Studies of diagnostic test accuracy

F. Farrokhyyar, MPhil, PhD

Learning objectives

- How to design a diagnostic test accuracy study?
- How to estimate and interpret the properties of diagnostic test?
- What is a ROC curve?
- How to decide on the cut-off levels?

Definition of a diagnostic test accuracy study?

- A diagnostic test accuracy study evaluates the accuracy of index test in correctly distinguishing between those with disease and those without disease compared to gold standard test.
- Gold standard test is also called criterion test or reference test.
- For a diagnostic test to be accurate, it should be both valid (accuracy) and reliable (precision).

To design a diagnostic test accuracy study

- Research question and objectives
- PICO:
  - Study population
  - Properties of the index test
  - Properties of the gold standard test.
  - Outcome (sensitivity, specificity, etc)
- Design
- Methods

Define research question & objectives

Example 1:
To evaluate the diagnostic ability of preop HbA1c plus FPG level in detecting glycometabolic abnormalities in elective CABG surgery patients - compared to the OGTT as the best available gold standard. (Ann Thorac Surg. 2010 May;89(5):1482-7)

Example 2:
To assess the diagnostic accuracy of combined urine dipstick, urine cytology and NMP22 assay in detecting the recurrent non-muscle invasive bladder cancer compared to cystoscopy as gold standard. (local study)

Study population

- Choosing the right population for a diagnostic test study is most important.
- Difficulties may arise:
  - When participants with disease are highly symptomatic (i.e. higher tumor stage).
  - When participants without disease have competing conditions that may cause a positive test.
  - When study participants are different from those to whom the test will be applied (generalizability).
What is the ‘right’ study population?

- Individuals in whom we are uncertain of diagnosis.
- Individuals in whom we will use the test in clinical practice setting to resolve the uncertainty.
- Individuals with the disease who have a wide spectrum of severity.
- Individuals without the disease who have symptoms commonly associated with the disease.

Example 1: all patients undergoing elective CABG surgery
Example 2: all patients with diagnosed bladder cancer.

Index test properties

- Justify the advantages and benefits of its use over the current gold standard test.
- Define if you are assessing the accuracy of a single test or a combination of tests as index.
  - Single test may result in low accuracy
  - Multiple tests – to increase accuracy
- Multiple tests can be applied in parallel or in serial testing.

Examples

Example 1: the indication for having glycometabolic abnormalities is if the results of HbA1c and FPG were positive.

"serial" testing.

Example 2: the indication for having recurrent non-muscle invasive bladder cancer is if the results of standard urine dipstick or urine cytology or NMP22 level were positive.

"parallel" testing

Assumption of independence: the accuracy of the results depends on whether the information contributed by each test is somewhat independent from each other.

Gold standard test properties

- The validity of an index test relies on the accuracy of the gold standard test.
- Paradox may arise if one imperfect test is tested against another imperfect test.
- The new test may seem inferior when in fact it is more accurate.

<table>
<thead>
<tr>
<th>Standard test results</th>
<th>Index test results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True positive</td>
</tr>
<tr>
<td>Negative</td>
<td>False positive</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Standard test results</th>
<th>Index test results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>TP</td>
</tr>
<tr>
<td>Negative</td>
<td>FN</td>
</tr>
</tbody>
</table>

Outcome measure

- Define how the outcomes of index test will be measured:
  - Dichotomous measure – i.e. CT scan, x-ray, MRI that yields positive/negative or normal/abnormal answers.
  - Continuous measure – such as biomarkers, blood pressure.

For continuous measures, we need to establish a cutoff point for a positive test.
### Design

Participants → Randomize → Index test → Positive / Negative

Participants → Index test and gold standard tests → Positive / Negative

<table>
<thead>
<tr>
<th>Standard test results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Index test results</strong></td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>True positive</td>
</tr>
<tr>
<td>False negative</td>
</tr>
</tbody>
</table>

### Methods

**Factors affecting sensitivity and specificity**

- Chance
- Bias due to methods of conducting tests
- Bias due to misanalysis of data
- Bias due to selection of standard test
- Spectrum of disease
- Intrinsic performance of index test

### Bias due to methods of conducting test

- Factors that contribute to the variation in the results:
  - Intrasubject variation
  - Intraobserver variation
  - Interobserver variation

### Creating 2x2 table

<table>
<thead>
<tr>
<th>Standard test results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Index test results</strong></td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

### Example

<table>
<thead>
<tr>
<th>Standard test results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Index test results</strong></td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

### Pre-test probability

- Pre-test probability – is the proportion of individuals with the disease in our study population – **prevalence**.
### Pretest probability or prevalence

<table>
<thead>
<tr>
<th>Index test results</th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>80</td>
<td>FP 30</td>
<td>110</td>
</tr>
<tr>
<td>Negative</td>
<td>TP 20</td>
<td>TN 270</td>
<td>290</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>300</td>
<td>400</td>
</tr>
</tbody>
</table>

Pretest Probability $= \frac{100}{400} = .25 = 25\%$

---

### Sensitivity and Specificity

<table>
<thead>
<tr>
<th>Index test results</th>
<th>Standard test results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive Negative Total</td>
</tr>
<tr>
<td>Positive</td>
<td>80 TP 30</td>
</tr>
<tr>
<td>Negative</td>
<td>20 TN 270</td>
</tr>
<tr>
<td>Total</td>
<td>100 300</td>
</tr>
</tbody>
</table>

Sensitivity $= \frac{TP}{TP + FN} = \frac{80}{100} = .8 = 80\%$

80% of individuals with disease are correctly identified by index test.

Specificity $= \frac{TN}{FP + TN} = \frac{270}{300} = .9 = 90\%$

90% of individuals without disease are correctly identified by index test.

---

### Accuracy

Accuracy $= \frac{80 + 270}{400} = .9 = 90\%$

---

### Posttest probabilities

- **Positive Predictive Value** - is posttest probability of having the disease if the test is positive.
- **Negative Predictive Value** - is posttest probability of NOT having the disease if the test is negative.

### Sensitivity and Specificity

<table>
<thead>
<tr>
<th>Index test results</th>
<th>Standard test results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive Negative Total</td>
</tr>
<tr>
<td>Positive</td>
<td>80 TP 30</td>
</tr>
<tr>
<td>Negative</td>
<td>20 TN 270</td>
</tr>
<tr>
<td>Total</td>
<td>100 300</td>
</tr>
</tbody>
</table>

PPV $= \frac{TP}{TP + FP} = \frac{80}{110} = .73 = 73\%$

73% of individuals with disease have positive test.

NPV $= \frac{TN}{FN + TN} = \frac{270}{290} = .93 = 93\%$

93% of individuals without disease have negative.
### Predictive values & Pretest probability

<table>
<thead>
<tr>
<th>Index test results</th>
<th>Standard test results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>Positive</td>
<td>240</td>
</tr>
<tr>
<td>Negative</td>
<td>60</td>
</tr>
<tr>
<td>Total</td>
<td>300</td>
</tr>
</tbody>
</table>

Pretest P = 300/600 = 50%

SEN = \( \frac{240}{300} = 0.8 = 80\% \)

SPE = \( \frac{270}{300} = 0.9 = 90\% \)

PPV = \( \frac{240}{270} = 0.9 = 90\% \)

NPV = \( \frac{270}{330} = 0.82 = 82\% \)

---

### Therefore...

- Pretest probability of the disease has no effect on the validity of the test (sensitivity & specificity).
- Pretest probability has an effect on the posttest probabilities of a test.
- Predictive values observed in one study do not apply universally and to other clinical settings.
- Therefore, predictive values must be interpreted in the context of pretest probability.

\[
PPV = \frac{sen \times P}{(sen \times P) + (1-spe) \times (1-P)}
\]

\[
NPP = \frac{spe \times (1-p)}{(spe \times (1-p) + (1-sen) \times p)}
\]

---

### Likelihood ratios

- The sensitivity and specificity of a test can be combined into one measure as likelihood ratio.
- Likelihood ratio – how many times more (or less) likely patients with the disease are to have positive test than patients without the disease.
- LR ratio for a positive test is:

\[
LR^+ = \frac{true \ positives}{false \ positives} = \frac{sensitivity}{1-specificity}
\]

- LR ratio for a negative test is:

\[
LR^- = \frac{false \ negatives}{true \ negatives} = \frac{1-sensitivity}{specificity}
\]

---

### Rule of thumb for interpreting LRs

- LR+ of \( \geq 10 \) — large change between pretest and posttest probability of having the disease (*rule in* disease)
- LR+ of \( 5 - 10 \) — moderate change …
- LR+ of <2 — poor change …
- LR- of \( \leq 0.1 \) — large change between pretest and posttest probability of NOT having the disease (*rule out* disease).
- LR- of \( 0.1 - 0.5 \) — moderate change …
- LR- of >0.8 — poor change …

### Samples

#### Study 1

- P = 25%
- SEN = 80%
- SPE = 90%
- PPV = 73%
- NPV = 93%

#### Study 2

- P = 50%
- SEN = 80%
- SPE = 90%
- PPV = 90%
- NPV = 82%
Use of LR for individual patients

- Likelihood ratios can also be used to calculate the probability of disease for individual patients (posttest probability).
- 1- Fagan nomogram is a graphical tool for estimating how much the result on a diagnostic test changes the probability that a patient has a disease.
- 2- Direct mathematical calculation
  
  posttest odds = pretest odds x likelihood ratio
  
  \[ \text{odds} = p / (1-p) \]
  
  \[ P = \frac{\text{odds}}{1+\text{odds}} \]
  
  \[ \text{odds} = 0.25 / 0.75 = 0.33 \]
  
  \[ P = 0.33 / 1.33 = 0.25 \]

How to choose a cut-off point

- We choose a cutoff point based on biological information rather than statistical consideration.
- Example 1, we would want to choose a cutoff point for HbA1c and FGP that is associated with increased risk of glycometabolic abnormalities.
- With any cutoff point we choose, it should be relatively easy to distinguish between extreme values.
- The uncertainty remains about cases that fall close to the cutoff point – called “gray zone”.

Blood glucose level and diabetes

Thus, there is a trade off between sensitivity and specificity.
**Receiver operator characteristic curve**

- ROC curve is a graph for assessing the ability of a test to discriminate between those with disease and those without disease.
- ROC curve is obtained by plotting sensitivity (true positives) against 1 – specificity (false positives) for every cut-off point.
- ROC curve allows a visual analysis of trade off between sensitivity and specificity.
- ROC curve may be used for three purposes:
  - Determination of the cut-off point at which optimal sensitivity and specificity are achieved.
  - Assessment of the diagnostic accuracy of a test.
  - The comparison of two or more diagnostic tests.

**ROC curve – determining cut-off point**

If we choose a lower cut-off point, we increase sensitivity and decrease specificity.

If we choose a higher cut-off point, we decrease sensitivity and increase specificity.