RATING QUALITY OF EVIDENCE AND STRENGTH OF RECOMMENDATIONS

GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies

The GRADE system can be used to grade the quality of evidence and strength of recommendations for diagnostic tests or strategies. This article explains how patient-important outcomes are taken into account in this process

In this fourth article of the five part series, we describe how guideline developers are using GRADE to rate the quality of evidence and move from evidence to a recommendation for diagnostic tests and strategies. Although recommendations on diagnosis share the fundamental logic of recommendations for other interventions, they present unique challenges. We will describe why guideline panels should be cautious when they use evidence of the accuracy of tests ("test accuracy") as the basis for recommendations and why evidence of test accuracy often provides low quality evidence for making recommendations.

AHolger J Schünemann professor, Department of Epidemiology, Italian National Cancer Institute Regina Elena, 00144 Rome, Italy and CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada L8N 3Z5

Andrew D Oxman researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, 0130 Oslo, Norway

Jan Brozek research fellow, Department of Epidemiology, Italian National Cancer Institute Regina Elena, 00144 Rome, Italy

Paul Glasziou professor, Centre for Evidence-Based Medicine, Department of Primary Health Care, University of Oxford, Oxford OX3 7LF

Roman Jaeschke clinical professor, Department of Medicine, McMaster University, 1200 Main Street West, Hamilton, Ontario, Canada L8N 3Z5

Gunn E Vist researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, 0130 Oslo, Norway

John W Williams Jr professor, Department of Medicine, Duke University and Durham VA Medical Center, Durham, NC 27705, USA

Regina Kunz associate professor, Basel Institute of Clinical Epidemiology, University Hospital Basel, Hebelstrasse 10, 4031 Basel, Switzerland

Jonathan Craig associate professor, Screening and Test Evaluation Program, School of Public Health, University of Sydney, Department of Nephrology, Children's Hospital at Westmead, Sydney, Australia

Authors continued on next page

This is the fourth in a series of five articles that explain the GRADE system for rating the quality of evidence and strength of recommendations

Testing makes a variety of contributions to patient care

Clinicians use tests—including signs and symptoms, imaging, and biochemistry—to identify physiological derangements, establish prognosis, monitor illness, and diagnose.[1] This article focuses on diagnosis: the use of tests to establish the presence or absence of a disease (such as tuberculosis), target condition (such as iron deficiency), or syndrome (such as Cushing's syndrome).

Clinicians often use diagnostic tests as a package or strategy. For example, in managing patients with apparently operable lung cancer, clinicians may proceed directly to thoracotomy or apply a strategy of imaging the brain, bone, liver, and adrenal glands, with subsequent management depending on the results. Thus, one can often think of evaluating or recommending not a single test, but a diagnostic strategy. Guideline panels considering a diagnostic test or strategy should begin by identifying the patients, diagnostic intervention (strategy), comparison, and outcomes of interest (box).[2 3]

Test accuracy is a surrogate for outcomes important to patients

The main contribution of this article is that it presents a framework for thinking about the quality of evidence for diagnostic tests in terms of their impact on outcomes important to patients ("patient-important outcomes"). Usually, when clinicians think about diagnostic tests, they focus on accuracy (sensitivity and specificity); that is, how well the test classifies patients correctly as having or not having a disease. The underlying assumption is, however, that obtaining a better idea of whether a target condition is present or absent will result in improved outcome. For patients who present with apparently operable lung cancer, the presumption is that additional tests will spare patients the morbidity and early mortality associated with futile thoracotomy. The example of computed tomography for coronary artery disease in the box illustrates another common rationale for a new test: replacement of another test (coronary computed tomography instead of conventional angiography) to avoid complications associated with a more invasive and expensive alternative.[6]

The best way to assess any diagnostic strategy—but in particular new strategies with putative superior accuracy—is a randomised controlled trial in which investigators randomise patients to experimental or control diagnostic approaches and measure mortality, morbidity, symptoms, and quality of life (figure).[7]

When diagnostic intervention studies—ideally randomised controlled trials but also observational studies—comparing the impact of alternative diagnostic

Table 1 | Examples and implications of different testing scenarios

Example of new test and reference test or strategy	Putative benefit of new test	Diagnostic accuracy	
		Sensitivity	Specificity
Shorter version of dementia test compared with original mini mental state exam for diagnosis of dementia	Simpler test, less time	Equal	Equal
Helical computed tomography for renal calculus compared with intravenous pyelogram (IVP)	Detection of more (but smaller) calculi	Greater	Equal
Computed tomography for coronary artery disease compared with coronary angiography	Less invasive testing, but misses some cases	Slightly less	Less

See text for explanations of terms.

Example of a sensible clinical question

In patients in whom coronary artery disease is suspected, does multislice spiral computed tomography of coronary arteries as a replacement for conventional invasive coronary angiography reduce complications with acceptable rates of false negatives associated with coronary events and false positives leading to unnecessary treatment and complications?[4 5]

strategies on patient-important outcomes are available, guideline panels can use the GRADE approach described in previous articles in this series.[12 13] When such studies are not available, guideline panels must focus on studies of test accuracy and make inferences about the likely impact on patient-important outcomes.[14] The key questions are whether a reduction in false negatives (cases missed) or false positives and corresponding increases in true positives and true negatives will occur, how accurately similar or different patients are classified by the alternative testing strategies, and what outcomes occur in both patients labelled as cases and those labelled as not having disease. Table 1 presents examples that illustrate these questions.

Using indirect evidence to make inferences about impact on patient-important outcomes

Inferring from data on accuracy that a diagnostic test or strategy improves patient-important outcomes requires the availability of effective treatment.[1] Alternatively, even without an effective treatment, an accurate test may be beneficial if it reduces test related adverse effects or anxiety, or if confirming a diagnosis improves patients' well-being through the prognostic information it imparts.

For instance, the results of genetic testing for Huntington's chorea, an untreatable condition, may provide either welcome reassurance that a patient will not have the condition or the ability to plan for the future knowing that he or she will develop the condition. The ability to plan is analogous to an effective treatment, and the benefits of planning need to be balanced against the downsides of receiving an early diagnosis.[15-17] We will now describe factors that influence the balance between desirable and undesirable consequences, focusing on the quality of evidence. We will use a simplified approach that classifies test results into yielding true positives, true negatives, false positives, and false negatives.

Victor M Montori associate professor, Knowledge and Encounter Research Unit, Department of Medicine, Mayo Clinic College of Medicine, Rochester, MN 55905, USA

Patrick Bossuyt professor, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam 1100 DE, Netherlands

Gordon H Guyatt professor, CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada L8N 3Z5

For the GRADE Working Group

Correspondence to:
schuneh@mcmaster.ca

Judgment about quality of underlying evidence

Study design and limitations (risk of bias)

GRADE's four categories of quality of evidence represent a gradient of confidence in estimates of the effect of a diagnostic test strategy on patient-important outcomes.[13] Table 2 describes how GRADE deals with the particular challenges of judging the quality of evidence of alternative diagnostic strategies. As we have noted, randomised trials of alternative diagnostic approaches represent the ideal study design for informing recommendations. Nevertheless, in the GRADE system, valid studies of test accuracy also start as high quality in the diagnostic framework. Such studies are, however, vulnerable to limitations and often provide low quality evidence for recommendations as a result of the indirect evidence they usually offer on impact on patient-important outcomes.

Valid studies of diagnostic test accuracy include representative and consecutive patients in whom legitimate diagnostic uncertainty exists—that is, the sort of patients to whom clinicians would apply the test in the course of regular clinical practice. If studies fail this criterion—and, for example, enrol severe cases and healthy

focusing on accuracy

Patients' outcomes and expected impact on management				Balance between presumed outcomes, test complications, and cost
True positives	True negatives	False positives	False negatives	
Presumed influence on patient-important outcomes:				Evidence of shorter time and similar test accuracy (and thus patients' outcomes) would generally support new test's usefulness
Uncertain benefit from earlier diagnosis and treatment	Almost certain benefit from reassurance	Likely anxiety and possible morbidity from additional testing and treatment	Possible detriment from delayed diagnosis	
Directness of evidence (test results) for outcomes important to patients:				
Some uncertainty	No uncertainty	Some uncertainty	Major uncertainty	
Presumed influence on patient-important outcomes:				Fewer complications and downsides compared with IVP would support new test's usefulness, but balance between desirable and undesirable effects is not clear in view of uncertain consequences of identifying smaller stones
Certain benefit for larger stones; less clear benefit for smaller stones, and unnecessary treatment can result	Almost certain benefit from avoiding unnecessary tests	Likely detriment from unnecessary additional invasive tests	Likely detriment for large stones; less certain for small stones, but possible detriment from unnecessary additional invasive tests for other potential causes of complaints	
Directness of evidence (test results) for patient-important outcomes:				
Some uncertainty	No uncertainty	No uncertainty	Major uncertainty	
Presumed influence on patient-important outcomes:				Undesirable consequences of more false positives and false negatives with computed tomography are not acceptable despite higher rate of rare complications (infarction and death) and higher cost of angiography
Benefit from treatment and fewer complications	Benefit from reassurance and fewer complications	Harm from unnecessary treatment	Detriment from delayed diagnosis or myocardial insult	
Directness of evidence (test results) for patient-important outcomes:				
No uncertainty	No uncertainty	No uncertainty	Some uncertainty	

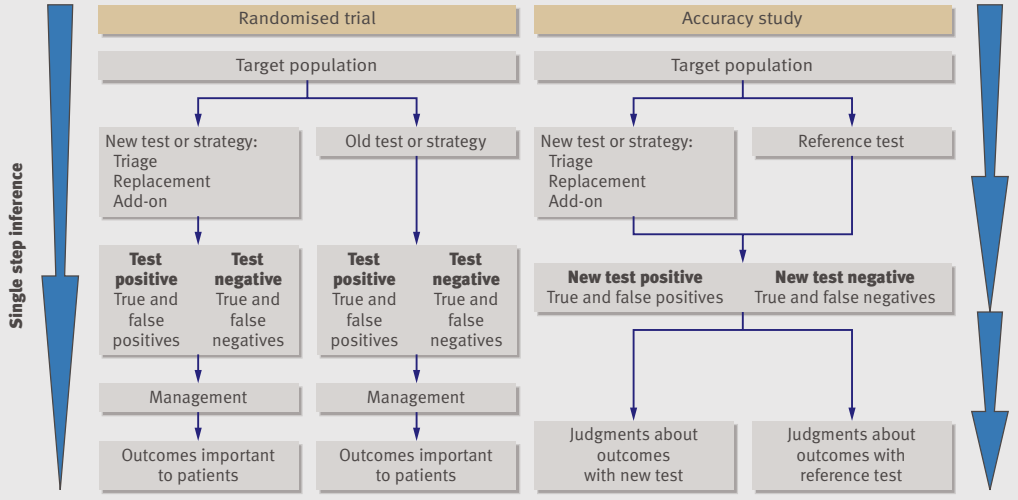
controls—the apparent accuracy of a test is likely to be misleadingly high.[18 19] Valid studies involve a comparison between the test or tests under consideration and an appropriate reference (sometimes called "gold") standard. Investigators' failure to make such a comparison in all patients increases the risk of bias. The risk of bias is further increased if the people who carry out or interpret the test are aware of the results of the reference or gold standard test or vice versa. Guideline panels can use existing instruments to assess the risk of bias in studies evaluating the accuracy of diagnostic tests and can downgrade the quality of evidence if serious limitations exist.[20-22]

Directness

Judging directness presents perhaps the greatest challenges for guideline panels making recommendations about diagnostic tests. For instance, a new test may be simpler to do, with lower risk and cost, but may produce false positives and false negatives. Consider the consequences of replacing invasive angiography with coronary computed tomography scanning for the diagnosis of

Table 2 | Factors that decrease quality of evidence for studies of diagnostic accuracy and how they differ from evidence for other interventions

Factors that determine and can decrease quality of evidence	Explanations and differences from quality of evidence for other interventions
Study design	Different criteria for accuracy studies—Cross sectional or cohort studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard are considered high quality and can move to moderate, low, or very low depending on other factors
Limitations (risk of bias)	Different criteria for accuracy studies—Consecutive patients should be recruited as a single cohort and not classified by disease state, and selection as well as referral process should be clearly described.[7] Tests should be done in all patients in the same patient population for new test and well described reference standard; evaluators should be blind to results of alternative test and reference standard
Indirectness:	
Outcomes	Similar criteria—Panels assessing diagnostic tests often face an absence of direct evidence about impact on patient-important outcomes. They must make deductions from studies of diagnostic tests about the balance between the presumed influences on patient-important outcomes of any differences in true and false positives and true and false negatives in relation to complications and costs of the test. Therefore, accuracy studies typically provide low quality evidence for making recommendations owing to indirectness of the outcomes, similar to surrogate outcomes for treatments
Patient populations, diagnostic test, comparison test, and indirect comparisons	Similar criteria—Quality of evidence can be reduced if important differences exist between populations studied and those for whom recommendation is intended (in previous testing, spectrum of disease or comorbidity); if important differences exist in tests studied and diagnostic expertise of people applying them in studies compared with settings for which recommendations are intended; or if tests being compared are each compared with a reference (gold) standard in different studies and not directly compared in same studies
Important inconsistency in study results	Similar criteria—For accuracy studies, unexplained inconsistency in sensitivity, specificity, or likelihood ratios (rather than relative risk or mean differences) can reduce quality of evidence
Imprecise evidence	Similar criteria—For accuracy studies, wide confidence intervals for estimates of test accuracy or true and false positive and negative rates can reduce quality of evidence
High probability of publication bias	Similar criteria—High risk of publication bias (for example, evidence from small studies for new intervention or test, or asymmetry in funnel plot) can lower quality of evidence



Example
Randomised control trials (RCTs) explored a diagnostic strategy guided by the use of B type natriuretic peptide (BNP)—designed to aid diagnosis of heart failure—compared with no use of BNP in patients presenting to the emergency department with acute dyspnoea.[8 9] As it turned out, the group randomised to receive BNP spent a shorter time in the hospital at lower cost, with no increased mortality or morbidity

Example
Consistent evidence from well designed studies shows fewer false negative results with non-contrast helical computed tomography (CT) than with intravenous pyelography (IVP) in the diagnosis of suspected acute urolithiasis.[10] However, the stones in the ureter that CT detects but IVP "misses" are smaller, and hence are likely to pass more easily. As RCTs evaluating the outcomes in patients treated for smaller stones are not available, the extent to which reduction in cases that are missed (false negatives) and follow-up of incidental findings unrelated to renal calculi with CT have important health benefits remains uncertain[11]

Two generic ways in which a test or diagnostic strategy can be evaluated. On the left, patients are randomised to a new test or strategy or to an old test or strategy. Those with a positive test result (cases detected) are randomised (or were previously randomised) to receive the best available management (second step of randomisation for management not shown). Investigators evaluate and compare patient-important outcomes in all patients in both groups.[6] On the right, patients receive both a new test and a reference test (old or comparator test or strategy). Investigators can then calculate the accuracy of the test compared with the reference test (first step). To make judgments about importance to patients of this information, patients with a positive test (or strategy) in either group are (or have been in previous studies) submitted to treatment or no treatment; investigators then evaluate and compare patient-important outcomes in all patients in both groups (second step)

Table 3 | Key findings of diagnostic accuracy studies—should multislice spiral computed tomography rather than conventional coronary angiography* be used to diagnose coronary artery disease in a population with a low (20%) pre-test probability?[5]

Measure	Test findings (95% CI)
Pooled sensitivity	0.96 (0.94 to 0.98)
Pooled specificity	0.74 (0.065 to 0.84)
Positive likelihood ratio†	5.4 (3.4 to 8.3)
Negative likelihood ratio†	0.05 (0.03 to 0.09)

*Assuming that the reference standard, angiography, does not yield false positives or false negatives.

†Average likelihood ratios from Hamon et al.[5]

coronary artery disease (tables 3 and 4). True positive results will lead to the administration of treatments of known effectiveness (drugs, angioplasty and stents, bypass surgery), and true negative results will spare patients the possible adverse effects of the reference standard test. On the other hand, false positive results will result in adverse effects (unnecessary drugs and interventions, including the possibility of follow-up angioplasty) without apparent benefit, and false negatives will result in patients not receiving the benefits of available interventions that help to reduce the subsequent risk of coronary events.

Thus, it is relatively certain that minimising false positives and false negatives will benefit patients. The impact of inconclusive test results is less clear, but they are clearly undesirable. Furthermore, the complications of invasive angiography—infarction and death—although rare, are undoubtedly important. When guideline panels balance the desirable and undesirable consequences of diagnostics tests, they should consider the importance of these consequences for patients. In this example of patients with a relatively low probability for coronary artery disease, computed tomography scanning results in a large number of false positives leading to unnecessary anxiety and further testing (table 4). It also leads to missing about 1% (false negatives) of patients who have coronary artery disease.

Guideline panels considering questions of diagnosis also face the same sort of challenges regarding indirectness as do panels making recommendations for other interventions.[2] Test accuracy may vary across populations of patients, so panels need to consider how well the populations included in the studies correspond to the population that is the focus of the recommendations. Similarly, panels need to consider how comparable new tests and reference tests are to the tests used in the settings for which the recommendations are made. Finally, when evaluating two or more alternative new tests or strategies, panels need to consider whether these diagnostic strategies were compared directly (in one study) or indirectly (in separate studies) with a common (reference) standard.[25-27]

Arriving at a bottom line for study quality

Table 5 shows the evidence summary and the quality assessment for all critical outcomes of computed tomography angiography as a replacement for invasive angiography. Little or no uncertainty exists about the directness of the evidence (for test results) for patient-important outcomes for true positives, false positives, and true negatives (table 1). However, some uncertainty about the extent to which limitations in test accuracy will have deleterious consequences on patient-important outcomes for false negatives led to downgrading the quality of evidence from high to moderate (table 5 see bmj.com). Unexplained heterogeneity in the results across studies further reduced the quality of evidence for all outcomes. Major uncertainty about the impact of false negative tests on patient-important outcomes would have led to downgrading the quality of evidence from high to low for the other examples in table 1.

Arriving at a recommendation

The balance of presumed patient-important outcomes as the result of true and false positives and negatives with test complications determine whether a guideline panel makes a recommendation for or against applying a test.[12] Other factors influencing the strength of a recommendation include the quality of the evidence, the uncertainty about values and preferences associated with the tests and presumed patient-important outcomes, and cost.

Coronary computed tomography scanning avoids the adverse consequences of invasive angiography, which can include myocardial infarction and death. These consequences are, however, very rare. As a result, a guideline panel evaluating coronary computed tomography as a replacement test for coronary angiography could, despite its lower cost, make a weak recommendation against its use in place of invasive coronary angiography. This recommendation follows from the large number of false positives and the risk of

Table 4 | Consequences of key findings of diagnostic accuracy studies—should multislice spiral computed tomography rather than conventional coronary angiography* be used to diagnose coronary artery disease in a population with a low (20%) pre-test probability?[6]

Consequences	No per 1000 patients	Importance†
True positive results‡	192	8
True negative results§	592	8
False positive results¶	208	7
False negative results**	8	9
Inconclusive results††§§	–	5
Complications‡‡§§	–	5
Cost[§§]	–	5

All results given per 1000 patients tested for prevalence of 20% and likelihood ratios shown in table 3.

*Assuming that the reference standard, angiography, does not yield false positives or false negatives.

†On a 9 point scale, GRADE recommends classifying these outcomes as not important (score 1-3), important (4-6), and critical (7-9) to a decision.[13 18 19]

‡Important because mandates drugs, angioplasty and stents, bypass surgery.

§Important because spares patients unnecessary interventions associated with adverse effects.

¶Important because patients are exposed to unnecessary potential adverse effects from drugs and invasive procedures.

**Important because increase risk of coronary events as a result of patients not receiving efficacious treatment.

††Uninterpretable, indeterminate, or intermediate test results; important because generate anxiety, uncertainty as to how to proceed, further testing, and possible negative consequences of either treating or not treating.

‡‡Not reliably reported; important because although rare, they can be serious.

§§Although the data for these consequences are not reported for simplicity or because they are not exactly known on the basis of the available data, they are important.

SUMMARY POINTS

As for other interventions, the GRADE approach to grading the quality of evidence and strength of recommendations for diagnostic tests or strategies provides a comprehensive and transparent approach for developing recommendations

Cross sectional or cohort studies can provide high quality evidence of test accuracy

However, test accuracy is a surrogate for patient-important outcomes, so such studies often provide low quality evidence for recommendations about diagnostic tests, even when the studies do not have serious limitations

Inferring from data on accuracy that a diagnostic test or strategy improves patient-important outcomes will require the availability of effective treatment, reduction of test related adverse effects or anxiety, or improvement of patients' wellbeing from prognostic information

Judgments are thus needed to assess the directness of test results in relation to consequences of diagnostic recommendations that are important to patients

missing patients with coronary artery disease who could be treated effectively (false negatives). It also follows from the evidence for the new test being only low quality and the consideration of values. Despite the general preference for less invasive tests with lower risks of complications, most patients would probably favour the more invasive approach (angiography), given the risks associated with false positives and negatives.

Conclusion

As for other management recommendations, the GRADE approach to grading the quality of evidence and strength of recommendations for diagnostic tests provides a comprehensive and transparent approach for developing these recommendations. Recognising that test results are surrogates for patient-important outcomes is central to this approach. The application of the approach requires a shift in clinicians' thinking to clearly recognise that, whatever their accuracy, diagnostic tests are of value only if they result in improved outcomes for patients.

We thank the many people and organisations that have contributed to the progress of the GRADE approach through funding of meetings and feedback on the work described in this article.

The members of the Grade Working Group are Phil Alderson, Pablo Alonso-Coello, Jeff Andrews, David Atkins, Hilda Bastian, Hans de Beer, Jan Brozek, Francoise Cluzeau, Jonathan Craig, Ben Djulbegovic, Yngve Falck-Ytter, Beatrice Fervers, Signe Flottorp, Paul Glasziou, Gordon H Guyatt, Robin Harbour, Margaret Haugh, Mark Helfand, Sue Hill, Roman Jaeschke, Katharine Jones, Ilkka Kunnamo, Regina Kunz, Alessandro Liberati, Nicola Magrini, Merce Marzo, James Mason, Jacek Mrukowics, Andrew D Oxman, Susan Norris, Vivian Robinson, Holger J Schünemann, Jane Thomas, Tessa Tan Torres, David Tovey, Peter Tugwell, Mariska Tuut, Helena Varonen, Gunn E Vist, Craig Wittington, John Williams, and James Woodcock.

Contributors: All listed authors, and other members of the GRADE working group, contributed to the development of the ideas in the manuscript, and read and approved the manuscript. HJS wrote the first draft and collated comments from authors and reviewers for subsequent iterations. All other listed authors contributed ideas about structure and content and provided feedback. HJS is the guarantor.

Funding: This work was partially funded by "The human factor, mobility and Marie Curie Actions Scientist Reintegration" European Commission Grant: IGR 42192-"GRADE" to HJS.

Competing interests: The authors are members of the GRADE Working Group. The work with this group probably advanced the careers of some or all of the authors and group members. Authors listed in the byline have received travel reimbursement and honorariums for presentations that included a review of GRADE's approach to grading the quality of evidence and strength of recommendations. GHG acts as a consultant to UpToDate; his work includes helping UpToDate in their use of GRADE. HJS is documents editor and methodologist for the American Thoracic Society; one of his roles in these positions is helping implement the use of GRADE; he supports the implementation of GRADE by organisations worldwide. VMM supports the implementation of GRADE in several North American not for profit professional organisations.

1 Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001;323:157-62.
2 Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ* 1988;138:697-703.
3 Mulrow C, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. *J Gen Intern Med* 1989;4:288-95.
4 Guyatt G, Montori V, Devereaux PJ, Schünemann H, Bhandari M. Patients at the center: in our practice, and in our use of language. *ACP J Club* 2004;140(1):A11-2.
5 Hamon M, Biondi-Zoccai GG, Malagutti P, Agostoni P, Morello R, Valgimigli M, et al. Diagnostic performance of multislice spiral computed tomography of coronary arteries as compared with conventional invasive coronary angiography: a meta-analysis. *J Am Coll Cardiol* 2006;48:1896-910.
6 Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006;332:1089-92.
7 Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000;356:1844-7.
8 Mueller C, Scholer A, Laule-Kilian K, Martina B, Schindler C, Buser P, et al. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. *N Engl J Med* 2004;350:647-54.
9 Moe G, Howlett J, Januzzi JL, Zowall H, Canadian multicenter improved management of patients with congestive heart failure (IMPROVE-CHF) Study Investigators. N-terminal pro-B-type natriuretic peptide testing improves the management of patients with suspected acute heart failure: primary results of the Canadian prospective randomized multicenter IMPROVE-CHF study. *Circulation* 2007;115:3103-10.
10 Worster A, Preyra I, Weaver B, Haines T. The accuracy of noncontrast helical computed tomography versus intravenous pyelography in the diagnosis of suspected acute urolithiasis: a meta-analysis. *Ann Emerg Med* 2002;40:280-6.
11 Worster A, Haines T. Does replacing intravenous pyelography with noncontrast helical computed tomography benefit patients with suspected acute urolithiasis? *Can Assoc Radiol J* 2002;53:144-8.
12 Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, Schünemann HJ. Going from evidence to recommendations. *BMJ* 2008, doi: 10.1136/bmj.39493.646875.AE.
13 Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ. What is "quality of evidence" and why is it important to clinicians? *BMJ* 2008, doi: 10.1136/bmj.39490.551019.BE.
14 Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med* 2006;144:850-5.
15 Maat-Kievit A, Vegter-van der Vlis M, Zoeteweij M, Losekoot M, van Haeringen A, Roos R. Paradox of a better test for Huntington's disease. *J Neurol Neurosurg Psychiatry* 2000;69:579-83.
16 Walker FO. Huntington's disease. *Semin Neurol* 2007;27:143-50.
17 Almqvist EW, Brinkman RR, Wiggins S, Hayden MR. Psychological consequences and predictors of adverse events in the first 5 years after predictive testing for Huntington's disease. *Clin Genet* 2003;64:300-9.
18 Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174:469-76.
19 Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
20 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Intern Med* 2003;138:40-4.
21 Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;3:25.
22 Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol* 2006;6:9.
23 Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1490.
24 Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, et al. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. *Am J Respir Crit Care Med* 2006;174:605-14.
25 Fletcher RH. Carcinoembryonic antigen. *Ann Intern Med* 1986;104:66-73.
26 Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography: multivariable analysis. *Am J Med* 1984;77:64-71.
27 Levy D, Labib SB, Anderson KM, Christiansen JC, Kannel WB, Castelli WP. Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. *Circulation* 1990;81:815-20.