

Studies of the accuracy of tests to rule in or rule out disease

Criterion	Green	Yellow	Red	Comments
Spectrum of patients enrolled in the study	Study population consists of patients likely to receive the test in clinical practice; the differential diagnosis reasonably includes the target disease, but also includes diseases which may present similarly, from which the target disease needs to be differentiated	Study population consists of patients whose differential diagnosis includes other diseases besides the target disease, but in whom the diagnosis is likely to be already apparent based on already available information	Study population consists of patients who clearly have the target disease based on available information, and patients who are clearly healthy and have a very low likelihood of having the target disease	Diagnostic tests are designed to resolve diagnostic uncertainties; if the positive test subjects have advanced disease, the sensitivity will be biased upwards; if the negative test subjects are clearly healthy, the specificity of the test will be biased upwards; this bias is reduced when consecutive patients who would be candidates for the test are enrolled, and increased when a case-control design is used
Reporting of results	All test results for all patients are reported, including the number of positive, negative, indeterminate, and uninterpretable results	Positive, negative, and indeterminate results are reported, but the number of uninterpretable results is not reported	Only positive and negative results are reported and used for calculation of sensitivity and specificity	The frequency with which the test does not return a definite result is required for estimation of its performance in practice
Reference standard (gold)	There is a recognized gold	There is a recognized gold	There is no gold standard	The readily applicable gold

Criterion	Green	Yellow	Red	Comments
standard)	standard which provides a definitive test of the presence of the disease, and which can be applied to all patients undergoing the diagnostic test being evaluated	standard for the disease, but it is not practical to apply to all patients undergoing the diagnostic test being evaluated	for the disease	standard test may be the exception rather than the rule; if it is an invasive or expensive test, application to all patients in a study may be impractical or unethical. It is acceptable to apply the gold standard to those who test positive, and to follow up those who test negative for subsequent developments, when the gold standard test is not practical
Test thresholds	Clearly defined cutoff points are given which distinguish the difference between a positive and a negative test result; when multiple cutoff points are possible, the sensitivity and specificity are reported for each, and a Receiver Operating Characteristic (ROC) curve is given, with area	Same criteria, but with area under ROC curve of 0.7 to 0.8	Cutoff points are unclear, or area under ROC curve is less than 0.7	This applies only when the test returns a continuous result, and the tradeoff of sensitivity and specificity can be expected to be displayed graphically

Criterion	Green	Yellow	Red	Comments
	under the curve of 0.8 or more			
Blinding of test interpreters	It is clearly stated that the interpreters of the test under evaluation were not aware of the results of the gold standard test, and that the interpreters of the gold standard test were unaware of the results of the test under evaluation	It is clear that the interpreters of the test under evaluation were unaware of the results of the gold standard test, but it is not clear that the interpreters of the gold standard test were unaware of the results of the test under evaluation	Blinding of the interpreters is not clear, or was not done	Large biases are introduced when test interpretation is influenced by knowledge of the results of other tests
Inter-rater reliability	The interpretation of the test is done by two or more assessors working independently, and there is a good agreement between them (Kappa is 0.6 or greater)	The interpretation of the test is done by two or more assessors working independently, and there is a fair agreement between them (Kappa is 0.4 to 0.6)	The interpretation of the test is done by two or more assessors working independently, and there is a slight or poor agreement between them (Kappa is less than 0.4), or there was no report of inter-rater reliability	Kappa may be biased if the prevalence of the disease in the study population is close to zero or is close to 100%; this should not happen if there is an appropriate spectrum of patients in the study sample
Test settings	The test has been applied in a wide variety of settings (primary care, specialty care, tertiary care, high and low prevalence of the disease)	The test has been applied in only a few settings	The test has been applied in only one setting	Test performance may vary with different settings, and a wide variety of settings is necessary for assessing its usefulness in

Criterion	Green	Yellow	Red	Comments
				clinical practice
Test performance measures are presented with measures of uncertainty (e.g., 95% confidence intervals)	Point estimates are given for sensitivity and for specificity, together with 95% confidence intervals for both measures, and are presented for two or more well-described cutoff points	Point estimates are given for sensitivity and for specificity, with confidence intervals, but cutoff points are either lacking or are unclear	Test performance is not clear from the data in the study	Sensitivity and specificity are the core performance measures; predictive values depend on population characteristics and are optionally reported
Likelihood ratios (true positive rate/false positive rate) are likely to produce useful shifts in the estimate of the probability of the presence of the disease	Likelihood ratio is 5 or greater	Likelihood ratio is between 2 and 5	Likelihood ratio is less than 2	Likelihood ratios are measures of how much more probable a positive test is in a person with a disease than in a person without the disease, and are a useful summary measure of the impact of the test result on the odds that a patient has the disease
Characteristics of test interpreters	Test interpreters are well characterized in terms of specialty training, experience, and expertise with executing and reading the test	There is some information about the test interpreters, but they are not fully described in their expertise and training	Information about the test interpreters is vague or missing	Test interpretation may involve subjective judgment, and a learning curve may be involved in reading or executing the test
Benefits of	Test results	Test results	Test results	More than one

Criterion	Green	Yellow	Red	Comments
receiving the test	clearly change patient management in ways that lead to fewer complications, faster recovery, and better final outcomes, due to the making of diagnoses with different treatment strategies	successfully diagnose the target disease, but there is equivocal benefit from the changes in management that result from making the diagnosis	make no difference in management or outcome	type of study may be required to make this determination; a randomized clinical trial is the most robust design to compare outcomes of patients who do and do not have the test
Incremental value of test	The test is clearly shown to have an advantage over simpler or cheaper tests, in having higher likelihood ratios, or in leading to better outcomes for patients who get the test	The test has better diagnostic performance than simpler or cheaper tests, but there is no evidence that doing it leads to better outcomes	The test adds nothing to what is already available for diagnostic investigations	Clinical investigations are expected to result in useful changes in management, not simply additional information