Diagnostic (Classification) Accuracy Studies, part 1

Evidence-Based Practice in Speech-Language Therapy
(SHSC 2033)

Session 8

Thomas Klee & Elizabeth Barrett

香 港 大 學
THE UNIVERSITY OF HONG KONG

Outline

1. Diagnostic accuracy of clinical tests and measures
2. Classification accuracy measures
3. Group discussion

Why do we assess clients?

1. To detect or rule out a condition (classify)
   - Screening
   - Diagnosis
   - Differential diagnosis
2. To track the clinical course of a condition
3. To measure intervention outcome (or progress)

A framework for diagnostic research[1]

Phase I

- Do those with the target disorder have different test results than those without the disorder?
- Results at the group level

Phase II

- Are those with certain test results more likely to have the target disorder than those with other test results?
- Results at the individual level

[1] Sackett and Haynes (2002)

A framework for diagnostic research[2]

Phase III

- Does the test distinguish those with and without the target disorder among those in whom it is clinically reasonable to suspect that the disorder is present?
- Results at the individual level

Phase IV

- Do those who undergo the diagnostic test fare better (in their ultimate health outcomes) than similar people who are not tested?

[2] Sackett and Haynes (2002)

Diagnostic quartet[3]

A valid diagnostic study...

1. Assembles an appropriate spectrum of patients
2. Applies both the **diagnostic test** ("index measure") and the **reference standard** to all of them
3. Interprets each blind to the other
4. Repeats itself in a second, independent ("test") set of patients (replication)

[3] Haynes, Sackett, Guyatt, and Tugwell (2006, p. 275)

Diagnostic accuracy studies compare. . .

Index measure

- The test or measure under investigation

Reference standard

- The way in which the target condition is defined
- Gold standard is a term used when there is widespread agreement on how the reference standard for a condition should be defined (i.e., a definitive diagnosis).

2 x 2 outcome table

	Condition present	Condition absent
Index test +	True positive	False positive
Index test −	False negative	True negative

2 x 2 outcome table

	Condition present	Condition absent
Index test +	a	b
Index test −	c	d

Example: early identification of language delay

- O How accurate is [4]
- I parent-based screening [5]
- P for identifying toddlers in need of further evaluation for suspected language delay
- C compared to the results of a clinical evaluation?

[4] Outcome measure is classification accuracy

[5] Index measure

Study details[6]

- 24-month-olds were screened using two questionnaires sent to their parents ($N = 306$).
  - Language Development Scale (Rescorla, 1998)
  - Written questionnaire asking about concerns and other things.
- Double-blind clinical evaluations were done within 1 month of the screening ($N = 64$).
- Concurrent and predictive validity of the screening approach were examined.

[6] Klee et al. (1998); Klee, Pearce, and Carson (2000)

Study measures

Index test

[$<$ 50 words OR no word combinations by 24 months on the LDS] AND [either parent concern OR $>$ 6 ear infections]

Reference standard

Clinical outcome (language delay, language normal) based on standardised test, play-based language sample and clinical judgement.

Screening outcomes[7]

	Language delay	Language normal	Total
Screen +	10	2	12
Screen −	1	51	52
Total	11	53	64

[7] Klee (2008); Klee et al. (2000)

Classification Accuracy Measures

Sensitivity

How accurately does the screen identify those with the condition?

	Language delay	Language normal	Total
Screen +	10	2	12
Screen −	1	51	52
Total	11	53	64

Sensitivity: $10/11 = .91$, 95% CI [.62, 1.00]

Specificity

How accurately does the screen identify those without the condition?

	Language delay	Language normal	Total
Screen +	10	2	12
Screen −	1	51	52
Total	11	53	64

Specificity: 51/53 = .96, 95% CI [.87, .99]

Positive predictive value

What proportion of positive tests are true positives?

	Language delay	Language normal	Total
Screen +	10	2	12
Screen −	1	51	52
Total	11	53	64

PPV: $10/12 = .83$, 95% CI [.55, .95]

Negative predictive value

What proportion of negative tests are true negatives?

	Language delay	Language normal	Total
Screen +	10	2	12
Screen −	1	51	52
Total	11	53	64

NPV: $51/52 = .98$, 95% CI [.90, 1.00]

Caveats

- All four measures are calculated from a sample of data. But what is important to clinicians is how the new test (or screening measure) will perform in the population.
- Although sensitivity and specificity values should reflect their population values, PPV and NPV will not, since they vary with prevalance.

Likelihood ratios

- Indicate how many times more likely particular test results occur in those with the condition than in those without the condition.
- Can be directly applied to give probabilistic statements concerning the likelihood of the condition in an individual.

Positive likelihood ratio (LR+)

- Likelihood ratio of a positive test result
- Indicates the number of times a positive test is likely to occur in those with the disorder compared to those without
- Proportion of positive screens in children with language delay / Proportion of positive screens in those without language delay
- $\text{LR+} = \textit{Sensitivity}/(1 - \textit{Specificity})$
- $\text{LR+} = .91/(1 - .96) = 24.1$, 95% CI [6.1, 95.0]
- When a child screens positive, he/she is 24 times more likely to have delayed language than not.
- The further LR+ is from 1, the more accurate the classification (diagnostic) ability of the test.

Negative likelihood ratio (LR–)

- Likelihood ratio of a negative test result
- Indicates the number of times a negative test is likely to occur in those with the disorder compared to those without
- Proportion of negative screens in children with language delay / Proportion of negative screens in those without language delay
- LR– $= (1 - \textit{Sensitivity})/\textit{Specificity}$
- LR– $= (1 - .91)/.96 = .09$, 95% CI [.02, .61]
- When a child screens negative, he/she is .09 times as likely to have delayed language as not.
- The further LR– is from 1, the more accurate the classification (diagnostic) ability of the test.

Interpreting LRs[8]

- LRs of > 10 or < 0.1 indicate large and often conclusive changes from pre- to post-test probability.
- LRs of 5 to 10 and 0.1 to 0.2 indicate moderate shifts in probability.
- LRs of 2 to 5 and 0.2 to 0.5 indicate small (but sometimes important) shifts in probability.
- LRs of 1 to 2 and 0.5 to 1 alter probability to a small (and rarely important) degree.

[8] Guyatt, Rennie, Meade, and Cook (2008, p. 208)

Here's the cool part.

- You can convert LRs to post-test probabilities. (woop, woop)
- Tells you the **probability** that your client has the disorder, given your test result.
- In screening, a large post-test probability would indicate the need for further, clinical assessment (e.g., language sample, informal or formal assessment).

Converting LRs to post-test probabilities

- Why? Probabilities are easier to interpret than odds,[9] being on a scale of 0-100.
- Post-test probabilities indicate the chance of having the condition given a positive or negative test result.
- Post-test probabilities can be calculated from various pre-test probabilities (prevalence figures).
- More information about this will be presented next session.

[9] . . . to me anyway!

Summary of measures

- Sensitivity and specificity tell you how accurate the test is in general.
- PPV and NPV tell you what a particular test result means for a particular individual. Be cautious interpreting these, since they change with prevalence.
- The LR nomogram lets you calculate the probability that your client has a disorder given a particular test outcome.
- Be sure to take the 95% CI into account when interpreting any of these measures.

Useful resources[10]

Diagnostic accuracy calculators

`https://www.medcalc.org/calc/diagnostic_test.php`
`https://ebm-tools.knowledgetranslation.net/calculator/diagnostic/`

Reporting standards for authors (STARD 2015)

`http://www.equator-network.org/reporting-guidelines/stard/`

Reporting standards for authors of SRs and MAs of diagnostic accuracy studies (PRISMA-DTA)

`http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.19163`

Critical appraisal checklists for readers (QUADAS-2)

`http://www.bristol.ac.uk/social-community-medicine/projects/quadas/` or `http://www.sign.ac.uk/checklists-and-notes.html`

[10] All links working on 2019-03-19

Group discussion

- Break up into your assigned groups.
- Use CADE (Dollaghan, 2007, p. 155) to critically appraise the research article.
- Document *where* you found information in the research article addressing each point.

References I

Dollaghan, C. A. (2007). *The handbook for evidence-based practice in communication disorders*. Baltimore, MD: Paul H. Brookes Publishing Co.

Guyatt, G., Rennie, D., Meade, M. O., & Cook, D. J. (2008). *Users' guides to the medical literature: essentials of evidence-based clinical practice* (2nd ed.). New York: McGraw Hill.

Haynes, R. B., Sackett, D. L., Guyatt, G. H., & Tugwell, P. (2006). *Clinical epidemiology: how to do clinical practice research* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.

Klee, T. (2008). Considerations for appraising diagnostic studies of communication disorders. *Evidence-Based Communication Assessment and Intervention*, *2*(1), 34–45. doi: 10.1080/17489530801927757

References II

Klee, T., Carson, D. K., Gavin, W. J., Hall, L., Kent, A., & Reece, S. (1998). Concurrent and predictive validity of an early language screening program. *Journal of Speech, Language, and Hearing Research*, *41*, 627–641.

Klee, T., Pearce, K., & Carson, D. K. (2000). Improving the positive predictive value of screening for developmental language disorder. *Journal of Speech, Language, and Hearing Research*, *43*, 821–833.

Sackett, D. L., & Haynes, R. B. (2002). The architecture of diagnostic research. In J. A. Knottnerus (Ed.), *The evidence base of clinical diagnosis* (pp. 19–38). London: BMJ Books.