **Comprehensive CAD calibration study** (“TDR” toolkit) – prospective or retrospective operational research
2. **Iterative threshold score calibration** (ITSC) – iterative cycles of operational research in parallel to implementation
3. **Reactive adjustment** – adjusting the threshold score in response to programmatic outcomes
4. **Set and forget** – choosing one threshold score using literature/past experience/manufacturer recommendation and remaining there.

The most appropriate strategy to use depends on the availability of resources, such as:

- Staff with the correct skills
- Time available
- Data collected
- Availability of confirmation tests

Threshold Score Selection Strategy



THRESHOLD SCORE IN THIS PROGRAMME

Threshold Score in [country name]– for NTP customization

The current threshold score is X.

This threshold score was arrived at [insert how: recommended by the manufacturer? Based on previous experience? Based on operational research?]

- The more details the better – could present research results.

All people with chest X-ray images with scores >X will be classified by “TB” as AI and need confirmatory testing.

Summary

- A threshold score is a numerical output score used by CAD to classify chest X-ray images as “No signs of TB” or “Possibility of TB” based on how the abnormality score compares to the threshold.
- If using classification alone to triage patients, the threshold score determines key programmatic outcomes for a CAD screening intervention.
- Low threshold scores result in higher sensitivity and needing to test more people, so there is reduced cost savings and increased likelihood of over-diagnosis.
- A threshold score can be chosen to meet a programmatic goal, but research using locally collected data is required to do this accurately.
- There are four strategies for selecting a threshold score. Some of these strategies require large amounts of data and detailed statistical analysis.
- **The threshold score selected for this program is X and this was chosen by [method]**

Question 1:

True or False

Although it is recommended, a universal threshold score has not yet been decided on for all CAD products.

Question 2. If a high threshold score is set, what are the potential impacts on screening? (Choose all that apply)

A. Over-diagnosis
B. Low sensitivity
C. Under-diagnosis
D. More people will need confirmatory tests