STUDIES OF THE ACCURACY OF DIAGNOSTIC TESTS:
(Relevant JAMA Users' Guide Numbers IIIA & B: references (5,6))

Introduction:

The most valid study design for assessing the accuracy of diagnostic tests is a non-experimental cross-sectional study that compares a test's classification of a diagnosis with a reference standard's classification, in a relevant study population.

The conceptual starting point of a diagnostic test study is to apply the reference (or gold) standard to determine which study participants have the disease or condition (D_E) - equivalent to exposed subgroup in other studies described in this module - and which participants don't have it (D_C) - equivalent to the comparison subgroup. In many diagnostic test studies information on test results rather than the reference standard are collected first, however applying the reference standard remains the conceptual starting point.

The outcome of interest in a diagnostic test study is the test result (N). This may initially appear counter-intuitive as the outcome of interest in most studies is the disease. In the simplest example illustrated in the PECOT diagram (page 12), the test result is either positive (N+) or negative (N-). If the test is positive in someone with the condition (i.e. reference standard positive) then we use the symbol $N+_E$; if the test is positive in someone without the condition (i.e. reference standard negative) then we use the symbol $N+_C$. Similarly we can derive test negative categories $N-_E$ and $N-_C$.

The "Outcomes" square in the PECOT diagram (page 12) is equivalent to the 2x2 table often described in texts and studies about diagnostic tests, however we have turned it on its side. For some reason most 2x2 tables have the reference standard results across the top of the table and the test results down the side of the table. We suggest you use our table format because when you draw the PECOT diagram, it is more obvious where the 2x2 table comes from.

The most useful single measure of accuracy of a diagnostic test is the likelihood ratio (LR). The LR is equivalent to a relative risk in other epidemiological studies and is calculated in the same way. However it is possible to calculate LRs for different test result (e.g. for a positive or a negative test result) – see boxes below for definitions.

These numbers can also be used to calculate sensitivity and specificity, which are the more traditionally described characteristics of a diagnostic test study. While they provide useful information (see definitions in boxes below), the LR has the advantage of combining sensitivity and specificity in one number. Moreover, as long as you remember that it is equivalent to a relative risk, it is easy to derive the LR from the PECOT diagram.

If you know the LRs for a test and you have an idea of the average disease prevalence in the group of patients you would apply the test to (known as the pre-test probability), you can also use a simple tool, called a likelihood ratio nomogram (reference 6, page 705 or reference 11, page 79), to estimate the probability that the patient has the disease once you have received the test result (known as the post-test probability of disease).

For those readers who feel more comfortable with sensitivity and specificity, the LR for a positive test is the sensitivity/(1 – specificity) and the LR for a negative test is (1-sensitivity/specificity.

The likelihood ratio for a positive test (LR+ve) is the ratio of: i.) the likelihood of a positive test in people with disease to: ii) the likelihood of a positive test in people without disease.

$$\text{Likelihood Ratio for positive test (LR+ve)} = \frac{\text{number of } N+_E \text{ outcomes / number in } D_E}{\text{number of } N+_C \text{ outcomes / number in } D_C}$$

The likelihood ratio for a negative test (LR-ve) is the ratio of: i.) the likelihood of a negative test in people with disease to: ii) the likelihood of a negative test in people without disease.

$$\text{Likelihood Ratio for negative test (LR-ve)} = \frac{\text{number of } N-_E \text{ outcomes / number in } D_E}{\text{number of } N-_C \text{ outcomes / number in } D_C}$$

The sensitivity of a test is its ability to detect people who have disease; it is the proportion of all people with disease who are identified as positive by the test.

Sensitivity = number of $N+_E$ outcomes / number in D_E

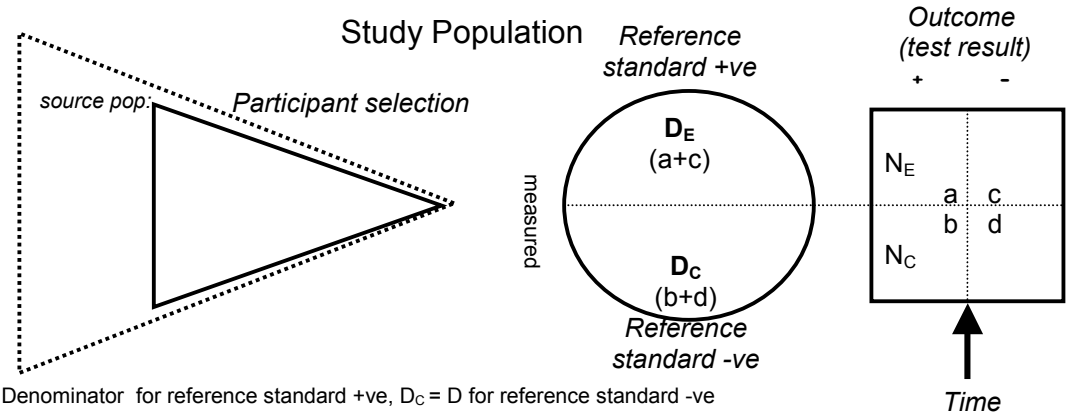
The specificity of a test is its ability to detect people who do not have disease; it is the proportion of all people without disease who are identified as negative by the test.

Specificity = number of $N-_C$ outcomes / number in D_C

The effectiveness of a diagnostic test in reducing the occurrence of a health problem (i.e. the effectiveness of screening with a diagnostic test) is best evaluated in a randomised controlled trial (see appraisal guide for experimental studies).

GATE Checklist for Diagnostic Test Studies (cross-sectional)

Study author, title, publication reference	*Key 5 part study question (PECOT). Was it focussed?*



D_E = Denominator for reference standard +ve, D_C = D for reference standard -ve

N_E = Numerator for reference standard +ve, N_C = N for reference standard -ve

SECTION 1: STUDY VALIDITY		Appraised by:	
Evaluation criterion		*How well was this criterion addressed?*	Quality ✓ ? x
Participants	What were the key selection (inclusion & exclusion) criteria? Were they well defined? Were they replicable?		
	Were selection criteria appropriate given study question?		
	Did selection lead to an appropriate spectrum of participants (like those assessed in practice)		
Exposure/ Comparison	What was the reference standard of diagnosis? Was it clearly defined, independent & valid?		
	Was the reference standard applied regardless of test result?		
	Was the reference standard assessed blind to test result?		
Outcomes	What tests were used? Were they well defined? Replicable?		
	Was the test applied regardless of the reference standard result?		
	Was test assessment blind to reference standard result?		
	Was the test validated in a second, independent group?		
QUALITY OF STUDY DESIGN: *How successfully do you think the study minimised bias? Very well = +, okay = ∅, poorly = -*			

SECTION 2: STUDY RESULTS: ACCURACY & PRECISION	
What measures of test accuracy were reported (sensitivity, specificity, LRs)?	
What measures of precision were reported (CIs, p-values)?	

THE NUMBERS TABLE: LIKELIHOODS, LIKELIHOOD RATIO ESTIMATES & PRECISION

TEST RESULT (N[O])	**IF REFERENCE STANDARD + VE: likelihood of a specific test result (N[O]) = L+ve = $(N[O]_E / D_E)$***	**IF REFERENCE STANDARD - VE: likelihood of a specific test result N[O]) = L–ve = $(N[O]_C / D_C)$***	**LIKELIHOOD RATIO LR = L+ve / L-ve (similar to RR)**	**± 95% CI**
+ve	**= sensitivity (a/a+c)**	**= 1 - specificty (b/b+d)**		
-ve	**= 1 - sensitivity (c/a+c)**	**= specificty (d/b+d)**		
etc				

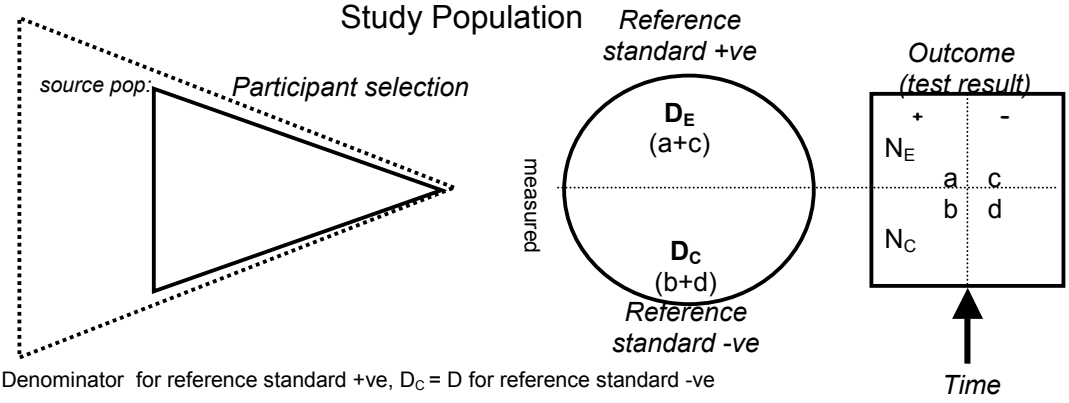
** N[O] represents the generic test result (e.g. +ve, -ve, or a level of a test)*

		Quality ✓ ? x
Could useful measures of test accuracy (i.e.likelihood ratios [LR]) be calculated?		
What was the magnitude of the LR estimates?		
Was the precision of the LR estimates sufficient?		
If no statistically significant associations detected, was there sufficient power?		
QUALITY OF STUDY RESULTS: *Useful, precise +/or sufficient power? Very good = +, okay = ∅, poor = -*		

SECTION 3: STUDY APPLICABILITY			
Participants	Was the source population for participants well described?		
	Were participants representative of source population?		
	Can the relevance of the participants to a specific target group(s) be determined?		
Exposures & Comparison	Were the characteristics of the study setting well described? *e.g. rural, urban, inpatient, primary care*		
	Can sensible estimates of individual patient's pre-test probabilities be determined from the study? (or from elsewhere?)		
Outcomes	Is the test available, affordable and reproducible in the target settings?		
	Will resulting post-test probabilities affect management and help patients? For which target group(s)?		
QUALITY OF STUDY APPLICABILITY: (a) Was it possible to determine applicability? *Very well = +, okay = ∅, poorly = - (b) Are findings applicable in your practice/setting? Very well = +, okay = ∅, poorly = -*			

USERS GUIDE for GATE Checklist for Diagnostic Test Studies

Study author, title, publication reference	*Key 5 part study question (PECOT). Was it focussed?*



D_E = Denominator for reference standard +ve, D_C = D for reference standard -ve

N_E = Numerator for reference standard +ve, N_C = N for reference standard -ve

SECTION 1: STUDY VALIDITY		Appraised by:	
Evaluation criterion		*How well was this criterion addressed?*	Quality ✓ ? x
Participants	What were the key selection (inclusion & exclusion) criteria? Were they well defined? Were they replicable?	List important selection criteria; e.g. age group, gender, risk profile, medical history. Usually in Methods section. There should be sufficient information in the paper (or referenced) to allow the reader to theoretically select a similar population	
	Were selection criteria appropriate given study question?	Are the participants a relevant group to apply the study intervention to? (e.g. diagnostic tests are not very helpful in people with a very high probability of disease).	
	Did selection lead to an appropriate spectrum of participants (like those assessed in practice)	Studies including participants with the range of common presentations of the target disorder and with commonly confused diagnoses are far more informative than studies that only include the extreme ends of the spectrum (florid cases & asymptomatic volunteers only	
Exposure / Comparison	What was the reference standard of diagnosis? Was it clearly defined, independent & valid?	The validity of the study requires that there is an accepted, valid and replicable reference (gold) standard of diagnosis. Readers should give careful and critical consideration to the authors' choice of a reference standard. In addition, those applying and interpreting the reference standard should ideally be unaware of the result of the test to avoid conscious or unconscious bias. This is not always possible, and can lead to over or under-interpretation of the reference standard results.	
	Was the reference standard applied regardless of test result?	Reference standards are often not applied to participants with negative tests, particularly if invasive. An alternative is to follow these participants for an extended period to identify any false negative cases.	
	Was the reference standard assessed blind to test result?	see above, reduces under and over-interpretation of reference standard	
Outcomes	What tests were used? Were they well defined? Replicable?	The methods for undertaking tests should be well described or referenced. It should be theoretically possible for the reader to replicate the process.	
	Was the test applied regardless of the reference standard result?	All participants who are assessed with the reference standard should be tested. Untested participants are equivalent to cases "lost to follow-up"	

	Was test assessment blind to reference standard result?	see above, reduces under and over-interpretation of test	
	Was the test validated in a second, independent group?	As diagnostic tests are predictors, not explainers,of diagnoses, it is possible that the findings in a participant group are related to the characteristics of those selected. Demonstration of test accuracy in a second participant group increases confidence in the findings.	
QUALITY OF STUDY DESIGN: *How successfully do you think the study minimised bias? Very well = +, okay = ∅, poorly = -*			

SECTION 2: STUDY RESULTS: ACCURACY & PRECISION	
What measures of test accuracy were reported (sensitivity, specificity, LRs)?	Some studies do not provide the relevant number of participants (D) in the study population who were assessed using the reference standard, the numbers who were tested (N), the proportions with various test results (N/D) in each reference stand group, or the relevant measures of test accuracy. If they are not reported or cannot be calculated, it is not possible to ascertain the accuracy of the test(s) - see definitions below in the Numbers Table below.
What measures of precision were reported (CIs, p-values)?	Either confidence intervals or p values for sensitivity, specificity & LRs should be reported or be possible to calculate

THE NUMBERS TABLE: LIKELIHOODS, LIKELIHOOD RATIO ESTIMATES & PRECISION

TEST RESULT **(N[O])**	**IF REFERENCE STANDARD + VE: likelihood of a specific test result (N[O]) = L+ve = $(N[O]_E / D_E)$***	**IF REFERENCE STANDARD - VE: likelihood of a specific test result N[O]) = L–ve = $(N[O]_C / D_C)$***	**LIKELIHOOD RATIO LR = L+ve / L-ve** **(similar to RR)**	**± 95% CI**
+ve	**= sensitivity (a/a+c)**	**= 1-specificty (b/b+d)**		
-ve	**=1-sensitivity (c/a+c)**	**= specificity (d(b+d)**		
etc				

** N[O] represents the generic test result (e.g. +ve, -ve, or a level of a test)*

		Quality ✓ ? x
Could useful measures of test accuracy (i.e.likelihood ratios [LR]) be calculated?	LRs should be reported or able to be calculated in the Numbers Table (above). If sensitivity & specificity are reported, it is possible to calculate LRs	
What was the magnitude of the LR estimates?	These numbers are the bottom line of every study. All other appraisal questions relate to the validity, precision and applicability of these numbers. The importance of these numbers in practice depends on the group to which they are applied (see Applicability - next section).	
Was the precision of the LR estimates sufficient?	If 95% confidence intervals are wide and include the no effect point (LR=1) or p-values are >> 0.05, then the precision of the estimates is likely to be poor & insufficient	
If no statistically significant associations detected, was there sufficient power?	If an LR estimate is not significantly different from 1 and the confidence interval is wide, the study is probably not large enough to determine if the test is accurate (i.e. a low power study). A non significant LR associated with a tight CI suggests the test is not useful and that the study has adequate power. Look for a power calculation in the methods section.	
QUALITY OF STUDY RESULTS: *Useful, precise +/or sufficient power? Very good = +, okay = ∅, poor = -*		

SECTION 3: STUDY APPLICABILITY			
Participants	Was the source population for participants well described?	If the source population is not well described it is not easy to assess the generalisability of the study findings to a target group or whether the study participants are a typical or atypical subset of the source population.	
	Were participants representative of source population?	As above	
	Can the relevance of the participants to a specific target group(s) be determined?	As above	
Exposures & Comparison	Were the characteristics of the study setting well described? *e.g. rural, urban, inpatient, primary care*	This helps determine the applicability of the test	
	Can sensible estimates of individual patient's pre-test probabilities be determined from the study? (or from elsewhere?)	The importance of a test depends to a large extent on the pre-test probability of the target condition (i.e. the prevalence of the condition) in the people to whom the test is applied in practice. This information is often difficult to find and readers often depend on the study to determine this.	
Outcomes	Is the test available, affordable and reproducible in the target settings?	The reproducibility of a test may depend on the expertise of those performing and evaluating the test. Information on reproducibility and training in the study setting can help determine reproducibility in other settings.	
	Will resulting post-test probabilities affect management and help patients? For which target group(s)?	The post-test probabilities of the target condition (i.e. the probability of having the target condition if the test is positive or if the test is negative) depends on both the pre-test probability in the whole group tested and the test accuracy (LR). As pre-test probabilities are likely to differ between groups, the usefulness of a test will vary from group to group.	
QUALITY OF STUDY APPLICABILITY: (a) Was it possible to determine applicability? *Very well = +, okay = ∅, poorly = - (b) Are findings applicable in your practice/setting? Very well = +, okay = ∅, poorly = -*			