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
Evaluation of test results is done under circumstances which closely resemble the circumstances under which they would be evaluated in everyday practice	The interpreter of the test results has the same kind of information that would be available to a clinician using the test in daily practice (has seen the patient, taken a history, done a physical	The test results are interpreted with only part of the information which would be available to a clinician using the test in daily practice	The test results are interpreted under circumstances which would rarely be seen in practice (interpreter has never seen the patient)	If the test is interpreted under highly artificial circumstances, the study may inaccurately describe how the test will perform in the real world; this is NOT to be confused with having the test results interpreted

Cervival and Low Back 2013

Criterion	Green	Yellow	Red	Comments
	examination, seen the routine laboratory tests, etc)			blinded to the results of the gold standard (see below)
Description of the test	Sufficient information about the test equipment and execution is provided to permit replication of the test	Partial information is given about how the test is executed	Insufficient information about the execution of the test is given	It is important to have enough description of test protocols to allow results to be compared between studies, and to decide whether the test technique being studied is the same as the test being considered for a guideline recommendation; it is acceptable to have technical details furnished in a separate document provided that the reference section point the reader to the source of the details
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 standard)	There is a recognized gold standard which provides a definitive test	There is a recognized gold standard for the disease, but it is not	There is no gold standard for the disease	The readily applicable gold standard test may be the exception rather than the

Cervival and Low Back 2018

Criterion	Green	Yellow	Red	Comments
	of the presence of the disease, and which can be applied to all patients undergoing the diagnostic test being evaluated	practical to apply to all patients undergoing the diagnostic test being evaluated		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
Gold standard applied to all patients who underwent the test being evaluated, or to a random sample of patients	All patients who had the test being evaluated, or a random sample of such patients, also received the test for the gold standard	Some patients who had the test being evaluated did not have the gold standard test, but there is no indication that the performance of the gold standard test was influenced by factors which may predict its result	The gold standard was applied in a manner which is influenced by factors which may be associated with the condition being diagnosed	If the gold standard test is invasive or expensive, it need not be applied to those with a negative result on the test being evaluated; follow-up and continued observation may be substituted
Withdrawals	There is sufficient information to determine whether all patients who entered the study are	Some ambiguity exists concerning what happened to all of the patients who entered the	The patients who participated at the various stages of the study are not reported	It is necessary to know how many patients who received the gold standard also received the test under consideration,

Cervival and Low Back 2013

Criterion	Green	Yellow	Red	Comments
	accounted for, including how many patients participated in each phase of the study (flow diagrams with numbers of patients at each stage of the study are ideal)	study; some patients are not accounted for at the end of the study		and vice versa; if many patients withdrew after participating in only one phase of the study, it is necessary to describe and account for them
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 under the curve of 0.8 or more	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
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,	There is ambiguity about whether the interpreters of one test were aware of the results of the other test; it is clear whether	Blinding of the interpreters is not clear, or was not done; sequence of tests cannot be determined	Large biases are introduced when test interpretation is influence by knowledge of the results of other tests; if tests are strictly

Cervical and Low Back 2013

Criterion	Green	Yellow	Red	Comments
	and that the interpreters of the gold standard test were unaware of the results of the test under evaluation; it is clear which test was applied first	the gold standard or the test under evaluation was applied first		numerical readings of instruments, this criterion is less important
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 clinical practice
Test performance measures are presented with measures of uncertainty (e.g., 95%	Point estimates are given for sensitivity and for specificity, together with 95% confidence	Point estimates are given for sensitivity and for specificity, with confidence intervals, but	Test performance is not clear from the data in the study	Sensitivity and specificity are the core performance measures; predictive values depend on

Cervival and Low Back 2013

Criterion	Green	Yellow	Red	Comments
confidence intervals)	intervals for both measures, and are presented for two or more well-described cutoff points	cutoff points are either lacking or are unclear		population characteristics and are optionally reported
Likelihood ratios (LR+) for a positive test (true positive rate/false positive rate) are likely to produce useful shifts in the estimate of the probability of the presence of the disease, with the potential to alter clinical decisions	LR+ is 10 or greater	LR+ is between 5 and 10	LR+ is less than 5	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
Likelihood ratios (LR-) for a negative test (false negative rate/true negative rate) are likely to produce useful shifts in the estimate of the probability of the presence of the disease	LR- is less than 0.1	LR- is between 0.1 and 0.2	LR- is greater than 0.2	As with LR for positive tests, a low LR- can alter clinical decisions regarding whether to consider a diagnosis improbable enough to look to other diagnoses of the clinical condition
Diagnostic odds ratio (DOR) can be calculated from (LR+/LR-) the likelihood	DOR of greater than 20, preferably even greater	DOR less than 20	DOR less than 20	DOR, unlike positive and negative predictive value, is relatively independent of

Cervival and Low Back 2013

Criterion	Green	Yellow	Red	Comments
ratios positive and negative				prevalence of the disease; it is sensitive to the spectrum of patients enrolled in the study
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 receiving the test	Test results 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	Test results successfully diagnose the target disease, but there is equivocal benefit from the changes in management that result from making the diagnosis	Test results make no difference in management or outcome	More than one 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	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

Cervical and Low Back 2013

Criterion	Green	Yellow	Red	Comments
	who get the test			
Purpose of test	There is a clear description of the setting in which the test is to be used, and the purposes to which it is intended	The setting and purpose are not stated, but may be inferred by the reader	The setting and purpose are not apparent	Sensitivity is crucial for screening tests but not for confirmatory tests; specificity is crucial for confirmatory but not for screening tests

Reference for likelihood ratios and diagnostic odds ratios:

Fisher JE, Bachmann LM, Haesche R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. Intensive Care Med 2003;29:1043 -1051

Cervival and Low Back 2015