Slide set II
Towards Good Review Practice: a training manual for the review of clinical performance of in vitro diagnostic medical devices

These materials were produced with support from Grand Challenges Canada through a grant for Improving Regulatory Oversight Of In-Vitro Diagnostics In The Developing World: Affordable access to in-vitro diagnostics through regulatory harmonization approaches and UNITAID through a grant for A Global Network to Improve Access and Quality of HIV Monitoring Technologies: Better regulation for point-of-care HIV devices
Contents

1. Introduction to assessing clinical performance data and good review practice (GRP)
2. Risk Classification
3. Intended use
4. Performance indicators
5. Standards and guidelines for the clinical validity of in vitro diagnostic medical devices
6. Study design and bias
7. Samples size
8. Reference standards
9. Accuracy and precision
10. Data interpretation
11. Case studies
12. Data from scientific literature and meta analysis
1
Introduction to assessing clinical performance data and good review practice (GRP)
The ideal IVD would be...

- Reliable – always work
- Accurate – provide a correct result
- Robust – compatible with working and storage conditions
- Affordable – to meet budget constraints
- Available – in sufficient supply to meet demand
- Compatible – appropriate for the population with which it will be used
Evaluating IVDs

R&D
Field testing
Optimisation

How reliable is the test?
Lab studies on analytical performance
Clinical performance studies

How useful is the test?
Impact & cost benefits

Which technologies? Where/when to use them?
Technology Assessment

Regulatory approval

Procurement
Analytical Performance Studies

Purpose

• How well does the test perform under the best circumstances?

Key measurements

• Analytical sensitivity
• Analytical specificity
• Reproducibility
• Others...
Clinical Performance Studies

Purpose

• To measure the ability of the test to detect and predict the condition that is associated with an analyte measurement

Key measurements

• Clinical sensitivity
• Clinical specificity
Clinical Utility Studies

Purpose

• To measure the risks and benefits associated with the use of a test: usefulness of information for patient management and for disease control
• To determine the cost-effectiveness of a diagnostic test, feasibility of test introduction, and strategies for scale up

Key measurements

• Reduction in mortality, morbidity, or improvement in quality of life associated with diagnostic use
• Adverse effects of diagnostic use or treatment
• Feasibility of test introduction
• Cost-effectiveness and strategies for scale-up
Clinical trials

Clinical trials are prospective studies which measure the impact of diagnostic tests (or algorithms) on patient health.

They are undertaken after the performance of the IVD has been studied and found safe.

Such studies should be registered with a clinical trials database e.g. https://clinicaltrials.gov/
Evaluating IVDs

R&D
Field testing
Optimisation

How reliable is the test?
Lab studies on analytical performance

Clinical performance studies

How useful is the test?
Impact & cost benefits

Which technologies? Where/when to use them?
Technology Assessment

Procurement

Regulatory approval
Regulating IVDs

- The primary goal is to protect public and individual health and safety
- Do the potential benefits outweigh risk of harm?
- Regulatory decisions must be based on strong and clear science, free of external influences and consistent with the directives of law
Definition

Clinical performance:

The ability of an IVD medical device to yield results that are correlated with a particular clinical condition/physiological state in accordance with target population and intended user.

Clinical Performance Study

A study undertaken to establish or confirm the clinical performance of an IVD medical device

GHTF/SG5/N6:2012

NOTE: This term is synonymous with ‘clinical trial’ and ‘clinical study’ but a clinical performance study for an IVD is not equivalent to a clinical trial for a medicine or vaccine
What is the purpose of a clinical performance study for an IVD?

To demonstrate, with objective evidence, how the IVD will be expected to perform in routine clinical practice.
Clinical performance studies

Setting
  • Settings of intended use, performed by intended users

Sample type and size
  • Prospectively collected clinic samples
  • Sample size depends on disease prevalence and expected difference in performance between reference standard test and the test under evaluation

Evaluation outcomes
  • Clinical performance
  • Operational characteristics
Assessing Clinical Performance
Study Design

- Study should be designed to provide data necessary to show clinical performance of the IVD
- Should account for potential risks
- Should follow ethical principles
- Should be compliant with all relevant legal and regulatory requirements
Assessing Clinical Performance Study Design

• Specimen collection and handling
• Clinical performance study site
• Statistical design
• Potential risks
• Ethical considerations
Clinical Performance Study Protocol

Document that states the rationale, objectives, design and proposed analysis, methodology, monitoring, conduct and record-keeping of the clinical performance study.

GHTF/SG5/N8:2012
Operational Characteristics that should be considered

Performance requirements

- Specimen types
- Operating temperatures for the IVD at site of intended use
- Sensitivity, specificity, etc., degree of confidence in these values, and definition of populations
- Interfering factors (substances and medical conditions)
Operational Characteristics that should be considered

• Cultural requirements
  • E.g., acceptability of bovine, porcine materials

• Usability and human factors
  • Ease of use
  • Ease of learning, training, staff technical ability
  • Accessibility
    • Color blindness
  • Power sources
Important Note

• Don’t reject a test because it doesn’t meet every requirement for every setting

• No test is perfect!
Good Review Practices

Definition

- Documented best practices for any aspect related to the process, format, content, and/or management of a medical product review

Goal

- Promote timeliness, predictability, consistency, transparency, clarity, efficiency, and high quality of the content and the management of reviews

Mechanism

- Development of review tools (e.g., SOPs, templates) and reviewer learning activities (e.g., training courses, mentoring, orientation packages, discussion sections)
- Evaluate and update on an ongoing basis
Principles of a Good Review

• Evidence-based
  Scientific and state-of-the-art

• Utilizes critical analyses
  Assesses scientific integrity, relevance, and completeness of data, labeling, and interpretation

• Identifies signals
  Highlights areas of concern

• Investigates and problem-solves
  Devise and recommend critical solutions and efficient alternatives where needed

• Makes linkages
  Integrated analysis across all aspects of an application
Principles of a Good Review

• Considers context
  Proposed conditions of use and storage, including perspectives from healthcare professionals and other Regulatory Authorities

• Involves consultation
  Internal and external

• Balanced
  Objective and unbiased

• Thorough

• Well-documented
  Professional, well-written, clear, neutral, respectful
Activities Critical to Good Review Practice: Managing the Review

- Project management
- Quality management
Activities Critical to Good Review Practice: Managing the Review

• SOPs, guidances, checklists, and templates

• Communication
  • Intra-agency (regulatory authority)
  • Inter-agency
  • With the applicant
  • With the public

SOPs = Standard Operating Procedures
Critical Activities

• Personnel competencies and training
  • Reviewer training
  • Critical thinking
  • External experts
  • Collaboration and leveraging

• Conducting the review
  • Defining a review strategy
    • Public health priority of product
    • Understanding other RAs action on the application
    • Understanding specific intrinsic and extrinsic factors
    • Assessment of submission quality
    • Identification of major scientific questions and their possible resolution
  • Evidence-based review with a risk-benefit result
Best Practice Assessment Principles

• Global Harmonization Task Force (GHTF)
  • Founded in 1992
  • Group of representatives from global national medical device regulatory authorities (Australia, Canada, EU, Japan and USA) and from industry
  • Purpose: Convergence of regulatory practice and identification of best regulatory practice
  • Publication of documents
## Resources:

http://www.imdrf.org/ghtf/ghtf-archives.asp

Study Group 5: Clinical Safety/Performance

<table>
<thead>
<tr>
<th>DOCUMENT ID</th>
<th>NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>GHTF/SG5/N8:2012</td>
<td>Clinical Evidence for IVD Medical Devices - Clinical Performance Studies for In Vitro Diagnostic Medical Devices</td>
</tr>
<tr>
<td>GHTF/SG5/N5:2012</td>
<td>Reportable Events During Pre-Market Clinical Investigations</td>
</tr>
<tr>
<td>GHTF/SG5/N4:2010</td>
<td>Post-Market Clinical Follow-Up Studies</td>
</tr>
<tr>
<td>GHTF/SG5/N3:2010</td>
<td>Clinical Investigations</td>
</tr>
<tr>
<td>GHTF/N2R8:2007</td>
<td>Clinical Evaluation</td>
</tr>
<tr>
<td>SG5/N1R8:2007</td>
<td>Clinical Evidence – Key Definitions and Concepts</td>
</tr>
</tbody>
</table>
Purpose of this Training

How to review clinical performance studies

• What to look for
  • Identify necessary elements and if they have been submitted for your assessment

• How to how to assess it
  • Determine if what is submitted is acceptable
    • Clear and complete presentation of data?
    • Accurate data summaries?
    • Appropriate conclusions (interpretation of data)?

How to apply what you’ve learned
End of Module 1

- Move to the next module
- Return to slide 1 of this module
- Return to the start of the course
- Move to the end of the course
2
Risk Classification
Definitions:

**Risk:** Combination of the probability of occurrence of harm and the severity of that harm

**Hazard:** Potential source of harm
Regulating IVDs

IVDs differ from medicines and vaccines

• Many more products on the market with shorter product life due to frequent addition of new/updated products
• IVDs are not ingested (reduced risk)

Guiding principles published 2001 (PAHO/US FDA)

2. Marketing Controls: proper use, accurate marketing.
3. Post-marketing Controls: to maintain vigilance and continued safety and quality of approved products.
Regulatory principles

• The primary goal is to protect public and individual health and safety

• A regulatory system should ensure that valuable new technologies are made available to the clinical community, patients and consumers expeditiously while preventing unsafe or ineffective devices from reaching the market

• **Regulatory control should be proportionate to the risk for harm**
Risk Classification

Based on regulatory principles:

i. Safety and performance is assessed through premarket scrutiny, adequate manufacturer quality management systems and implementation of effective post-market surveillance mechanisms

ii. Level of scrutiny depends upon the risk of harm
Risk Classification

The Global Harmonization Task Force (GHTF) created risk classification rules to determine the level of premarket regulatory assessment required.

GHTF/SG1/N045:2008
“Principles of In Vitro Diagnostic (IVD) Medical Devices Classification”
Risk Classification

• Classifies IVDs according to the risk they pose to personal and public health

• Takes into account potential outcomes if the test does not perform properly

• Determines the level of scrutiny applied for the assessment

• Ensures that resources are focused on those IVDs associated with the greatest potential risk

• Moderates degree of premarket assessment of IVDs considered low risk
## Risk Classes

### GHTF Risk Classes and Risk Level

<table>
<thead>
<tr>
<th>Classification</th>
<th>Personal Health Risk</th>
<th>Public Health Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class A IVD</td>
<td>Low</td>
<td>and Low</td>
</tr>
<tr>
<td>Class B IVD</td>
<td>Moderate</td>
<td>and Low</td>
</tr>
<tr>
<td>Class C IVD</td>
<td>High and/or</td>
<td>Moderate</td>
</tr>
<tr>
<td>Class D IVD</td>
<td>High</td>
<td>and High</td>
</tr>
</tbody>
</table>
## Regulatory requirements

<table>
<thead>
<tr>
<th>RISK CLASS</th>
<th>Regulatory requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Premarket review of Summary Technical Documentation (STED)</td>
</tr>
<tr>
<td>B</td>
<td>Premarket review of Summary Technical Documentation (STED)</td>
</tr>
<tr>
<td>C</td>
<td>Premarket review of Summary Technical Documentation (STED)</td>
</tr>
<tr>
<td>D</td>
<td>Premarket review of Summary Technical Documentation (STED)</td>
</tr>
</tbody>
</table>
Rules

Risk Classification Rule 1

IVD medical devices intended for the following purposes are classified as Class D

• Devices intended to be used to detect the presence of, or exposure to, a transmissible agent in blood, blood components, blood derivatives, cells, tissues or organs in order to assess their suitability for transfusion or transplantation

• Devices intended to be used to detect the presence of, or exposure to, a transmissible agent that causes a life-threatening, often incurable, disease with a high risk of propagation
Risk Classification Rule 3

IVD medical devices are classified as Class C if they are intended for use (abbreviated list):

• In detecting the presence of, or exposure to, a sexually transmitted agent

• In detecting the presence of an infectious agent where there is a significant risk that an erroneous result would cause death or severe disability to the individual or fetus being tested

• In the management of patients suffering from a life-threatening infectious disease.
Differences between Class C and D

• Documentation for a Class C will contain less elaborate information than the STED for a Class D device.

• Main difference for a Class D STED would be in the level of details in the clinical/performance data and details of the manufacturer’s QC release program.

• The Regulatory Authority/Conformance Assessment Body should, in the review process, not normally require more elaborate information for a Class C device. However, this does not preclude the Regulatory Authority/Conformity Assessment Body from requesting such information in specific cases.

GHTF/SG1/N046:2008)
Global situation

- Australia, Canada, Tanzania and WHO PQ use the GHTF model

- EU (CE-Mark) Currently not based on a stated risk classification system but proposed new regulations will adopt GHTF model

- US FDA follows a three tier system

<table>
<thead>
<tr>
<th>Risk Class</th>
<th>Level of Risk</th>
<th>Corresponding GHTF Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class I</td>
<td>Low to moderate</td>
<td>A</td>
</tr>
<tr>
<td>Class II</td>
<td>Moderate to high</td>
<td>B or C</td>
</tr>
<tr>
<td>Class III</td>
<td>High</td>
<td>C or D</td>
</tr>
</tbody>
</table>
Determining Risk to Health: Asking the Right Questions

What is the endemicity and prevalence of the infectious agent or condition that the IVD is detecting?
Determining Risk to Health:

What is the level of care available for a patient with the disease?
Determining Risk to Health:

What is the availability of additional or follow-up testing?
Determining Risk to Health:

How competent are the practitioners/users of the test?
Determining Risk to Health:

How important is the test result to the diagnosis (sole determinant or one of several)?
Determining Risk to Health:

What is the impact of the result (true or false) to the individual and/or to public health?
Example:

HIV Rapid Diagnostic Test (RDT)

• Purpose and role: Test used as a basis for diagnosis or as an aid in the diagnosis of HIV infection.

• Conclusion: High public health risk and high individual health risk = Class D

• Specifically identified as Class D in risk classification rules
Malaria RDT

Purpose and role:

• High income countries: May be one of several factors used to evaluate whether a patient has malaria

• Low-income countries: May be sole basis for diagnosis of malaria

Other factors:

• High income countries: Readily available healthcare and additional testing, low endemicity, no vectors (no opportunity for transmission)

• Low income countries: Poor access to healthcare and additional testing, high endemicity and possibility for transmission
Malaria RDT

**Low-income Country**
- Moderate public health risk
- High individual health risk

= Class C

**High-income Country**
- Low public health risk
- Moderate individual health risk

= Class B
What risk classification would you assign to

i. A HIV viral load assay

ii. A HIV early infant diagnosis assay

iii. A CD4 assay
HIV

Individual risk

- **CD4** (if used to initiate therapy)
- **EID**
- **RDT for HIV infection**

Viral load

Public health risk

- **RDT for HIV infection**
- **CD4**
- **Viral load**
- **EID**
TB

Individual risk

*Rapid POC diagnostic test
Xpert MTB/RIF
Stains
Culture media
Tests for latent TB

Public health risk

*Rapid POC diagnostic test
Xpert MTB/RIF
Stains
Culture media
Tests for latent TB

* Not yet available
Determining Risk:

• Risk classification may vary depending on the setting in which a product is intended to be used
• Factors may differ between resource-limited settings compared to high-income countries
• Takes into account potential outcomes if the test does not perform properly or is not available
Further reading

www.who.int/diagnostics_laboratory/evaluations/en/
End of Module 2

Move to the next module

Return to slide 1 of this module

Return to the start of the course

Move to the end of the course
3

Intended use
Definition: Intended Use

How the test is to be used: what condition, what sample, what patient?
Includes indications for use: Why test?

The objective intent of the manufacturer regarding the use of a product, process or service as reflected in the specifications, instructions and information provided by the manufacturer. (GHTF/SC/N4: 2011)
Performance evaluation:

The assessment and analysis of data to establish or verify the ability of an in vitro diagnostic medical device to achieve its intended use.
When reviewing clinical performance data it is important to establish the intended use of the test.

Did the sampling and study design reflect the intended use?
Diagnostic tests have different uses

<table>
<thead>
<tr>
<th><strong>Role of Diagnostics</strong></th>
<th><strong>Examples</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Personal health:</strong></td>
<td></td>
</tr>
<tr>
<td>• Diagnosis of infection</td>
<td>Malaria</td>
</tr>
<tr>
<td>• Aid to diagnosis (guide clinical management</td>
<td>Urine dipstick, Dengue, influenza</td>
</tr>
<tr>
<td>decisions)</td>
<td></td>
</tr>
<tr>
<td>• Screening for asymptomatic infections</td>
<td>Tuberculosis, HIV, Syphilis</td>
</tr>
<tr>
<td>• Initiate treatment</td>
<td>CD4 count</td>
</tr>
<tr>
<td>• Monitor treatment effectiveness</td>
<td>Viral load assay for HIV</td>
</tr>
<tr>
<td>• Predisposition</td>
<td>Human papilloma virus</td>
</tr>
<tr>
<td>• Determination of physiological state</td>
<td>Blood glucose testing</td>
</tr>
<tr>
<td><strong>Public health:</strong></td>
<td></td>
</tr>
<tr>
<td>• Identify infection to aid preventing further</td>
<td>Tuberculosis, sexually transmitted infections</td>
</tr>
<tr>
<td>transmission</td>
<td></td>
</tr>
<tr>
<td>• Aid outbreak investigations and disease control</td>
<td>Ebola, influenza, dengue</td>
</tr>
<tr>
<td>• Surveillance and aiding eradication programs</td>
<td>Schistosomiasis, African trypanosomiasis, Leishmaniases</td>
</tr>
</tbody>
</table>
Indications for Use for IVDs
GHTF/SG5/N8:2012

• Diagnosis
• Aid to diagnosis
• Screening
• Monitoring
• Predisposition
• Prognosis
• Prediction of treatment
• Determination of physiological status
Diagnosis

• Used as *the sole basis for making a clinical decision*

• Designed to evaluate a *patient’s current state*

• Example: Malaria RDT when used as the only basis for making a clinical conclusion
Aid to Diagnosis

• Used to provide *additional information to make a clinical decision* (not the sole basis)
• Designed to evaluate a *patient’s current state*
• Example: HIV RDT when used with other factors to determine infection
Screening

• Used to determine the status of a disease, disorder, or other physiological state in an asymptomatic individual

• Designed to evaluate an individual’s current state

• Example: Tests to detect infectious agents, or evidence of infection, in donated blood
Monitoring

• Used for the measurement of analyte levels for the purpose of adjusting treatments/interventions as required

• Designed to evaluate an individual’s state or changes in an individual’s state

• Example: Viral load testing of patients known to be infected with HIV to determine treatment response and adjust therapy, if necessary
Predisposition

• Used to determine *likelihood of disease onset in presymptomatic patients*
• Designed to evaluate a *patient’s future state*
• Example: Determine gene mutation status to assess risk of disease
Prognosis

• Used to measure factors that determine *likelihood of patient responses* to a specific therapy (companion diagnostics, personalized medicine)
• Designed to evaluate a *patient’s future state*
• Example: Measurement of baseline HIV-1 RNA level to assess patient prognosis
Prediction
(of treatment response or reaction)

• Used to measure factors that determine the *likelihood of patient responses or adverse reactions to a specific therapy* (companion diagnostics, personalized medicine)

• Designed to evaluate a *patient’s future state*

• Example: Identification of metabolizer status to determine potential benefits and/or adverse reactions to a treatment
Determination of Physiological Status

• Used to evaluate the physiological state of an individual to *identify a human condition or characteristic*

• Designed to evaluate a *patient’s current state*

• Example: hCG test for determination of pregnancy

hCG = human chorionic gonadotropin, a hormone elevated in pregnancy
Factors that Impact Study Design: Indications for Use

**DIAGNOSIS or MONITORING**
(Symptomatic or Known)

- **DIFFERENT INCLUSION/EXCLUSION CRITERIA**
- **SMALLER sample size**

**SCREENING**
(Asymptomatic)

- **LARGER sample size**

*Diagnostic tests require high specificity*
*Screening tests require high sensitivity*
**Intended use**

X is a single-use immunochromatographic test for the detection of antibodies to Human Immunodeficiency Virus Type 1 (HIV-1) and Type 2 (HIV-2) in fingerstick whole blood, venous whole blood, serum or plasma specimens.

**Indications for use**

as a **point-of-care test** to **aid in the diagnosis** of infection with HIV-1 and HIV-2...
The Xpert MTB/RIF Assay is intended for use with specimens from patients for whom there is clinical suspicion of tuberculosis and who have received no anti-tuberculosis therapy, or less than 3 days of therapy. This test is intended as an aid in the diagnosis of pulmonary tuberculosis when used in conjunction with clinical and other laboratory findings.

The Xpert MTB/RIF Assay does not provide confirmation of rifampin susceptibility...must have results confirmed by a reference laboratory.

...should also be tested for the presence of genetic mutations associated with resistance to other drugs.

...must be used in conjunction with mycobacterial culture...

http://www.accessdata.fda.gov/cdrh_docs/reviews/K131706.pdf
End of Module 3

Move to the next module
Return to slide 1 of this module
Return to the start of the course
Move to the end of the course
4

Performance Indicators
Evaluating IVDs

R&D
Field testing
Optimisation

How reliable is the test?
Lab studies on analytical performance
Clinical performance studies

How useful is the test?
Impact & cost benefits

Which technologies? Where/when to use them?
Technology Assessment

Procurement

Regulatory approval
Evaluating IVDs

**Analytical accuracy**
Lab studies using well characterised samples and specimens.

**Clinical accuracy**
Performance when compared to the ‘reference standard’ when testing population/site of intended use

**Utility and impact**
Once safety has been established: impact on individual/public health, QALYs, DALYs, health system, economics etc (modelling)
What are the expected outputs of a clinical performance study?

Results from *clinical evaluation* of the IVD according to the study protocol

Generate *clinical data*

Which serve as *clinical evidence*

To support the *clinical performance*
Definition:

Clinical Evidence for an IVD

All the information that supports the scientific validity and performance for its use as intended by the manufacturer.

GHTF/SG1/N68:2012
GHTF/SG5/N6:2012
Definition

Clinical performance:

The ability of an IVD medical device to yield results that are correlated with a particular clinical condition/physiological state in accordance with target population and intended user.

Accuracy:
A combination of trueness and precision (ISO 3534-2)
Closeness of agreement between a measured quantity value and a true quantity value of what is being measured. (VIM/JCGM_200_2012)

Precision: The closeness of agreement between independent test results obtained under stipulated conditions (ISO 3534-1).
Is expressed quantitatively in terms of imprecision (standard deviation SD or coefficient of variation CV).
Clinical validity

Clinical validity refers to the accuracy with which a test can predict the presence or absence of the clinical disease or condition it is intended to detect.

Diagnostic accuracy

• The ability of a diagnostic test to differentiate between diseased and non-diseased subjects, or between two or more clinical states (CLSI/EP24-A2)

• The extent of agreement between the information from the test under evaluation and the diagnostic accuracy criteria (CLSI /EP 12-A2)
Clinical Performance Study Outputs

- Sensitivity
- Specificity
- Positive predictive value
- Negative predictive value
- Invalid rate
- Accuracy
Sensitivity:

The ability of a test to give a positive result for subjects who have the disease or condition for which they are being tested.

It is measured as the ratio of positive test results in those that have the condition to the total number who have the condition, and is often expressed as a percentage. (CLSI/EP24-A2)

i.e. % of infected people that have a positive test result
Sensitivity

\[
\frac{\text{# of True Positives}}{\text{# of True Positives} + \text{# of False Negatives}} \times 100
\]

e.g. Of 200 people with the disease 180 were found positive by the test.

Test sensitivity = 90%
Specificity:

The ability of a test to give a negative result for subjects who do not have the disease or condition for which they are being tested.

It is measured as the ratio of negative test results in those unaffected by the condition to the total number of condition-free subjects and is often expressed as a percentage. (CSLI/EP24-A2)

i.e. % not infected that have a negative test result
Specificity

\[
\text{Test specificity} = \frac{\# \text{ of True Negatives}}{\# \text{ of True Negatives} + \# \text{ of False Positives}} \times 100
\]

e.g. If of 500 people that do not have the disease 10 have a false positive test result

Test specificity = 98%
Positive predictive value PPV

• The probability that a positive result accurately indicates the presence of infection.
• The percentage of subjects with a positive test result who have the target condition.
• Measures the reliability of a positive test result
• PPV must be interpreted in context with the prevalence of the condition of interest (as determined by the diagnostic accuracy criteria).
  CLSI/EP12-A2
Positive Predictive Value (PPV)

% with a positive test result and who are infected

\[
\frac{\text{# of True Positives}}{\text{# of True Positives} + \text{# of False Positives}} \times 100
\]
Negative predictive value NPV

- The probability that a negative result accurately indicates the absence of infection.
- The percentage of subjects with a negative test result who do not have the target condition.
- Indicates the reliability of a negative test result.
- NPV must be interpreted in context with the prevalence of the condition of interest (as determined by the diagnostic accuracy criteria).

CLSI/EP12-A2
Negative Predictive Value (NPV)

% with a negative test result and who are not infected

\[\frac{\text{# of True Negatives}}{\text{# of True Negatives + # of False Negatives}} \times 100\]
2x2 tables

<table>
<thead>
<tr>
<th>Index test</th>
<th>Reference test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+ve</td>
</tr>
<tr>
<td>+ve</td>
<td>a</td>
</tr>
<tr>
<td>-ve</td>
<td>c</td>
</tr>
</tbody>
</table>

Sensitivity = $a/(a+c)$
Specificity = $d/(b+d)$
PPV = $a/(a+b)$
NPV = $d/(c+d)$
Sensitivity and Specificity vs. Positive and Negative Predictive Values

• Sensitivity and specificity are inherent characteristics of a test
• Positive and negative predictive values vary by prevalence in the population
Example:

• For a test that is 99% sensitive and 99% specific, and a sample of 10,000 individuals:

• If the prevalence of the analyte is 10%:
  \[ \text{PPV} = \frac{900}{900 + 90} = 90.1\% \]

• If the prevalence of the analyte is 1%:
  \[ \text{PPV} = \frac{99}{99 + 99} = 50\% \]
Example:

A test for resistance to the anti-tuberculosis drug rifampicin has a specificity of 98% = 2 false positives in 100 tests

- If the incidence of resistance among TB patients is 50%, 100 tests will give **50 true positive** and **2 false positive** results  \( \text{PPV} = 96\% \)

- If the incidence of resistance among TB patients is 1%, 100 tests will give **1 true positive** and **2 false positive** results  \( \text{PPV} = 33\% \)
### Example: Malaria

<table>
<thead>
<tr>
<th>Region</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highland epidemic</td>
<td>4.0%</td>
</tr>
<tr>
<td>Lake endemic</td>
<td>50.4%</td>
</tr>
<tr>
<td>Coast endemic</td>
<td>6.0%</td>
</tr>
<tr>
<td>Semi-arid and seasonal</td>
<td>0.7%</td>
</tr>
<tr>
<td>Low risk</td>
<td>0.6%</td>
</tr>
</tbody>
</table>

#### 2010 Kenya MALARIA Indicator Survey

- A new test has a sensitivity of 99% and specificity of 98%.
- Used in the lakes region 1000 tests will give **499 true positives** and **10 false positives** \( \text{PPV} = 96\% \).
- Used in Nairobi 1000 tests will give **6 true positives** and **20 false positives** \( \text{PPV} = 23\% \).
Point estimates

• The values for sensitivity, specificity etc derived from clinical performance studies are estimates (best guesses)

• The use of sample data to calculate a single value (statistic) to serve as a best guess for a population is called point estimation

• When reviewing statistical data it is important to consider the reliability of the point estimates presented

• Confidence in the statistics is affected by the number of data points (sample size) and the variability of the data
Confidence intervals (CI)

- Statisticians use a confidence interval to express the degree of uncertainty associated with a statistic.
- Confidence intervals are a measure of the reliability of an estimate; they indicate the precision of the estimate and the uncertainty of the estimate.
- They are usually (but not always) stated at the 95% confidence level.

Estimated sensitivity 85% (CI95: 72,98) - low confidence
Estimated sensitivity 85% (CI95: 83,87) - high confidence
## Other indicators

<table>
<thead>
<tr>
<th>False-negative result</th>
<th>Negative test result for subject in whom the disease or condition of interest is present (CLSI/EP24-A2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>False-positive result</td>
<td>Positive test result for subject in whom the disease or condition of interest is absent (CLSI/EP24-A2)</td>
</tr>
</tbody>
</table>
| False-negative fraction | • Ratio of subjects that have the disease, but have a negative test result, to all subjects that have the disease.  
• False negative/(false negative + true positive)  
• Equivalent to (1 - sensitivity) (CLSI/EP24-A2) |
| False-positive fraction | • Ratio of subjects that do not have the disease but have a positive test result, to all subjects that do not have the disease.  
• False positive/false positives + true negatives.  
• Equivalent to (1 - specificity) (CLSI/EP24-A2) |
Invalid Test Rate

An invalid test is one that does not meet the test requirements for an interpretable test result

\[
\frac{\text{# of Invalid Tests}}{\text{# of Invalid Tests} + \text{Total # of Tests}} \times 100
\]
How to review

• Are the numbers correct?

• Are all data taken into account?

• Ensure that there is a justification for any data point not included in calculations

• Is there an excessive number of invalid test results?
  • Source of invalids (test, operator)?
End of Module 4

Move to the next module

Return to slide 1 of this module

Return to the start of the course

Move to the end of the course
5
Standards and guidelines for the clinical validity of *in vitro* diagnostic medical devices
Clinical validity

Clinical validity (diagnostic accuracy) refers to the accuracy with which a test can predict the presence or absence of the clinical disease or condition it is intended to detect:

- The ability of a diagnostic test to differentiate between diseased and non-diseased subjects, or between two or more clinical states (CLSI/EP24-A2)
- The extent of agreement between the information from the test under evaluation and the diagnostic accuracy criteria (CLSI /EP 12-A2)
GHTF Guidelines

The Global Harmonization Task Force
http://www.imdrf.org/documents/documents.asp

<table>
<thead>
<tr>
<th>DOCUMENT ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GHTF/SG5/N8:2012</td>
<td>Clinical Evidence for IVD Medical Devices - Clinical Performance Studies...</td>
</tr>
<tr>
<td>GHTF/SG5/N5:2012</td>
<td>Reportable Events During Pre-Market Clinical Investigations</td>
</tr>
<tr>
<td>GHTF/SG5/N4:2010</td>
<td>Post-Market Clinical Follow-Up Studies</td>
</tr>
<tr>
<td>GHTF/SG5/N3:2010</td>
<td>Clinical Investigations</td>
</tr>
<tr>
<td>GHTF/N2R8:2007</td>
<td>Clinical Evaluation</td>
</tr>
<tr>
<td>SG5/N1R8:2007</td>
<td>Clinical Evidence – Key Definitions and Concepts</td>
</tr>
</tbody>
</table>
Standards or guidance?

Standard:

A document that provides requirements, specifications, guidelines or characteristics that can be used to ensure that materials, products, processes and services are fit for their designated purpose.

*GHTF definition:* A document, established by consensus and approved by a recognized body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context. (*GHTF/SG1/N044:2008*)
Standards

International Standards Organisation

ISO/TC 212 Clinical laboratory testing and *in vitro* diagnostic test systems

Standardization and guidance in the field of laboratory medicine and in vitro diagnostic test systems. This includes, for example, quality management, pre- and post-analytical procedures, analytical performance, laboratory safety, reference systems and quality assurance.

25 documents
Standards and guidelines

Clinical and Laboratory Standards Institute (CLSI)
National Committee for Clinical Laboratory Standards (NCCLS).

CLSI publishes and sells voluntary consensus standards and guidelines. Some guidelines are specific to a single technology, or a single technical challenge but others are more general.


- Clinical Utility (*screening, diagnostic or confirmatory testing*)
- Device Familiarization and Training
- Evaluation Materials (*Controls and sample collection and handling*)
- Bias and Imprecision
- Comparison of methods (*Test specimens, sample size, duration, monitoring, discrepant results and reference specimen panels*)
- Data analysis, including an appendix on statistical methods
CLSI documents


MM17-A Verification and Validation of Multiplex Nucleic Acid Assays; Approved Guideline (2008)
Guidelines

DEEP  Diagnostic Evaluation Expert Panel

Series of open access articles available from www.nature.com/nrmicro/supplements/index.html
Standards on reporting

The STARD Initiative

Standards for the Reporting of Diagnostic Accuracy Studies

Concern about the quality of diagnostic test evaluations reported in scientific journals led to the STARD Initiative which aims to improve the accuracy and completeness of reporting of studies of diagnostic accuracy, thus allowing readers to assess the potential for bias in the study (internal validity) and to evaluate its generalisability (external validity).

A check list and flow chart have been developed to aid the design and reporting of diagnostic test evaluation studies. Guidance and examples are provided on the STARD website http://www.stard-statement.org/
STARD flow chart
STARD checklist

• State the research questions or **study aims**, e.g. estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.

• Describe the **study population**: inclusion and exclusion criteria, setting and locations where the data were collected.

• Describe participant **recruitment**: Was recruitment based on presenting symptoms, results from previous tests . . .

• Describe participant **sampling**: Was the study population a consecutive series of participants. If not, specify how participants were selected.

• Describe **data collection**: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?
• Describe the **reference standard** and its rationale.

• Describe technical specifications of **material and methods** involved including how and when measurements were taken, and/or cite references for index tests and reference standard.

• Describe definition of and rationale for the **units, cut-offs** and/or categories of the results of the index tests and the reference standard.

• Describe the number, **training and expertise** of the persons executing and reading the index tests and the reference standard.

• Describe whether or not the readers of the index tests and reference standard were **blind (masked)** to the results of the other test and describe any other clinical information available to the readers.
• Describe statistical methods used
• Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals)
• Describe methods for calculating test reproducibility, if done.

• Report when study was done, including beginning and ending dates of recruitment.

• Report clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, co morbidity, current treatments, recruitment centers).
STARD checklist

- Report **time interval** from the index tests to the reference standard, and any **treatment administered** between.

- Report distribution of **severity of disease** (define criteria) in those with the target condition; **other diagnoses** in participants without the target condition.

- Report a cross tabulation of the results of the index tests (including **indeterminate and missing results**) by the results of the reference standard; for continuous results, the **distribution of the test results** by the results of the reference standard.

- Report any **adverse events** from performing the index tests or the reference standard.
• Report estimates of **diagnostic accuracy** and measures of **statistical uncertainty** (e.g. 95% confidence intervals).

• Report how **indeterminate results, missing responses** and **outliers** of the index tests were handled.

• Report estimates of **variability** of diagnostic accuracy **between subgroups** of participants, readers or centers, if done.

• Report estimates of **test reproducibility**, if done.
End of Module 5

Move to the next module

Return to slide 1 of this module

Return to the start of the course

Move to the end of the course
6

Study design and bias
Essential Principles (EP)

Essential Principles of Safety and Performance of Medical Devices (GHTF/SG1/N68:2012)

• A manufacturer of a medical device is expected to design and manufacture a product that is safe and performs as intended.

• This guidance document describes fundamental design and manufacturing requirements, referred to as ‘Essential Principles of Safety and Performance’, to ensure this outcome.

• When followed, EP would provide assurance that the product is safe and performs as expected.
Clinical Performance Studies

Purpose

• To demonstrate, **with objective evidence**, how the IVD will be expected to perform in routine clinical practice
• To measure the ability of the test to detect and predict the condition that is associated with an analyte measurement in the intended target population, performed by the intended users

Key measurements

• Clinical sensitivity
• Clinical specificity
• Positive predictive value
• Negative predictive value
Clinical performance studies should be conducted according to the intended use of the IVD

- Users
- Setting
- Population
- Specimens
- Test version
Clinical Performance Studies

Setting
  • Settings of intended use, performed by intended users

Sample type and size
  • Prospectively collected clinic samples
  • Sample size depends on disease prevalence and expected difference in performance between reference standard test and the test under evaluation

Evaluation outcomes
  • Clinical performance
  • Operational characteristics (e.g. ease of use, robustness)
Study Design

• Study design will depend on:
  • Study objectives
  • Intended use
    • Test purpose (indications for use)
    • Target populations (e.g., age, race, gender, geography, clinical condition)
    • Specimen type
    • Intended users
  • Established analytical characteristics (e.g., precision, interference, measuring range, cutoff)
Study Site Location

• To simulate real world use, best to conduct clinical performance studies at sites other than the manufacturer (for multiple sites, manufacturer may be one)
  • Intended user, intended use environment

• If testing is only done at the manufacturers, then justification is needed

• Procedures for specimen collection and/or testing at the site?
  • Minimise bias
Indications for Use

• When should specimen be collected (prior to treatment or during treatment)?
• Is the study designed to include patient follow-up to determine clinical endpoint or outcome?
  • For tests for predisposition, prognosis, and prediction
Specimen Collection and Handling

• Specimen source
  • Specimens taken from patients with the *intention to use them in a particular clinical performance study*
    • Tested immediately (fresh) or stored (e.g., refrigerated or frozen) for later use
  • *Leftover specimens* collected for routine diagnostic testing that would otherwise be discarded, or specimens collected for research purposes
  • *Archived specimens* that were collected in the past and were stored in repositories

• Issues to consider
  • Specimen integrity (especially for leftover and archived)
  • Measures taken to avoid selection and testing bias
Statistical Design

• Is sampling based on sound statistical principles and methodology?

• Are the statistical methods used for data analysis (confidence interval calculations, significance calculations, etc.) appropriate?

• Appropriate subject and specimen inclusion and exclusion criteria
Statistical Issues

- Criteria for resolution of discrepant results (testing algorithm to establish truth)
- Criteria for data exclusion
- Methods of analysis
- Clinically relevant performance measures
- Minimization of bias
Bias

A flaw in the study design or the method of collecting or interpreting information that can lead to incorrect conclusions about what the study showed.

• Spectrum composition
• Workup bias
• Review bias (blinding)
Spectrum Composition Bias

• What it is:
  • Bias that results from the clinical performance studies not representing the intended use or intended user population

• How to assess studies for spectrum composition bias
  • Should contain information on 3 of the following 4 criteria:
    • Age distribution
    • Sex distribution
    • Summary of presenting clinical symptoms and/or disease stage
    • Eligibility criteria for study subjects and users
  • Pertinent subgroups
    • Sensitivity and specificity may represent average values for a population
### Example of Spectrum Bias

<table>
<thead>
<tr>
<th>Sputum smear microscopy score</th>
<th>TB Reference Laboratory</th>
<th>Rural health centre</th>
</tr>
</thead>
<tbody>
<tr>
<td>3+</td>
<td>• • • • •</td>
<td>•</td>
</tr>
<tr>
<td>2+</td>
<td>• • •</td>
<td>• •</td>
</tr>
<tr>
<td>1+</td>
<td>• • • •</td>
<td>• • •</td>
</tr>
<tr>
<td>Scanty</td>
<td></td>
<td>• •</td>
</tr>
<tr>
<td>Smear Negative</td>
<td></td>
<td>• • • •</td>
</tr>
</tbody>
</table>

- Sample from culture positive TB patient

A new test with a similar sensitivity to smear microscopy might record high sensitivity compared to culture at the referral centre but would perform less well at the rural clinic.
Workup Bias

• What it is:
  • When patients with positive or negative test results are preferentially referred to receive verification of diagnosis by the reference procedure

• How to assess studies for workup bias
  • Ensure that all samples are tested using the reference test
Review Bias (Blinding)

• What it is:
  • Bias introduced by operators who are aware of a prior test result or the clinical status of the individual from whom the test specimen was obtained

• How to assess studies for review bias
  • Were investigational test and reference test interpreted separately by persons unaware of the results of the other?
  • Ensure that only blinded specimens were included in the analysis
Ethical Considerations

• Informed consent
  • For specimens or personal data collected specifically for the study or for specimens/data that can be traced to an individual

• Ethics Committee
  • Required by some jurisdictions to review, approve, and monitor studies to protect human subject rights and welfare

• Communicating test results outside of the study
  • Mechanism to report results to physicians or public health authorities cleared by Ethics Committee?
Study Protocol Contents

• Identifies key factors that may impact completeness and significance of results:
  • Intended participant follow-up
  • Testing algorithms and discrepancy resolution
  • Procedures to mask/blind testers
  • Approaches to statistical analyses
  • Methods for recording endpoints/outcomes
  • Procedures for communication of test results (as appropriate)
Assessing the Protocol

• Did the study design reflect the diversity of real-world use of the product?
  • Intended users
    • Level of training?
  • Intended use setting
    • Laboratory or point-of-care?
    • Under conditions in which test expected to be used?
  • Intended use population
    • Reflects diversity of individuals who would be tested?
  • Intended specimens
    • Specimens tested are those for which a testing claim is sought?
Assessing the Protocol

• How did the study determine truth (what the correct test result should be)?
  • Appropriate comparator test (“gold standard”)?
  • Testing algorithm to establish clinical status of patient

• Were controls in place to minimize bias?
  • Blinding users to previous test results or clinical status of individual being tested
Study Participant Protection

• Was informed consent needed?
• If needed, was the informed consent adequate to inform the study participant of:
  • Nature and objectives of the study
  • What will be done with the specimen
  • Risks to the individual as a result of participating
  • Contact information for questions or concerns
Monitoring

• Who monitored the study?
• Was the monitoring independent of the test developers/manufacturers?
End of Module 6

Move to the next module

Return to slide 1 of this module

Return to the start of the course

Move to the end of the course
7

Reference standards
Establishing Truth

• Purpose of the clinical performance study is to establish the performance characteristics of a new IVD

• There needs to be a way to determine the true status of a sample
Reference Standard Test

• A diagnostic test that is considered the most accurate test for a particular condition.
  • If the reference standard test is positive, it is highly likely that the person has the disease (or analyte).
  • If the reference standard test is negative, it is highly likely that the person does not have the disease (or analyte).
Diagnostic accuracy criteria

• The best currently available criteria for establishing the presence of absence of the condition (CLSI/EP1- A2 )

• May be referred to as the reference standard by regulators.

• Can comprise laboratory testing, imaging and clinical data.
Suitability of Reference Standard

• Does the test have a known and appropriate level of clinical performance with the same specimen types?

• Is the test well-characterized with known state-of-the-art performance characteristics that is manufactured under a stringent quality management system?
The Testing Algorithm

• What is a testing algorithm?
  • A predetermined order of testing to establish the true status (positive or negative, correct quantitation, etc.) of a specimen

• Characteristics of testing algorithms
  • All specimens should be tested using both the test under investigation and the reference standard
  • In the case of a binary result, negative results on both tests may be considered negative, any positive result should be tested further
  • Testing methods should go as far as possible to establish truth

• For assessment: Is the testing algorithm adequate to establish truth?
Example of a Testing Algorithm

Rapid HIV antibody test as an aid in the diagnosis of HIV infection

New Test Result
Reference Test Result

AB-NEGATIVE

CONFIRMATORY TEST (WESTERN BLOT)

AB-POSITIVE

HIV-NEGATIVE

NAT

HIV-POSITIVE
Reference Standard Influences

• In assessment, ensure that the performance of the reference standard be as high as possible

• The “Slippery Slope”: Clinical performance may only be as good as the reference test

  • Negatives on both the test under investigation and the reference standard are not tested further (concordant negatives)
    • Done for practical reasons – most results negative in prospective clinical performance studies
Example of the “Slippery Slope”

- Reference test: 95% sensitive
- Clinical studies show that test under investigation is 95% sensitive compared to the reference test
  - Concordant negatives are not tested further, so false negative on both tests will go undetected
- Actual sensitivity of investigational test would be: $0.95 \times 0.95 = 0.90 = 90\%$
- If investigational test accepted as reference standard for a future study, then number of false negative identified continues to increase rapidly, despite claims of higher sensitivity
End of Module 7

- Move to the next module
- Return to slide 1 of this module
- Return to the start of the course
- Move to the end of the course
8

Precision and accuracy
Precision

• The closeness of agreement between independent test results obtained under stipulated conditions (ISO 3534-1).

• Is expressed quantitatively in terms of imprecision (standard deviation SD or coefficient of variation CV).
Definition: Accuracy

• Closeness of the agreement between the result of a measurement and a true value of the measurand (VIM/JCGM_200_2012).

• Accuracy refers to a combination of trueness and precision. (ISO 3534-2)

• The percentage of correct results obtained by the test under evaluation compared with the results of a reference test (aka the diagnostic accuracy criteria).

• Usually expressed as the number of correct results divided by the total number of results, multiplied by 100.
Diagnostic accuracy

• The ability of a diagnostic test to differentiate between diseased and non-diseased subjects, or between two or more clinical states (CLSI/EP24-A2)

• The extent of agreement between the information from the test under evaluation and the diagnostic accuracy criteria (CLSI /EP 12-A2)
Assessing accuracy

• How much difference is there in the quantitative result produced by the test method compared to the reference standard reference test?
• Where there is a difference, is it consistent?
Figure 9. Forest plots of sensitivity and specificity, all studies, extrapulmonary TB.

<table>
<thead>
<tr>
<th>Study</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alifano 1998a</td>
<td>31</td>
<td>3</td>
<td>11</td>
<td>41</td>
<td>0.74 [0.58, 0.86]</td>
<td>0.93 [0.81, 0.99]</td>
</tr>
<tr>
<td>Alifano 1998b</td>
<td>29</td>
<td>5</td>
<td>13</td>
<td>39</td>
<td>0.69 [0.53, 0.82]</td>
<td>0.89 [0.75, 0.96]</td>
</tr>
<tr>
<td>Banerjee 2003a</td>
<td>13</td>
<td>13</td>
<td>17</td>
<td>19</td>
<td>0.43 [0.25, 0.63]</td>
<td>0.59 [0.41, 0.76]</td>
</tr>
<tr>
<td>Banerjee 2003b</td>
<td>23</td>
<td>4</td>
<td>7</td>
<td>28</td>
<td>0.77 [0.58, 0.90]</td>
<td>0.88 [0.71, 0.96]</td>
</tr>
<tr>
<td>Caminero 1993</td>
<td>16</td>
<td>0</td>
<td>14</td>
<td>48</td>
<td>0.53 [0.34, 0.72]</td>
<td>1.00 [0.93, 1.00]</td>
</tr>
<tr>
<td>Caminero 1994</td>
<td>18</td>
<td>2</td>
<td>38</td>
<td>29</td>
<td>0.32 [0.20, 0.46]</td>
<td>0.94 [0.79, 0.99]</td>
</tr>
<tr>
<td>Chierakul 2001</td>
<td>31</td>
<td>21</td>
<td>36</td>
<td>31</td>
<td>0.46 [0.34, 0.59]</td>
<td>0.60 [0.45, 0.73]</td>
</tr>
<tr>
<td>Gevauaud 1992a</td>
<td>26</td>
<td>47</td>
<td>0</td>
<td>147</td>
<td>1.00 [0.87, 1.00]</td>
<td>0.76 [0.69, 0.82]</td>
</tr>
<tr>
<td>Gevauaud 1992b</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>185</td>
<td>0.23 [0.09, 0.44]</td>
<td>0.95 [0.91, 0.98]</td>
</tr>
<tr>
<td>Gevauaud 1992c</td>
<td>53</td>
<td>47</td>
<td>3</td>
<td>147</td>
<td>0.95 [0.85, 0.99]</td>
<td>0.76 [0.69, 0.82]</td>
</tr>
<tr>
<td>Gevauaud 1992d</td>
<td>18</td>
<td>9</td>
<td>38</td>
<td>185</td>
<td>0.32 [0.20, 0.46]</td>
<td>0.95 [0.91, 0.98]</td>
</tr>
<tr>
<td>Gevauaud 1992e</td>
<td>25</td>
<td>47</td>
<td>0</td>
<td>147</td>
<td>1.00 [0.86, 1.00]</td>
<td>0.76 [0.69, 0.82]</td>
</tr>
<tr>
<td>Gevauaud 1992f</td>
<td>0</td>
<td>9</td>
<td>25</td>
<td>185</td>
<td>0.00 [0.00, 0.14]</td>
<td>0.95 [0.91, 0.98]</td>
</tr>
<tr>
<td>Gevauaud 1992g</td>
<td>34</td>
<td>47</td>
<td>0</td>
<td>147</td>
<td>1.00 [0.90, 1.00]</td>
<td>0.76 [0.69, 0.82]</td>
</tr>
<tr>
<td>Gevauaud 1992h</td>
<td>0</td>
<td>9</td>
<td>34</td>
<td>185</td>
<td>0.00 [0.00, 0.10]</td>
<td>0.95 [0.91, 0.98]</td>
</tr>
<tr>
<td>Kunter 2003a</td>
<td>23</td>
<td>5</td>
<td>65</td>
<td>32</td>
<td>0.26 [0.17, 0.37]</td>
<td>0.86 [0.71, 0.95]</td>
</tr>
<tr>
<td>Kunter 2003b</td>
<td>42</td>
<td>3</td>
<td>46</td>
<td>34</td>
<td>0.48 [0.37, 0.59]</td>
<td>0.92 [0.78, 0.98]</td>
</tr>
<tr>
<td>Kunter 2003c</td>
<td>52</td>
<td>7</td>
<td>36</td>
<td>30</td>
<td>0.59 [0.48, 0.69]</td>
<td>0.81 [0.65, 0.92]</td>
</tr>
<tr>
<td>Luh 1996</td>
<td>28</td>
<td>17</td>
<td>30</td>
<td>207</td>
<td>0.48 [0.35, 0.62]</td>
<td>0.92 [0.88, 0.96]</td>
</tr>
<tr>
<td>McConkey 2002</td>
<td>27</td>
<td>13</td>
<td>29</td>
<td>61</td>
<td>0.48 [0.35, 0.62]</td>
<td>0.82 [0.72, 0.90]</td>
</tr>
<tr>
<td>Nsanze 1997a</td>
<td>18</td>
<td>0</td>
<td>17</td>
<td>35</td>
<td>0.51 [0.34, 0.69]</td>
<td>1.00 [0.90, 1.00]</td>
</tr>
<tr>
<td>Nsanze 1997b</td>
<td>4</td>
<td>0</td>
<td>31</td>
<td>35</td>
<td>0.11 [0.03, 0.27]</td>
<td>1.00 [0.90, 1.00]</td>
</tr>
<tr>
<td>Ratanasuwang 1997a</td>
<td>13</td>
<td>4</td>
<td>27</td>
<td>149</td>
<td>0.33 [0.19, 0.49]</td>
<td>0.97 [0.93, 0.99]</td>
</tr>
<tr>
<td>Ratanasuwang 1997b</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>151</td>
<td>0.80 [0.44, 0.97]</td>
<td>0.97 [0.94, 0.99]</td>
</tr>
<tr>
<td>Senol 2007</td>
<td>4</td>
<td>4</td>
<td>13</td>
<td>56</td>
<td>0.24 [0.07, 0.50]</td>
<td>0.93 [0.84, 0.98]</td>
</tr>
</tbody>
</table>


http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.1001062

Forest plot of Xpert sensitivity and specificity for tuberculosis detection in lymph node samples (tissue or aspirate) with a) culture reference standard and b) composite reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true positive; FP: false positive; FN: false negative; TN: true negative.
End of Module 8

Move to the next module

Return to slide 1 of this module

Return to the start of the course

Move to the end of the course
9

Sample Size
Importance of Sample Size

• How much confidence that the data from the clinical performance studies will be what the user can expect
• The generally accepted way to express data from studies is to identify the actual values (referred to as point estimates) along with 95% confidence intervals
95% Confidence Intervals

• Definition
  • There is a 95% likelihood that the values obtained by users will fall within the statistically determined range

• Calculation

\[ P \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}} \]

\( p \) = sensitivity (or specificity) measured as a proportion
\( n \) = number of samples from infected people
(or, for specificity, from uninfected people).
Confidence Intervals

• Basic concept:
  The larger the number of samples tested (N), the smaller the confidence intervals around the point estimate (i.e., the more confidence there is in the point estimate, and the less uncertainty)
### Sample Size and Confidence Intervals

<table>
<thead>
<tr>
<th>Number of infected (non-infected) subjects required (as defined by ref. test)</th>
<th>Estimated Test Sensitivity (or Specificity)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td>50</td>
<td>+/- 14%</td>
</tr>
<tr>
<td>100</td>
<td>+/- 10%</td>
</tr>
<tr>
<td>150</td>
<td>+/- 8%</td>
</tr>
<tr>
<td>200</td>
<td>+/- 7%</td>
</tr>
<tr>
<td>500</td>
<td>+/- 4%</td>
</tr>
<tr>
<td>1000</td>
<td>+/- 3%</td>
</tr>
</tbody>
</table>

95% conf. interval around the estimated sensitivity
Example

• We estimate that the sensitivity of a new test is 80% and we want the confidence interval to be ± 6%

• We will need to recruit, or have archived specimens from, 170 infected study subjects by the reference test

• If the prevalence of infection in the study population is 10%, then there will be 10 infected subjects per 100 patients seen at the clinic.

• So, to have 170 infected subjects, we will need to recruit 1,700 patients.
Example: Forest plot of 20 reports on sensitivity of microscopy for diagnosing TB, compared to culture

Note how confidence in the accuracy of the result is related to n (sample size)

Lower n = wider CI
Higher n = narrower CI
Factors that Impact Study Design: Indications for Use

DIAGNOSIS or MONITORING (Symptomatic or Known)

SCREENING (Asymptomatic)

DIFFERENT INCLUSION/EXCLUSION CRITERIA

SMALLER sample size

LARGER sample size

Diagnostic tests require high specificity
Screening tests require high sensitivity
Influence of Indications for Use on Sample Size

• A screening test requires high sensitivity, specificity is less critical as positive test results can be confirmed using a more specific test

• Evaluations in low prevalence settings
  • Most people tested will be negative for the analyte
  • Need to screen a large number of individuals to detect a positive
  • Supplement with known positives (but blinded to the user to avoid bias)
Influence of Indications for Use on Sample Size

• Diagnostic indication
  • Testing of individuals with signs and symptoms
  • More limited numbers realistically possible, and larger confidence intervals may be acceptable (per risk analysis)
How to Assess Clinical Performance Study Trial Size

• Look at the lower bound of the 95% confidence intervals derived from the data,
  • This will give you the lowest level of performance expected 95% of the time
  • Is this acceptable in terms of performance?
End of Module 9

Move to the next module

Return to slide 1 of this module

Return to the start of the course

Move to the end of the course
10
Data interpretation
Binary Test Results

• Results are Yes/No, Reactive/Non-reactive, Positive/Negative
  • Enzyme immunoassays: Signal to cutoff ratios
  • Rapid Diagnostics: Presence or absence of a line or spot
  • Early Infant diagnosis test: Above or below a cutoff value
Performance Measures Associated with Binary Test Results

• Sensitivity
• Specificity
• Positive predictive value
• Negative predictive value
• Invalid rate
Distribution of Test Results

A perfect Test:

Most Tests:
Adjusting Thresholds

Specificity=0.90; Sensitivity=0.90

Specificity=0.85; Sensitivity=0.95
Cutoff Values: Additional Comment

• The manufacturer may modify the cutoff after the clinical performance study and re-analyse the data to optimize the sensitivity and specificity

• Before regulatory approval!
How to Review Binary Data

Best presented and summarized in a 2 x 2 table

<table>
<thead>
<tr>
<th>Reference Standard</th>
<th>+ve</th>
<th>-ve</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ve</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>-ve</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

PPV = \( \frac{a}{a+b} \)

NPV = \( \frac{d}{c+d} \)

Sensitivity: \( \frac{a}{a+c} \)

Specificity: \( \frac{d}{b+d} \)
How to Review Binary Data

Best presented and summarized in a 2 x 2 table

<table>
<thead>
<tr>
<th>Reference Standard</th>
<th>+ve</th>
<th>-ve</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ve</td>
<td>74</td>
<td>12</td>
</tr>
<tr>
<td>-ve</td>
<td>7</td>
<td>249</td>
</tr>
</tbody>
</table>

PPV = \( \frac{a}{a+b} \)

NPV = \( \frac{d}{c+d} \)

Sensitivity: \( \frac{a}{a+c} \)

Specificity: \( \frac{d}{b+d} \)
How to Review Binary Data

Best presented and summarized in a 2 x 2 table

<table>
<thead>
<tr>
<th>Index Test</th>
<th>Reference Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+ve</td>
</tr>
<tr>
<td>+ve</td>
<td>74</td>
</tr>
<tr>
<td>-ve</td>
<td>7</td>
</tr>
</tbody>
</table>

Sensitivity: \( \frac{a}{a+c} \)  
Specificity: \( \frac{d}{b+d} \)

- Sensitivity: \( \frac{74}{74+7} = 91.4\% \)
- Specificity: \( \frac{249}{12+249} = 95.4\% \)
<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>91.4%</td>
<td>95.4%</td>
</tr>
</tbody>
</table>

**PPV**  true positive test results/total number of positive test results

**NPV**  true negative test results/total number negative test results

Predictive values vary according to the prevalence of the infection of condition.
Example: Malaria

RDT Malaria rates in Kenyan children (3m-14 yr)

- Highland epidemic: 4.0%
- Lake endemic: 50.4%
- Coast endemic: 6.0%
- Semi-arid and seasonal: 0.7%
- Low risk: 0.6%

2010 Kenya MALARIA Indicator Survey

- A new test has a sensitivity of 99 % and specificity of 98%
- Used in the lakes region 1000 tests will give **499 true positives** and **10 false positives**  \( \text{PPV} = 96\% \)
- Used in Nairobi 1000 tests will give **6 true positives** and **20 false positives**  \( \text{PPV} = 23\% \)
What is a rule in test?

What is a rule out test?
What is a rule in test?  
High specificity

What is a rule out test?  
High sensitivity
Analyzing Data

• Statistical methods

• Set acceptance levels for sensitivity and/or specificity
  • Best to base on lower boundary of the 95% confidence interval to take into account the size of the studies used to generate the data
  • Should be based on a risk analysis to determine if the benefits outweigh the risks
Risk Analysis

• Do the benefits of the test outweigh the risks?

• Inputs into decision-making
  • Number of false negatives vs. number of positive people tested who would not be tested otherwise
  • Number of false positives vs. number of negative people tested who would not be tested otherwise
  • Cost/resource factors that contribute to greater benefit?

• There are risk models that can be applied to assist in decision-making
Methods to Analyze Binary Data: Receiver Operating Characteristic (ROC)

What is it?

- A method to determine how well the test correctly classifies those with and without the analyte (what it detects)
- Plots the true positive rate (sensitivity) as a function of either false positive rate (1 – specificity) or decreasing specificity
Examples of ROC Curves

AUC: 96.9% (95.4%...98.3%)

AUC: 83.2% (76.0%...90.4%)
Receiver Operating Characteristic (ROC)

How to interpret the results

- Calculate area under the curve that is generated
- Results can be interpreted as follows:

<table>
<thead>
<tr>
<th>RANGE OF AREA</th>
<th>GUIDE TO INTERPRETATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90 – 1.00</td>
<td>Excellent</td>
</tr>
<tr>
<td>0.80 – 0.90</td>
<td>Good</td>
</tr>
<tr>
<td>0.70 – 0.80</td>
<td>Fair</td>
</tr>
<tr>
<td>0.60 – 0.70</td>
<td>Poor</td>
</tr>
<tr>
<td>0.50 – 0.60</td>
<td>Fail</td>
</tr>
</tbody>
</table>
Quantitative Test Results

• Numerical readout that is interpreted
  • Viral load assays: Copies of HIV RNA/mL
  • CD4 assays: Quantitation of cells
Accuracy

• How much difference is there in the quantitative result produced by the test method compared to the reference standard reference test?
• Where there is a difference, is it consistent?
Linear regression (a, c, e) and Bland-Altman (b, d, f) analyses of absolute CD4+ T cell counts between the MyT4 and BD FACSCount (a and b); the MyT4 and BD FACSCalibur (c and d); and BD FACSCount and BD FACSCalibur (e and f).

http://www.plosone.org/article/info:doi/10.1371/journal.pone.0107410
Assessing Differences between Quantitative Tests: Linear Regression

What is it?

• A method to compare how similar the results of two tests are to each other
• Plot of test result as a function of the reference test result
• Fit best line to data
Linear Regression

• Results to be interpreted
  • Slope (with confidence intervals) \((m)\)
    • As close to 1.00 as possible
  • Correlation coefficient \((R^2)\)
    • How closely data fall along a straight line
    • As close to 1.00 as possible
  • \(y\)-intercept (with confidence intervals) \((b)\)
Models Based on Linear Regression (account for errors)

- Passing-Bablock
- Deming
Example of a Linear Regression Using Passing-Bablock

- \( m = 0.97 \) (95% CI 0.92-1.01)
- \( R^2 = 0.936 \)
- \( b = -0.05 \) log copies/mL (95% CI -0.22 to 0.14)
A Note on Comparing Quantitative Nucleic Acid Tests

• Readout of nucleic acid tests for viral load is copies of RNA/mL

• “Copy number” may vary significantly from test to test using the same sample

• To compare results in this case, it is necessary to establish the number of copies per unit of RNA when such a standard exists
  • Testing of WHO standards and convert copies to units to normalize the test results
Assessing Differences between Quantitative Tests: Bland-Altman

What is it?

• A plot that assesses the difference between 2 measurements

• Often used to evaluate the agreement among two different instruments or two measurements techniques
Bland-Altman

How to do it

• Plot the the difference between the 2 measurements on a given sample (the test under investigation and the reference test) as a function of the mean of the 2 test results

\[ S(x, y) = \left( \frac{S_1 + S_2}{2}, (S_1 - S_2) \right) \]

• Difference values also have 95% confidence intervals calculated (mean difference +/- 1.96 x Standard Deviation), as well as the statistical significance of that mean bias
Bland-Altman Interpretation

Ask the following questions:

• How big is the average difference between the 2 methods and does this make a difference clinically?

• How wide are the limits of agreement?
  • Wide means the results are ambiguous
  • Narrow means the two methods are essentially equivalent

• Is there a trend in the size of the difference or on which side of the line the difference is? This would show a bias toward one of the tests.

• Is the variability (scatter) consistent across the graph?
Example: Bland-Altman Difference Analysis

From: SAMPLE PRODUCT DOSSIER for WHO Prequalification
How to Review Quantitative Data

• Are the numbers correct?
  • Numbers in summary tables and text
  • Calculations
  • Confidence intervals and results of other statistical methods

• Are all data taken into account?
  • Ensure that there is a justification for any data point not included in calculations
How to Review Quantitative Data

• Are there an excessive number of invalid test results?
  • Source of invalids (test, operator)?

• Are the performance numbers (including confidence intervals, results of statistical analysis, etc.) acceptable

• Decisions can be based on the implications of invalid and incorrect test results
Linear regression analysis (a) and Bland-Altman analysis (b) of absolute CD4+ T lymphocyte counts between FACSCalibur and FACSCount using whole blood.

http://www.plosone.org/article/info:doi/10.1371/journal.pone.0067612
Linear regression analysis and Bland-Altman analysis of absolute CD4+ T lymphocyte counts between PIMA and FACSCalibur using whole blood samples (a & b), PIMA and FACSCount using whole blood samples (c & d), and PIMA (capillary blood samples) and FACSCount (whole blood samples) (e & f) respectively.

http://www.plosone.org/article/info:doi/10.1371/journal.pone.0067612
End of Module 10

Move to the next module

Return to slide 1 of this module

Return to the start of the course

Move to the end of the course
11

Dossier Review Case Studies

1. Risk vs Benefit

2. Reference Standard
Case Study 1: Oral HIV Test

Background

• An HIV infected person does not show any remarkable symptoms in the early stages of infection but is capable of transmitting the virus to others.

• A key component of the prevention and control of HIV is the early detection of infection through a test, and counseling or treatment to interrupt the chain of transmission within the population.

• Simple rapid HIV tests have made testing accessible beyond clinic settings, but at-risk individuals, who are typically drivers of the HIV epidemic, frequently remain stigmatized and marginalized from care.

• In the US, ~ 1.2 million people are infected with HIV, of whom 18% do not know their HIV status and are continuing to transmit their infection to their sexual partners.
### Oral HIV Test Performance

<table>
<thead>
<tr>
<th>Performance Measure</th>
<th>Professional Use OraQuick Test Performance (2-sided 95% CI*)</th>
<th>Minimum Recommended Performance</th>
<th>Over-the-Counter OraQuick Test Performance (2-sided 95% CI*)</th>
<th>Minimum Recommended Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>98% (lower bound of the 2-sided 95% CI)</td>
<td></td>
<td>95% (lower bound of the 2-sided 95% CI)</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>98% (lower bound of the 2-sided 95% CI)</td>
<td></td>
<td>95% (lower bound of the 2-sided 95% CI)</td>
<td></td>
</tr>
</tbody>
</table>

*95% CI = 95% Confidence Interval*
### Oral HIV Test Performance

<table>
<thead>
<tr>
<th>Performance Measure</th>
<th>Professional Use OraQuick Test Performance (2-sided 95% CI*)</th>
<th>Minimum Recommended Performance</th>
<th>Over-the-Counter OraQuick Test Performance (2-sided 95% CI*)</th>
<th>Minimum Recommended Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>99.3% (98.4 - 99.7%)</td>
<td>98% (lower bound of the 2-sided 95% CI)</td>
<td>92.98% (86.64 – 96.92%)</td>
<td>95% (lower bound of the 2-sided 95% CI)</td>
</tr>
<tr>
<td>Specificity</td>
<td>99.8% (99.6 – 99.9%)</td>
<td>98% (lower bound of the 2-sided 95% CI)</td>
<td>99.98% (99.90 – 100%)</td>
<td>95% (lower bound of the 2-sided 95% CI)</td>
</tr>
</tbody>
</table>

*95%CI = 95% Confidence Interval
Group Work

• Do you think that the oral HIV test should be approved and why?

• If yes, what additional safeguards should be built in to minimise risks to individuals and public health?
Case Study 2: Ebola Tests

• Current tests to confirm ebola infection in patients presenting with fever:
  • Nucleic acid amplified tests (NAATs) that amplify and detect pathogen DNA or RNA
  • Lab based procedure that require highly trained lab staff to perform
  • Require 2-2.5 hours to get a result

• Companies are developing NAATs that require less time to get a result and use non-invasive samples
Group Work

• New ebola test:
  • Benefit: a safer and faster test of high accuracy
  • Risks:
    • a false positive result will mean that a patient with fever but not infected with ebola will be sent into an ebola ward where the risk of acquiring infection is increased
    • A false negative result will mean that an ebola infected patient will be released into the community and possibly infect many people

• Given that no test is 100% sensitivity and specificity, is high sensitivity more or less important that high specificity and why?
Reference Standard in Clinical Performance Studies

Case Study 3: Dengue Diagnostic Tests

Rosanna W Peeling
Professor and Chair, Diagnostic Research
London School of Hygiene & Tropical Medicine
Don’t Believe Everything You Read: Dengue IgM Rapid Tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Sens</th>
<th>Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>C</td>
<td>80</td>
<td>&gt;99</td>
</tr>
<tr>
<td>D</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>E</td>
<td>70</td>
<td>100</td>
</tr>
<tr>
<td>F</td>
<td>93</td>
<td>100</td>
</tr>
<tr>
<td>G</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

NS = not stated

<table>
<thead>
<tr>
<th>Test</th>
<th>Sens</th>
<th>Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>23</td>
<td>99</td>
</tr>
<tr>
<td>B</td>
<td>18</td>
<td>98</td>
</tr>
<tr>
<td>C</td>
<td>63</td>
<td>69</td>
</tr>
<tr>
<td>D</td>
<td>9</td>
<td>100</td>
</tr>
<tr>
<td>E</td>
<td>65</td>
<td>98</td>
</tr>
<tr>
<td>F</td>
<td>22</td>
<td>99</td>
</tr>
<tr>
<td>G</td>
<td>6</td>
<td>99</td>
</tr>
</tbody>
</table>

Utility of Different Diagnostic tests in the Detection of Dengue Infection

Days after onset of fever

Acute illness

NS1

viremia

IgM

IgG
Use of Dengue Diagnostic Tests

Case Management:

Aim to confirm clinical diagnosis

- **Confirmed diagnosis:**
  - Virus isolation
  - Nucleic acid detection
  - (Antigen detection)
  - Seroconversion for IgM
  - 4-fold rise in IgG titres

- **Highly suggestive:**
  - IgM positive

Surveillance/monitor impact of interventions:

- IgM positivity
- Virus isolation/nucleic acid detection

Outbreak investigations:

- IgM positivity
- Virus isolation/nucleic acid detection to identify genotype

Vaccine/drug trials:

- Best/most feasible diagnostic methods to define a dengue infected patient (and to identify the genotype)
What should be the reference standard for:

- Dengue IgM test - in Rapid or ELISA format?
- Dengue NS1 test - in Rapid and ELISA format?
End of Module 11

Move to the next module

Return to slide 1 of this module

Return to the start of the course

Move to the end of the course
12
Data from scientific literature and meta analysis
Role of scientific literature

There are two areas where scientific literature is used to assist test validation:

i. Scientific validity of the analyte: does detection/measurement of the analyte truly predict the condition under investigation?

ii. Clinical performance: how well does the test perform when used as intended by the manufacturer?
The need for clinical performance evidence will depend on the test:

- For an established and standardized test, clinical performance will not be required to demonstrate conformity with relevant EPs.
- For an established and non-standardized test, clinical performance will often be required to demonstrate conformity with relevant EPs.
- For a novel test, clinical performance most likely will be required to demonstrate conformity with relevant EPs.
Sources of clinical performance data

• Clinical Performance studies presented by manufacturers

• Published data

• Experience gained from routine diagnostic testing (in other jurisdictions)

GHTF/SG5/N7:2012 Clinical Evidence for IVD medical devices - Scientific Validity Determination and Performance Evaluation
Literature Data

Literature searching may be used to identify published clinical performance data that is not in the possession of the manufacturer that may assist the manufacturer in establishing acceptable clinical performance of an IVD medical device.

**Literature sources used to identify data**

Scientific databases – bibliographic (e.g. MEDLINE, EMBASE, PUBMED, GOOGLE Scholar);

Specialized databases (e.g. MEDION)

Systematic review databases (e.g. Cochrane Collaboration);

Clinical trial registers (e.g. CENTRAL, NIH);

Adverse event report databases (e.g. MAUDE, IRIS)

Reference texts
Literature Data

The data generated through literature searching should relate directly to the IVD medical device in question or earlier versions, with justification as to why the data for the earlier versions are applicable.

The quality of the evidence should be assessed (Use the STARD checklist)
Literature Search Report

Potentially relevant literature identified through search

Literature retrieved for more detailed assessment

Literature excluded, with reasons

Literature with relevant, usable data included in the Scientific Validity Determination and/or Performance Evaluation

Literature excluded from Scientific Validity Determination and/or Performance Evaluation, with reasons
Review considerations

• Should be specific to the IVD in question and reflect intended use.

• The information provided should be of sufficient quality to enable a rational and objective assessment of the clinical performance.

• Attention should be given to repeated publications on the same group of patients by the same authors, in order to avoid over-weighting.
Review considerations

• The different data sets should be reviewed for consistency of results across multiple sites and as appropriate, the intended target populations.

• If different results are observed across the data sets, it will be helpful to determine the reason for such differences.

• All data sets should be included
Meta-Analysis

• Meta-analysis refers to statistical methods for contrasting and combining results from different studies to aggregate data in order to achieve a higher statistical power.

• Meta-analysis is done by identifying a common statistical measure that is shared between studies, and calculating a weighted average of that common measure.

• This weighting is usually related to the sample sizes, although it can also include other factors, such as study quality.
Meta-Analysis

Pitfalls to avoid:

• Combining several poorly designed studies increases sample size but does not improve data quality

• Publication bias: positive findings are more likely to be published than negative findings

• Intentional bias where interested parties exclude studies for reasons other than scientific merit of the data
A forest plot is a means of graphically representing a meta-analysis, a means of displaying data from multiple studies. Confidence intervals are often incorporated.

Forest plot of Xpert MTB/RIF sensitivity and specificity for TB detection in lymph node samples. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true positive; FP: false positive; FN: false negative; TN: true negative.

Denkinger CM et al Eur Respir J. 2014
Methodological quality graph, all studies, pulmonary TB.

http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.1001062
Forest plots of sensitivity and specificity, all studies, extrapulmonary TB.


http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.1001062
Example 2

Head-to-head comparison of accuracy of a rapid point-of-care HIV test with oral versus whole-blood specimens: a systematic review and meta-analysis

Nitika Pant Pai, Bhairavi Balram, Sushmita Shivkumar, Jorge Luis Martinez-Cajas, Christiane Claessens, Gilles Lambert, Rosanna W Peeling, Lawrence Joseph

The Lancet January 24, 2012
DOI:10.1016/S1473-3099(11)70368-1
370 records identified through search of databases

242 after duplicates removed

114 excluded
  4 assessed other rapid test (urine)
  73 not relevant
  35 manufacturer reports
  2 non-English

128 screened

67 excluded for other outcomes (operational, patient centred, and economic)

61 full-text articles assessed for eligibility

16 on false reactive results

45 included

Accuracy synthesis

21 excluded from quantitative synthesis

24 included in quantitative synthesis (meta-analysis)

Positive predictive value analyses

22 excluded from positive predictive value analyses

23 included in positive predictive value analyses

Figure 1: Study selection
Results

Figure 2: HSROC curves for each subgroup. Curves for studies with oral mucosal transudate within-study comparisons (A), finger-stick blood within-study comparisons (B), oral mucosal transudate samples only (C), and finger-stick blood samples only (D). HSROC - hierarchical summary receiver operating characteristic.
End of Module 12

Return to slide 1 of this module

Return to the start of the course

These training materials were produced with support from **Grand Challenges Canada** through a grant for *Improving Regulatory Oversight Of In-Vitro Diagnostics In The Developing World: Affordable access to in-vitro diagnostics through regulatory harmonization approaches* and **UNITAID** through a grant for *A Global Network to Improve Access and Quality of HIV Monitoring Technologies: Better regulation for point-of-care HIV devices*

Please send comments or suggestions to improve these materials to Ruth.Mcnerney@lshtm.ac.uk

© LSHTM The London School of Hygiene & Tropical Medicine will undertake to give non-exclusive permission to reproduce these materials free of charge for personal, academic, educational or non-commercial use subject to appropriate acknowledgement. Any use that is directly intended to generate sales or profit shall require prior permission.