The importance and challenges of achieving high sensitivity and specificity in outcome assessment in randomized controlled trials

CISMAC-CIH webinar
3 October 2018
Lead: Catherine Schwinger, Ranadip Chowhury, and Halvor Sommerfelt
Terminology

- Precision/ reproducibility / reliability
- Accuracy/ validity

Assume that the center is the true value!

Inaccurate

BUT precise!!!

Imprecise
Poor accuracy

Inaccurate
BUT precise!!!
Poor precision

Imprecise
**Examples of measures**

<table>
<thead>
<tr>
<th>Continuous variables</th>
<th>Validity (=accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average bias</td>
</tr>
<tr>
<td></td>
<td>(mean difference between test and gold standard)</td>
</tr>
<tr>
<td>Categorical variables</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sensitivity &amp; Specificity</td>
</tr>
</tbody>
</table>
Definitions

Sensitivity:
- The ability of a test to identify correctly those who have the disease/condition

Specificity
- The ability of a test to identify correctly those who do NOT have the disease/condition
The values of sensitivity and specificity are independent of the prevalence of the disease. But their statistical precision (95% CI) depends on the number of observations.
Importance of sensitivity and specificity in RCTs
Types of misclassification

• Misclassification of outcome
  – Differential (related to exposure)
  – Non-differential (unrelated to exposure)

• Misclassification of exposure
  – Differential (related to outcome)
  – Non-differential (unrelated to outcome)
## Possible/likely misclassification in Trials

<table>
<thead>
<tr>
<th>Study type</th>
<th>Randomized Controlled Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome</strong></td>
<td></td>
</tr>
<tr>
<td>Non-differential</td>
<td>+</td>
</tr>
<tr>
<td>Differential</td>
<td>-(blinded RCT) + (unblinded RCT)</td>
</tr>
<tr>
<td><strong>Exposure</strong></td>
<td></td>
</tr>
<tr>
<td>Non-differential</td>
<td>NA °</td>
</tr>
<tr>
<td>Differential</td>
<td>NA</td>
</tr>
</tbody>
</table>

°NA = Not applicable
Objective: To measure to what extent routine daily zinc (1 RDA) supplementation reduces the incidence of PNEUMONIA

Possible definitions of pneumonia:

• Cough and fast breathing (WHO criteria) for ALRI
  – Easy to use even by field worker
  – High sensitivity: 95% (misses very few real cases of pneumonia)
  – BUT low specificity: 50%

• Crepitations on auscultation by physician
  – Difficult to use in field conditions
  – Low sensitivity (50%)
  – High specificity (95%)
If the diagnostic test was 100% sensitive and 100% specific and the intervention was 50% efficacious

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc</td>
<td>50</td>
<td>950</td>
<td>1000</td>
</tr>
<tr>
<td>Placebo</td>
<td>100</td>
<td>900</td>
<td>1000</td>
</tr>
</tbody>
</table>

$$RR = \frac{50/1000}{100/1000} = 0.5 \ (0.36-0.69)$$
If the diagnostic test was **50% sensitive** and **100% specific** and the intervention was **50% efficacious**

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Zinc</strong></td>
<td>50</td>
<td>950</td>
<td>1000</td>
</tr>
<tr>
<td><strong>Placebo</strong></td>
<td>100</td>
<td>900</td>
<td>1000</td>
</tr>
</tbody>
</table>

\[
RR = ??
\]
If the diagnostic test was 50% sensitive and 100% specific and the intervention was 50% efficacious

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc</td>
<td>25</td>
<td>975</td>
<td>1000</td>
</tr>
<tr>
<td>Placebo</td>
<td>50</td>
<td>950</td>
<td>1000</td>
</tr>
</tbody>
</table>

$$RR = \frac{25/1000}{50/1000} = 0.5 \ (0.31-0.80)$$
If the diagnostic test was 100% sensitive and **95% specific** and the intervention was 50% efficacious

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Zinc</strong></td>
<td>50</td>
<td>950</td>
<td>1000</td>
</tr>
<tr>
<td><strong>Placebo</strong></td>
<td>100</td>
<td>900</td>
<td>1000</td>
</tr>
</tbody>
</table>

RR = ???
If the diagnostic test was 100% sensitive and 95% specific and the intervention was 50% efficacious

<table>
<thead>
<tr>
<th>Pneumonia</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc</td>
<td>98</td>
<td>902</td>
<td>1000</td>
</tr>
<tr>
<td>Placebo</td>
<td>145</td>
<td>855</td>
<td>1000</td>
</tr>
</tbody>
</table>

\[
RR = \frac{98/1000}{145/1000} = 0.68 \quad (0.53-0.86) \\
(0.36-0.69)\]
If the diagnostic test was 50% sensitive and 95% specific and the intervention was 50% efficacious

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc</td>
<td>73</td>
<td>927</td>
<td>1000</td>
</tr>
<tr>
<td>Placebo</td>
<td>95</td>
<td>905</td>
<td>1000</td>
</tr>
</tbody>
</table>

\[
RR = \frac{73/1000}{95/1000} = 0.77 \quad (0.57 - 1.03)
\]
Effects of poor validity

Sensitivity↓ or specificity↓ depending on direction of the systematic measurement error.

Example:
Body temperature measurements (continuous) dichotomized into fever or normal body temperature
Validity and reliability of axillary and tympanic body temperature measurements among infants at Mulago national referral hospital, Uganda

By
Dr. Olive Jackie Namugga
PhD candidate

Supervisors
Dr. Victoria Nankabirwa
Prof. Halvor Sommerfelt
Methods

- Study design
  - cross-sectional study

- Study site
  - Mulago national referral hospital, Uganda

- Study Population.
  - two infant categories;
    - 349 infants aged 0 to 28 days (neonates)
    - 211 young infants aged 29 to 60 days.

- Data collection
  - Four temperature readings are taken sequentially using electronic digital thermometers.
  - For each body site (axilla, and rectum) two - first and second – readings are taken.
  - All the temperature measurements are taken by one trained examiner to reduce examiner variability.
### Effects of poor validity

<table>
<thead>
<tr>
<th>Axillary readings</th>
<th>Rectal (Gold standard)</th>
<th>Normal temperature</th>
<th>Fever</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥37.5°C</td>
<td>≥38.0°C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;37.5°C</td>
<td>36 - 38.0°C</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Rectal temperature:
  - ≥38.0°C: Fever
  - 36 - 38.0°C: Normal temperature

- Axillary temperature:
  - ≥37.5°C: Normal temperature
  - <37.5°C: Normal temperature
Rectal readings:
- ≥38.0°C: 85
- 36 - 38.0°C: 11
- ≥37.5°C: 72
- <37.5°C: 17

Axillary readings:
- ≥37.5°C: 113
- <37.5°C: 4

Systematic negative error (-0.2°C)
- Se = 95.5%
- Sp = 91.1%
- Se = 80.9%
- Sp = 96.0%
### Rectal Readings

<table>
<thead>
<tr>
<th>≥38.0°C</th>
<th>36 - 38.0°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>≥37.5°C</th>
<th>&lt;37.5°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>4</td>
</tr>
</tbody>
</table>

### Axillary Readings

<table>
<thead>
<tr>
<th>≥37.5°C</th>
<th>&lt;37.5°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>113</td>
</tr>
</tbody>
</table>

### Systematic Positive Error (+0.2°C)

#### Rectal Readings

<table>
<thead>
<tr>
<th>≥38.0°C</th>
<th>36 - 38.0°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>≥37.5°C</th>
<th>&lt;37.5°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>1</td>
</tr>
</tbody>
</table>

### Axillary Readings

<table>
<thead>
<tr>
<th>≥37.5°C</th>
<th>&lt;37.5°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>106</td>
</tr>
</tbody>
</table>

#### Se = 95.5%
#### Sp = 91.1%

#### Se = 98.9%
#### Sp = 85.5%
Effects of poor reproducibility

(intra-observer or inter-observer)

\textbf{Sensitivity}↓ & \textbf{Specificity}↓

Example (length dichotomized to stunting):
Measurements close to the cut-off (2 Z-scores) will at random be distributed on the "wrong side", the randomness ensuring that the probability of above-cut-off misclassification is the same as that of below-cut-off misclassification.
Effects of poor reproducibility

< «Truly» stunted

<table>
<thead>
<tr>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

Height-for-age Z-score

-2 SD
There is ALWAYS some error!

- Underestimation of mean
- Inflation of variance
Challenges of achieving high levels of sensitivity and specificity
Research stages

Challenges occur and can be addressed at different stages of the research cycle:

– Design
– Implementation
– Data analysis
Actions at the design stage

• Plan for repeated measurements
  – Length measurement

• Definition of outcome
  – Clinical vs. laboratory
  – History from participant vs. machine
  – Anthropometry vs. clinical observation

• Independent outcome assessment team
Actions during implementation

• Test-retest sessions / duplicate measures by supervisor
• Training refresher
• Calibration of the instruments
  – Weighing machine/Stadiometer
• Quality check of laboratory measurements
Sources of error in measurement

Random and systematic errors may result from:

1. Problems with the instrument quality
2. Inadequate measurement environment; interference with measurement
3. Poor subject preparation / Un-cooperative subject
4. Poor skills of the assessors
Actions during analysis

• Multiple Imputation (MI) Framework
  – Measurement error as partially missing information and completely missing values as an extreme form of measurement error.

• Multiple overimputation (MO)
  – Data values as either observed without error, observed with (conditionally random) error, or missing.
Actions during reporting

- Is there any digit preference?
- Did you do any imputations?

→ Report!