STUDIES OF DIAGNOSIS AND SCREENING

Chris Dollaghan

CPRI 2012

THE OUTLINE

- Know what your (assessment) problem is
  - Understanding? (e.g., latent structure)
  - Improving clinical outcomes?
    - Better treatment plans
    - Improved classification (category assignment)
- Clearly state question and research phase
- Avoid biases
- Design with effect size and CIs in mind

REASONS FOR CLINICAL ASSESSMENT

- Better treatment planning
- Improved classification people into clinically relevant groups
  - Faster, cheaper, less intrusive, more reliable, more accurate

JUDGING AN ASSESSMENT TOOL

- For planning treatment
  - Compare outcomes for those whose treatment was based on the assessment, with outcomes for those whose treatment was not
  - If the assessment has value for planning treatment, outcomes should be better for those whose treatment was based on its results

JUDGING AN ASSESSMENT TOOL (CONT.)

- For classifying people into clinically relevant groups
  - An effective assessment tool is one that classifies accurately, putting people into the correct group

ASSESSING TO CLASSIFY

- Screening measures
  - No clinical concern; dismiss without further assessment
  - Some clinical concern; further assessment needed
- Diagnostic measures
  - Disorder is present
  - Disorder is absent

EVALUATING CLASSIFICATION MEASURES

- Relatively "early-phase" studies
  - Do cases and controls perform differently on the measure?
    - Standardized group mean comparison (*d*) with associated confidence interval
  - Does the measure correlate with another measure intended to classify?
    - *r*-value with associated confidence interval

EVALUATING CLASSIFICATION MEASURES (CONT.)

- Relatively "later-phase" (accuracy) studies
  - How closely do the classification decisions of the new measure match the classification decisions of a reference ("gold standard") measure, when both are administered to the same people?
  - Accuracy metrics (sensitivity, specificity, likelihood ratios with their associated confidence intervals

AN IMBALANCE

- Group mean comparison studies and correlational studies of measures are MUCH more common than accuracy studies

ACCURACY STUDIES

- Both the new (index) measure and an accepted (reference) measure are given to a group of people, some of whom have the disorder
- Each person's category assignment on the reference measure is taken as his or her "true" status, to which his or her category assignment on the index measure is compared

PREFERRED ACCURACY METRICS

- Positive likelihood ratio (LR+)
  - How likely are those who score in the "disordered" range on the new test to have the disorder?
- Negative likelihood ratio (LR-)
  - How likely are those who score in the "normal" range on the new test to be free of the disorder?

INTERPRETING LIKELIHOOD RATIOS

- Positive likelihood ratio (LR+)
  - ≥ 10: a person with a score of "disordered" is virtually certain to have the disorder
  - 3-9: a person with a score of "disordered" might have the disorder, but other testing is needed to confirm the diagnosis
  - 1: a person with a score of "disordered" is just as likely NOT to have the disorder as to have it (such a score has no diagnostic value)

INTERPRETING LIKELIHOOD RATIOS (CONT.)

- Negative likelihood ratio (LR-)
  - $\leq$ 0.10: a person with a score of "normal" is virtually certain not to have the disorder
  - 0.30-0.90: a person with a score WNL might not have the disorder, but other testing is needed to confirm this
  - 1: a person with a score WNL is just as likely to have the disorder as not to have it (i.e., such a score has no diagnostic value)

WHERE DO GET LIKELIHOOD RATIOS?

- From sensitivity and specificity
- If sensitivity and specificity are reported you can calculate the LRs by hand or (better yet) have an on-line program calculate them, with their CIs
- If sensitivity and specificity are not reported, often you can figure them out from raw data using a simple 2x2 table

“True” status on reference test

Status on new test	Disordered (+)	Not disordered (-)
Disordered	a	b
Not disordered	c	d

Sensitivity = a/(a+c)

Specificity = d/(b+d)

CALCULATIONS

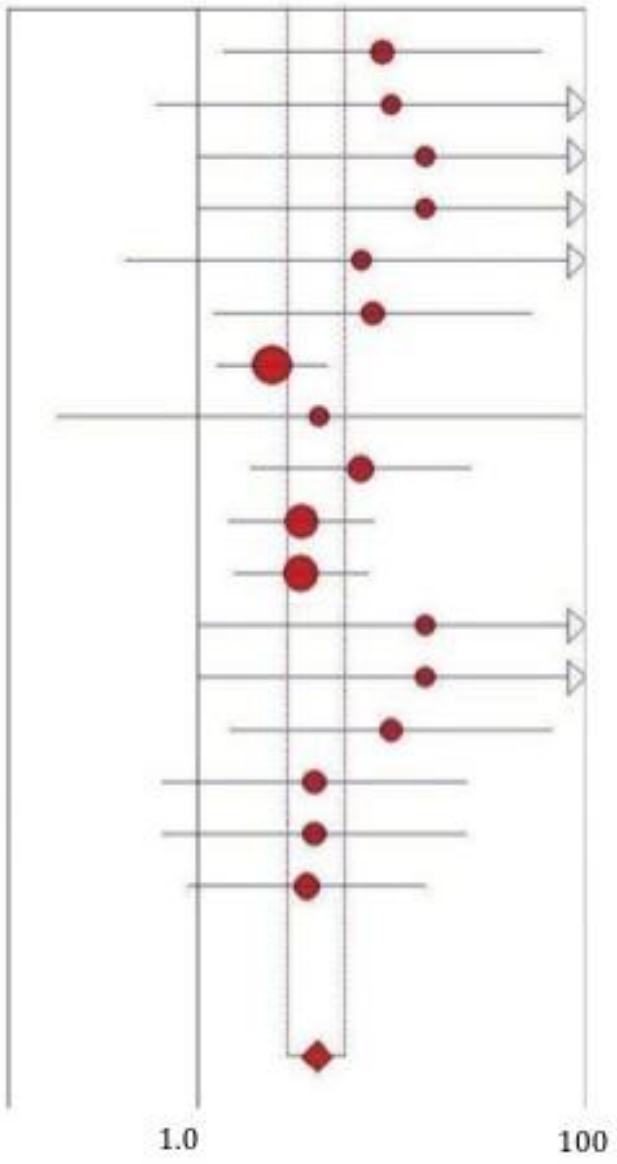
- By hand
  - LR+ = Sensitivity/(1-Specificity)
  - LR- = (1-Sensitivity)/Specificity
- By on-line calculator, with data from 2x2 table
  - http://spph.ubc.ca/sites/healthcare/files/calc/bayes.html (or www.cebm.utoronto.ca)
  - Also provides 95% CIs

WHAT DO CIs ADD TO LRs?

- Ideally, the lower bound of the CI for LR+ would be above 10
- Ideally, the upper bound of the CI for LR- would be below 0.10

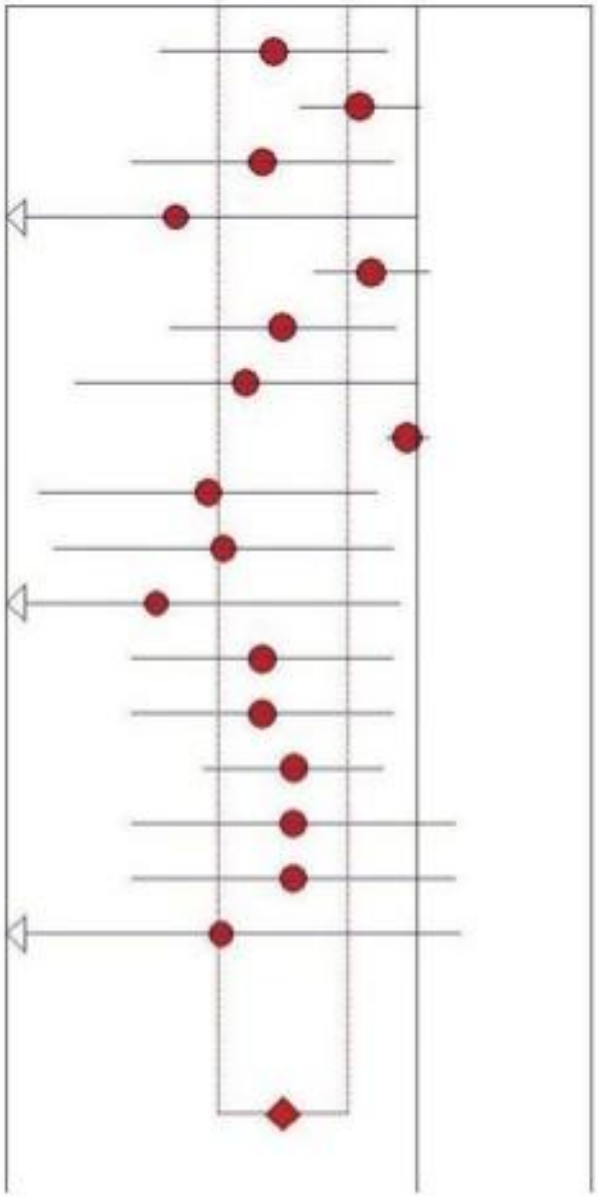
LRs AND CIs CAN SET BENCHMARKS FOR NEW CLASSIFICATION MEASURES

- Inspired by MCID idea that RR described, but applied to studies of diagnostic accuracy
- Forest plots of LR+ and LR- from previous studies to establish the target level a proposed new measure should beat
- e.g., Dollaghan, C. & Horner, E. (2011). Bilingual language assessment: a meta-analysis of diagnostic accuracy. *JSLHR, 54,* 1077-1088.



Measure	LR+	95% CI
Nonword repetition[a]	9.00	1.36, 59.54
Formal word definition[b]	10.00	0.62, 161.11
Spanish Morphosyntax Test – age 4[c]	15.00	1.00, 225.33
Spanish Morphosyntax Test – age 5[c]	15.00	1.02, 220.92
Spanish Morphosyntax Test – age 6[c]	7.00	0.43, 114.71
English Morphosyntax Test[d]	8.00	1.21, 52.69
Finite verb marking[e]	2.42	1.25, 4.69
Obligatory Subject[e]	4.25	0.19, 95.68
Past tense – regular verb[f]	6.88	1.87, 25.26
Past tense – irregular verb[f]	3.44	1.46, 8.09
Past tense – nonce verb[f]	3.42	1.55, 7.54
PR+ETU[g]	15.00	1.00, 225.33
PR+ETU+LTU+FH[g]	15.00	1.00, 225.33
Invented morpheme[h]	10.00	1.49, 67.29
MLU+UNGRAM[i]	4.00	0.66, 24.37
CLITIC+V+ART[i]	4.00	0.66, 24.37
MLU+THEME+DITRAN[i]	3.67	0.90, 14.97
Pooled LR+ (Random effects model)	4.12	2.94, 5.78

Inconsistency (I-square) = 0.0%; Cochrane-Q = 11.97 (df = 16, p = 0.75)



Measure	LR-	95% CI
Nonword repetition[a]	0.20	0.06, 0.71
Formal definition[b]	0.53	0.27, 1.03
Spanish Morphosyntax Test – age 4[c]	0.18	0.04, 0.77
Spanish Morphosyntax Test – age 5[c]	0.07	0.00, 0.98
Spanish Morphosyntax Test – age 6[c]	0.60	0.32, 1.13
English Morphosyntax Test[d]	0.22	0.06, 0.78
Finite verb marking[e]	0.15	0.02, 0.98
Obligatory Subject[e]	0.90	0.72, 1.13
Past tense – regular verb[f]	0.10	0.01, 0.63
Past tense – irregular verb[f]	0.11	0.02, 0.76
Past tense – nonce verb[f]	0.05	0.00, 0.82
PR+ETU[g]	0.18	0.04, 0.77
PR+ETU+LTU+FH[g]	0.18	0.04, 0.77
Invented morpheme[h]	0.25	0.09, 0.68
MLU+UNGRAM[i]	0.25	0.04, 1.52
CLITIC+V+ART[i]	0.25	0.04, 1.52
MLU+THEME+DITRAN[i]	0.11	0.01, 1.64
Pooled LR- (Random effects model)	0.22	0.11, 0.46

Inconsistency (I-square) = 86.7%; Cochrane-Q = 119.93 (df = 16, p = 0.00)

DESIGNING A STRONG STUDY OF A NEW CLASSIFICATION MEASURE

- Accuracy rather than pre-accuracy
  - Category assignments from the reference measure and the index measure for each participant
- Controls for subjective bias
  - Independent, blinded administration of the two measures to each person

DESIGNING A STRONG STUDY (CONT.)

- Avoid biases associated with non-randomized designs
  - Sampling/spectrum bias
    - Use one-gate rather than two-gate (“case-control”) sampling strategy
    - Sample should be representative of the full spectrum of severity encountered in the actual classification task

DESIGNING A STRONG STUDY (CONT.)

- Differential verification bias
  - Ensure that exactly the same reference standard, and identical procedures, are used to establish the trued status of every participant
  - Incorporation bias (a special case of the above)
    - Ensure that results of the index measure play no role in determining true status
    - Index measure derived from discriminant function analysis must be tested in a separate sample

DESIGNING A STRONG STUDY (CONT.)

- Choose a defensible reference standard
  - Need not be a “gold” standard
  - Does need to be plausible
    - e.g., prior enrollment in treatment (some face validity)
  - If nothing else, at least some evidence of inter-examiner reliability
    - e.g., independent, blinded ratings by experts

DESIGNING A STRONG STUDY (CONT.)

- Set a target for LRs and CIs, and power the study accordingly

SOME PRINCIPLES FOR SCIENTISTS

- *You must not fool yourself and you are the easiest person to fool. (Feynman)*
- *Science is what we have learned about how to keep from fooling ourselves. (Feynman)*
- *Less is more, except for sample size. (Cohen)*
- *You can always make a forest plot. (Dollaghan)*

THANKS!

SOME USEFUL REFERENCES

Battaglia M, Bucher H, Egger M, Grossenbacher F, Minder C, &Pewsner D (2002). *The Bayes Library of Diagnostic Studies and Reviews* (2nd edition). See www.ispm.unibe.ch/files/file/261.Bayes_**library**_handbook.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HCW, &Lijmer JG (2003). The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clinical Chemistry, 49:1,* 7-18.

Sackett DL, Straus SE, Richardson WS., Rosenberg W, & Haynes RB. (2000). *Evidence-based medicine: How to practice and teach EBM.* Edinburgh: Churchill Livingstone.

Straus SE, Richardson WS, Glasziou P & Haynes RB (2005). Evidence-based medicine (3rd ed.). Edinburgh, Scotland: Elsevier.

Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy, A (2006). Meta-DiSc: a software for meta-analysis of test accuracy data. *BMC Medical Research Methodology6:*31doi:10.1186/1471-2288-6-31.