Introduction to Threshold Selection

MODULE 4

Stop TB Partnership

Course Outline

- Threshold score selection
- How to select a threshold score suitable for the local context
- How to analyze program data for threshold selection
- Planning for screening

INTRODUCTION

This module introduces the key concepts of threshold score selection when using CAD and proposes several different strategies for how a threshold score should be selected that is suitable for the local context.

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.
- Understand some of the current strategies for adapting and optimizing a threshold in the local context.

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:

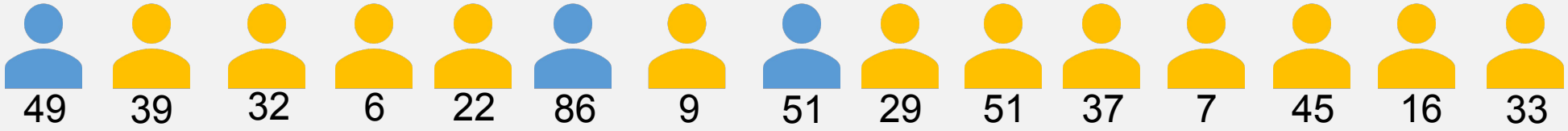
- **High** sensitivity but **low** specificity
  - More X-rays will have scores above the threshold, but a smaller proportion of these will have TB based on a diagnostic test.
- Needing to test **more** people to find a positive case, and therefore needing more diagnostic tests
- Increasing likelihood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



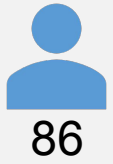
Person without TB

Person with TB

Population with "Possibility of TB" according to CAD

Situation A:
Saving on diagnostic tests

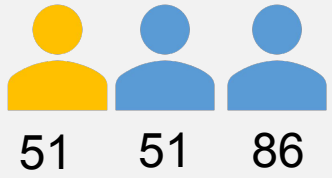
Threshold score is 75.



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

Situation B:
Optimizing sensitivity with resource constraint

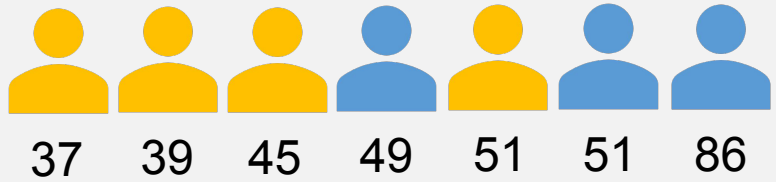
Threshold score is 50.



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 SELECT A THRESHOLD SCORE SUITABLE FOR THE LOCAL CONTEXT

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.

How to Choose a Threshold Score

There are four main strategies for selecting a threshold score:

1. Set and forget
2. Reactive adjustment
3. Iterative threshold score calibration (ITSC)
4. Comprehensive CAD calibration study (“TDR” toolkit)

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

Note: The threshold score can only be modified by the manufacturer in the back-end of the CAD software. Contact the manufacturer directly for this.

Set and Forget

The selection of a threshold score is kept for the duration of the implementation.

Sources of initial threshold could include:

- Prior experience with CAD products (ideally, the same product)
- Research using CAD literature (ideally, the same product and similar population)
- Recommended or default score from the CAD supplier

This strategy rests on the (unlikely) assumption that CAD performance will be the same in the target population as in the population used in the source of the threshold selection (e.g., the study in the CAD literature on which a chosen threshold is based).

Ideally, thresholds selected in this way should be optimized (using the prior strategies).

“Set and Forget” may be a practical compromise if resources are not available.

Note: Shortcomings in CAD Literature

If planning to use CAD literature to select the initial threshold score, it is important to be aware of a several limitations:

New versions of CAD software become available rapidly and require evaluation as the underlying AI model is likely different in newer versions compared to older versions.

CAD's ability to detect **non-TB abnormalities** has not been validated, even though products' ability to do this is often marketed.

Many studies are based on the area under the receiver operating characteristic curve. **More precise and implementation-relevant measures** should be explored.

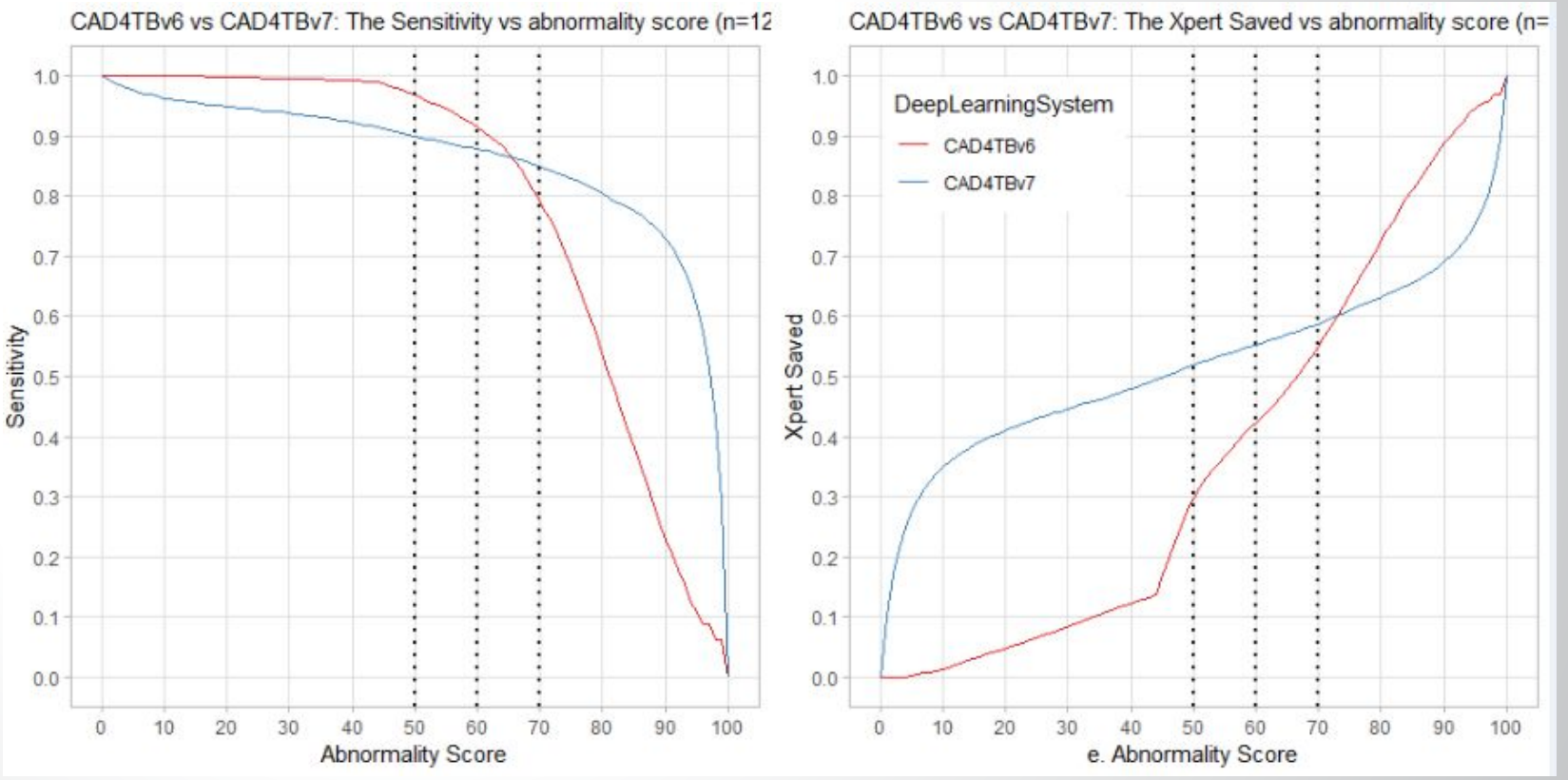
The performance of CAD in **children, risk groups,** and **TB key populations** needs more examination.

Many studies in CAD literature are **conducted with the involvement of the manufacturer** and focus on one product (CAD4TB) in particular.

Note: 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



Reactive Adjustment

Adjustment of a threshold score already selected (e.g., the one recommended by the manufacturer) by small increments in reaction to the occurrence of undesirable outcomes

- Undesirable outcomes: for example, CAD missing large numbers of people with TB, or a low positive confirmation test rate

- Performed **in parallel** to the implementation
- Similar to ITSC but **without concrete statistical methodology** and therefore potentially less accurate
- Requires **less statistical expertise** than previous strategies

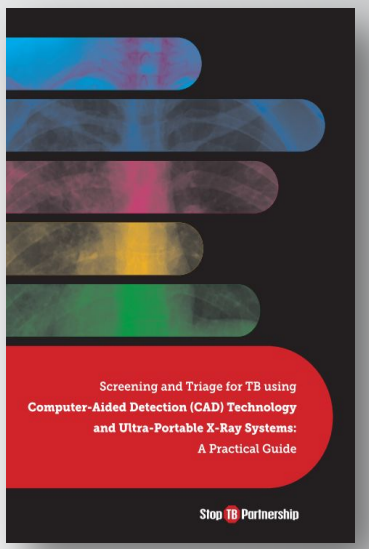
Data required:

- Participant demographic and clinical information
- Digital chest X-ray
- CAD score
- Confirmatory diagnostic test data (e.g., Xpert results)

Iterative Threshold Score Calibration (ITSC)

Strategy proposed by the Stop TB Partnership and Google.

- Requires setting an initial threshold score, then refining the initial score through ongoing rounds of data analysis until a target outcome is reached
- Can be performed **in parallel** to implementation
- If done correctly, selects a threshold score based on a targeted outcome (Xpert testing rate, Xpert positive rate, sensitivity, or confirmatory tests saved, for example)
- **Substantial statistical data analysis skills** required (may be necessary to engage an expert)



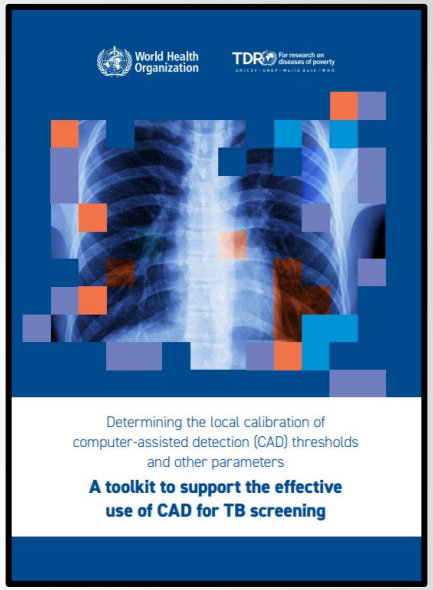
Data required:

- Participant demographic and clinical information
- Digital chest X-ray
- CAD score
- Confirmatory diagnostic test data (e.g., Xpert results) for all participants, or only for those with abnormality scores greater than the threshold

Comprehensive CAD Calibration Study

Strategy proposed by WHO and the Special Programme for Research and Training in Tropical Diseases (TDR)

- Involves conducting research in the population **in which CAD will be used**
- Can be **prospective** (before implementation) or **retrospective** (after implementation, to revise threshold score), depending on data resources
- If done correctly, selects a threshold score optimized for the population and use case
  - If TB prevalence in the population is low, large numbers of individuals may have to be screened to provide a sufficient sample.
- Also requires **substantial statistical analysis skills**



Data required:

- Participant demographic and clinical information
- Digital chest X-ray image
- CAD output score
- Confirmatory diagnostic test data for all participants (e.g., Xpert results)

Comprehensive CAD Calibration Study

Must be conducted in the same groups and regions where the tool will be used

Types of study:

- Cross-sectional
- Case-control

Do not forget any required ethical reviews!

Calibration study should not be used to make clinical decisions.

Comprehensive CAD Calibration Study

Cross-sectional study

Prospective study conducted with target groups and sites

Each eligible* participants will undergo:

- **Collection of key demographic and clinical patient information** (TDR toolkit has a data collection template)
- **Digital chest X-ray and reading** with CAD product
- **Collection of sputum samples** for testing with reference standard test (culture, WRD, etc.)

*** "Eligible participants" are ALL individuals in the selected use groups and sites.**

Sample Size Required for Cross-sectional Study and Level of Sensitivity (Based on 5 Percent Precision)

Cross-sectional Design	Sensitivity				
	50%	**60%**	**70%**	**80%**	**90%**
Number of **confirmed TB cases** required (assuming TB prevalence of 100%)	384	369	323	246	138
Number of **persons to screen** for reaching the expected number of TB cases, if the TB prevalence is **200**/100,000 persons	384 x 100,000 / 200 = 192,000	369 x 100,000 / 200 = 184,000	323 x 100,000 / 200 = 161,500	246 x 100,000 / 200 = 123,000	138 x 100,000 / 200 = 69,000
Number of persons to screen for reaching the expected number of TB cases, if the TB prevalence is **500**/100,000 persons	384 x 100,000 / 500 = 76,800	369 x 100,000 / 500 = 73,800	323 x 100,000 / 500 = 64,600	246 x 100,000 / 500 = 49,200	138 x 100,000 / 500 = 27,600

Comprehensive CAD Calibration Study

Case-control study

- Retrospective methodology conducted using data from the target groups and sites
- Individuals selected **separately and intentionally** on the basis of their **TB status** (cases or controls)
- Uses pre-existing patient data (outpatient department records, clinic records, prevalence surveys, and community screening) to conduct the calibration study
  - But must use data representative of the population intended to screen
- *May* be faster than a prospective study

Sample Size Required for Case-Control Study and Level of Sensitivity (Based on 5 Percent Precision)

Case-control design	Sensitivity				
	50%	**60%**	**70%**	**80%**	**90%**
Number of **confirmed TB cases** required (assuming TB prevalence of 100%)	384	369	323	246	138
Number of **confirmed non-TB cases** required (assuming the same precision and similar specificity as for the **cross-sectional** study)	384	369	323	246	138
Overall enrollment size required	**768**	**738**	**646**	**492**	**276**

Comprehensive CAD Calibration Study

Following the study (either cross-sectional or case-control), consider defining different CAD thresholds for sub-groups, such as:

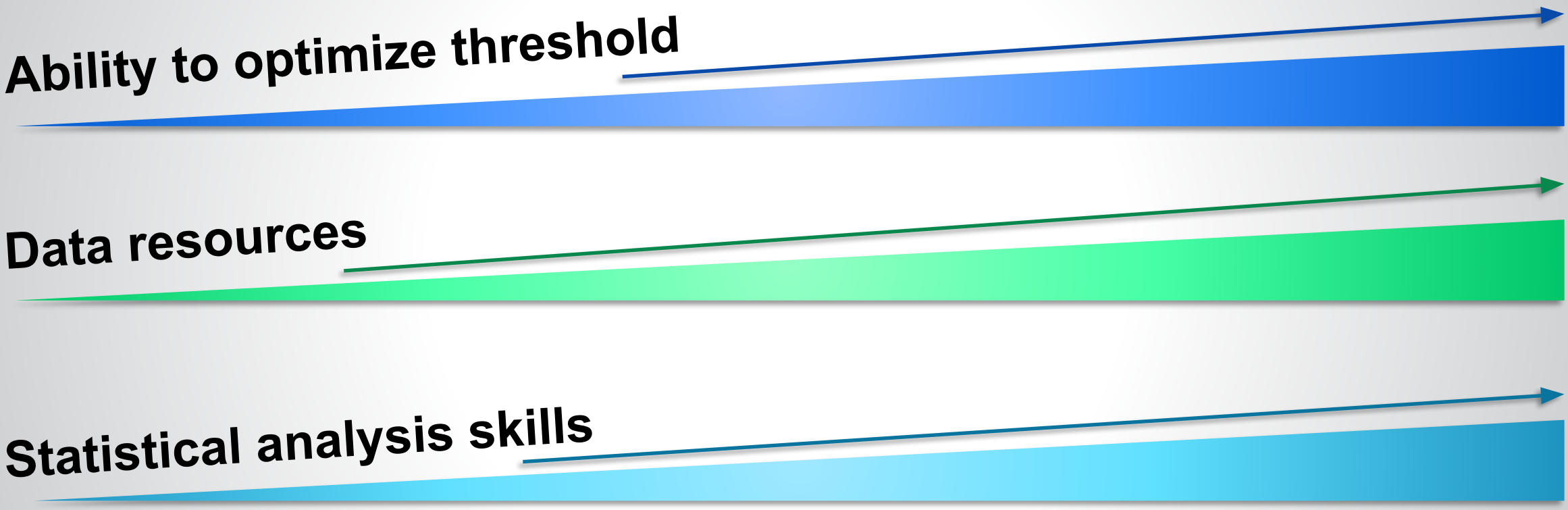
- Patient age
- HIV status
- Prior TB history
- Local prevalence

Comprehensive CAD Calibration Study

Exercise: Compare study designs

Cross-sectional study	Case-control study
General conditions for selecting study design	
Sample size	
Benefits (advantages)	
Weaknesses (disadvantages)	
Requirements for study	

Threshold Score Selection Strategy



Note: Manufacturer Calibration Study

- When procuring through GDF, a small scale CAD threshold score calibration study is included in the package.
  - Delft analyze a **maximum** of 200 chest X-ray images (100 containing TB, 100 “normal” images).
  - Analysis is in line with the WHO TDR toolkit method.
  - Based on the results, they can advise on a good threshold score to operate at.
- **Pros:** Uses local data, useful when just starting out with CAD so not enough sample size to do a larger scale study, helpful where there are no resources.
- **Cons:** Small scale, only offered once- will need to use another method to tailor threshold score if screening population changes.

HOW TO ANALYZE PROGRAM DATA FOR THRESHOLD SELECTION OPTIMIZATION

Data Analysis for Threshold Optimization

The Decision Analysis Framework can be used to monitor the accuracy and programmatic implications of using CAD software and may be used to inform threshold optimization.

The Framework uses three indicators, each relating directly to a programmatic goal:

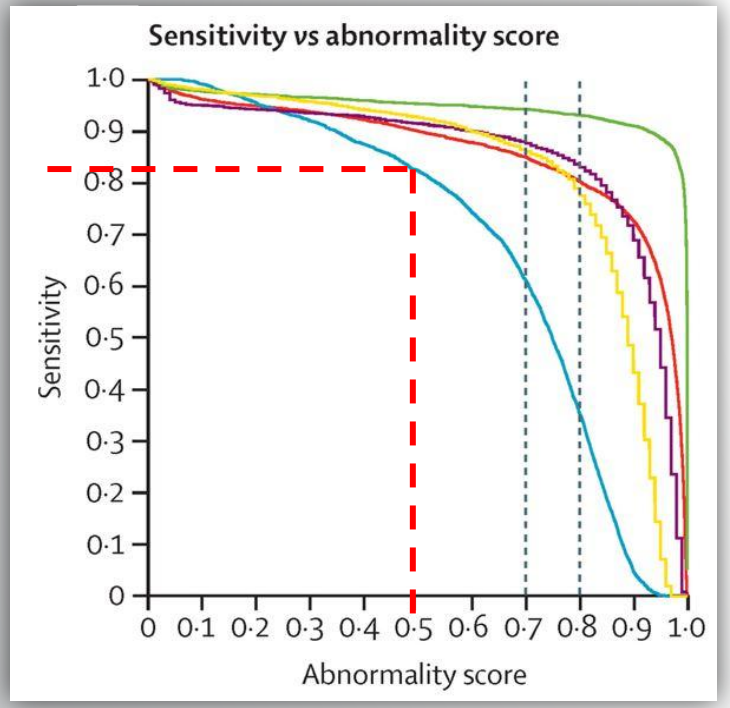
Indicator	Definition	Related performance/ programmatic goal
Sensitivity	True positive rate, ability of CAD to correctly identify people with TB in the population	High accuracy, maximizing TB cases detected
Number needed to test (NNT)	The number of people with a CAD score higher than the threshold who would need to be tested to find one person with TB	CAD's ability to triage
Proportion of confirmatory tests saved	The proportion of confirmatory tests that would be needed when using CAD as a triage tool, compared to the number without using CAD as a triage tool	CAD's cost-effectiveness

The effect of operating CAD at every threshold score in its range is modeled for the three indicators, and this is visualized as four key graphs.

Data Analysis for Threshold Optimization—Example

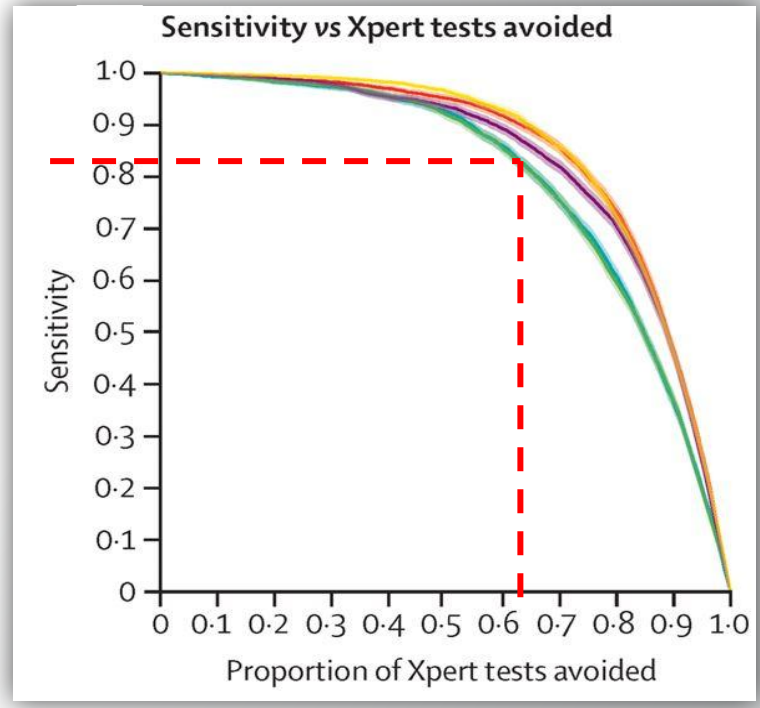
This example cites an application of the Framework to data from TB screening centers in Bangladesh. Different colored lines represent different CAD products.

It is possible to read the graphs for the effect of setting the threshold at different values.



A threshold score of 0.5 (or 50) for the **blue** product would result in sensitivity >80 percent (around 82 percent).

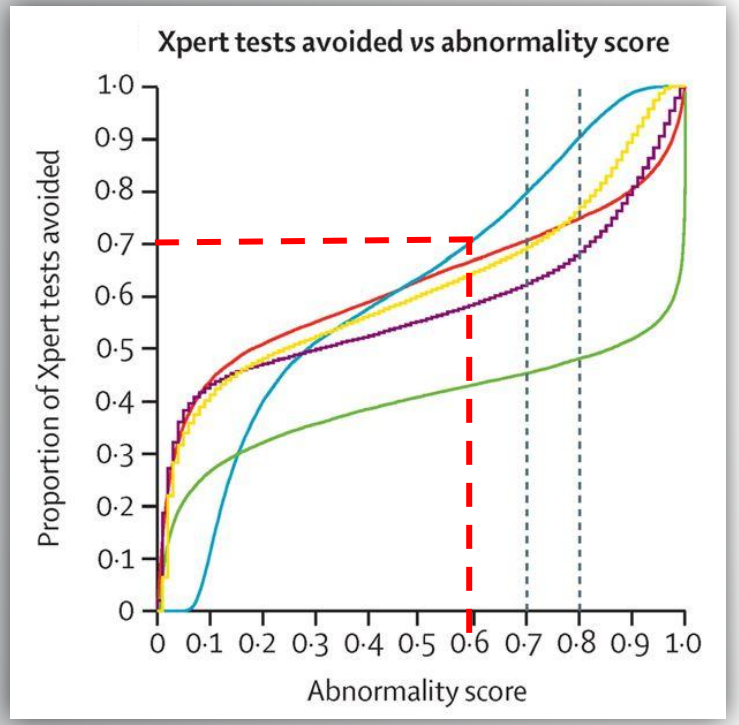
At this sensitivity for the **green** product, just over 0.6 (or 60 percent) of Xpert tests would be saved by using CAD as a triage tool.



Data Analysis for Threshold Optimization—Example

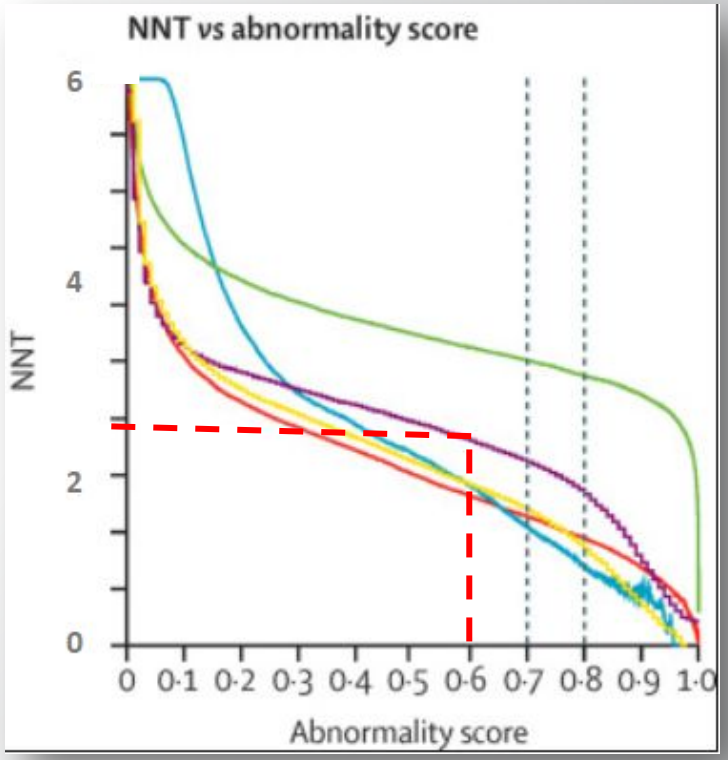
This example cites an application of the Framework to data from TB screening centers in Bangladesh. Different colored lines represent different CAD products.

It is possible to read the graphs for the effect of setting the threshold at different values.



If wanting to save 70 percent of Xpert tests and using the **blue** product, the threshold score should be set at around 0.6.

If using a threshold of 0.6 with the **purple** product, the NNT would be around 2.6.



Exercise—Selecting Thresholds in line with Programmatic Goals

Use the graphs provided to determine what threshold to use for each of these products in the following scenarios:

1. An active case finding project with limited budget for confirmatory (Xpert) tests that would like to reduce Xpert testing by 60 percent
2. An immigration screening program that needs to achieve at least 95 percent sensitivity

A program would like to operate at the WHO target sensitivity for a TB triage test (90 percent sensitivity). Use the graphs to tell:

3. What is the threshold score they should use?
4. What is the NNT?
5. What is the proportion/percentage of Xpert tests that would be saved?

PLANNING FOR SCREENING

Start How You Want to End

- Setting the threshold score will impact the cascade of care.
- If you increase the number of presumptive TB patients requiring follow-on testing, how will you meet the additional need?
- If you reduce the number of presumptive TB patients requiring follow-on testing, will you have additional testing capacity to deploy?

Start How You Want to End

- Are any infrastructural changes needed to accommodate additional testing needs?
- Does existing infrastructure limit the number of people we can test?

You Can Revisit the Threshold Score

- Over time, you may find the originally selected threshold score is no longer reflective of programmatic goals.
- Routine reviews of the CAD threshold score and the implications on sensitivity and specificity should be considered, especially as retrospective data (the case-control model) accumulates.

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 Decision Analysis Framework suggests some key indicators that can be calculated to monitor a CAD intervention and may be used to optimize the threshold score.