Allergy 2009 DOI: 10.1111/j.1398-9995.2009.02083.x

© 2009 John Wiley & Sons A/S

DOI: 10.1111/j.1398-9995.2009.02083.x

Review article

Grading quality of evidence and strength of recommendations in clinical practice guidelines

Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies

The GRADE approach to grading the quality of evidence and strength of recommendations provides a comprehensive and transparent approach for developing clinical recommendations about using diagnostic tests or diagnostic strategies. Although grading the quality of evidence and strength of recommendations about using tests shares the logic of grading recommendations for treatment, it presents unique challenges. Guideline panels and clinicians should be alert to these special challenges when using the evidence about the accuracy of tests as the basis for clinical decisions. In the GRADE system, valid diagnostic accuracy studies can provide high quality evidence of test accuracy. However, such studies often provide only low quality evidence for the development of recommendations about diagnostic testing, as test accuracy is a surrogate for patient-important outcomes at best. Inferring from data on accuracy that using a test improves outcomes that are important to patients requires availability of an effective treatment, improved patients' wellbeing through prognostic information, or – by excluding an ominous diagnosis – reduction of anxiety and the opportunity for earlier search for an alternative diagnosis for which beneficial treatment can be available. Assessing the directness of evidence supporting the use of a diagnostic test requires judgments about the relationship between test results and patient-important consequences. Well-designed and conducted studies of allergy tests in parallel with efforts to evaluate allergy treatments critically will encourage improved guideline development for allergic diseases.

J. L. Brożek[1,2], E. A. Akl[3], R. Jaeschke[4], D. M. Lang[5], P. Bossuyt[6], P. Glasziou[7], M. Helfand[8], E. Ueffing[9], P. Alonso-Coello[10,11], J. Meerpohl[12,13], B. Phillips[14], A. R. Horvath[15], J. Bousquet[16], G. H. Guyatt[4,17], H. J. Schünemann[3,17] for the GRADE Working Group

[1]Department of Epidemiology, Italian National Cancer Institute Regina Elena, Rome, Italy; [2]Department of Medicine, Jagiellonian University School of Medicine, Krakow, Poland; [3]Department of Medicine and Family Medicine, School of Medicine and Biomedical Sciences, State University of New York at Buffalo, Buffalo, NY, USA; [4]Department of Medicine, McMaster University, Hamilton, ON, Canada; [5]Allergy/Immunology Section, Asthma Center Respiratory Institute, Cleveland Clinic, Cleveland, OH, USA; [6]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands; [7]Centre for Evidence-Based Medicine, Department of Primary Health Care, University of Oxford, Oxford, United Kingdom; [8]Portland Veterans Administration Medical Center and Department of Medicine, Oregon Health & Science University, Portland, OR, USA; [9]Centre for Global Health, Institute of Population Health, University of Ottawa, Ottawa, ON, Canada; [10]Iberoamerican Cochrane Center, Servicio de Epidemiología Clínica y Salud Pública, Hospital de Sant Pau, Barcelona, Spain; [11]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Spain; [12]German Cochrane Centre, Institute of Medical Biometry and Medical Informatics, University Hospital Freiburg, Freiburg, Germany; [13]Division of Pediatric Hematology & Oncology, Department of Pediatrics, University Hospital Freiburg, Freiburg, Germany; [14]Centre for Reviews and Dissemination, University of York, York, United Kingdom; [15]Department of Clinical Chemistry, University of Szeged, Szeged, Hungary; [16]Service des Maladies Respiratoires, Hôpital Arnaud de Villeneuve, INSERM, Montpellier, France; [17]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada

Key words: clinical practice guidelines; diagnosis; evidence based medicine; grading.

Holger J. Schünemann, MD, PhD
Department of Clinical Epidemiology and Biostatistics
Faculty of Health Sciences
McMaster University
1200 Main Street W
Hamilton, ON
Canada

Accepted for publication 10 March 2009

Clinicians use clinical practice guidelines in daily practice. Guidelines make recommendations about appropriate use of therapeutic interventions and about the optimal use of diagnostic tests. Usually, when clinicians face the decision to use or not to use a test, they consider its sensitivity and specificity. They reason that the more sensitive and specific the test, the more likely they are to use it. But... should clinicians still use the test even if there is no effective treatment available? What if performing the test will not induce any change in the management? Would the results of the test then make any difference even if it was very accurate? If recommendations to use diagnostic tests are based on test accuracy alone, are clinicians providing best care or a disservice to their patients when they follow these recommendations? Are the patients better off? Acting in the best interest of the patients, should clinicians follow these recommendations at all?

What do patients want?

The Oxford Dictionary of English defines *diagnosis* as 'the identification of the nature of an illness or other problem by examination of the symptoms' (1). In the medical context, this implies evaluation of patient history, examination and review of laboratory and imaging data. In the Oxford Concise Medical Dictionary, the definition is further amended with a remark: 'unlike therapeutic procedures, diagnostic processes usually do not directly benefit the patient in terms of treatment' (2). This seemingly obvious observation captures the key concept that applying a 'good diagnostic test' by itself does not imply improved patient outcomes.

In a previous article in this series, we emphasized that what patients really want to know are the benefits and downsides of consenting to the diagnostic or therapeutic procedure they have been offered (3). Making an accurate diagnosis, i.e. accurately classifying patients as having or not having a given condition, does not necessarily imply that any of the outcomes important to them will be affected. However, when clinicians think about diagnostic testing, they focus on test performance (accuracy) – how well the test classifies patients correctly as having or not having a target condition. Their underlying assumption is that determining whether a target condition is present or absent will result in superior management of patients and improved outcomes. For instance, in testing for workplace triggers in patients with suspected work-related asthma, the assumption is that negative test results will spare patients the unnecessary removal from the workplace and its consequences or save resources that otherwise would be used to reduce the exposure.

Inferring from the evidence of test accuracy that applying a test improves patient-important outcomes usually requires the availability of an effective treatment (4). Alternatively, even without an effective treatment, an accurate test may be beneficial if it improves patients' wellbeing through prognostic information, spares patients further unnecessary testing, or – by excluding an ominous diagnosi\s – reduces anxiety or enables an earlier search for an alternative condition for which beneficial treatment may be available (5). In the field of allergy and asthma, there are few untreatable conditions, though some may not be curable. Consider the results of testing for peanut allergy. Although there is no treatment available yet, other than avoidance of peanuts (6), the information from diagnostic (cutaneous or *in vitro*) testing is still helpful. Testing may provide reassurance that a cause of prior allergic reaction has been found, while intentional avoidance of peanuts (the 'treatment' or management) will substantially reduce, if not eliminate, risk for future allergic/anaphylactic reactions.

Special challenges when considering the use of medical tests

Clinicians use tests that are usually referred to as 'diagnostic' – including information from patient's history, signs, and symptoms, as well as imaging, biochemical, pathology and psychological tests – with various objec-

Abbreviations: GRADE, Grades of Recommendation, Assessment, Development and Evaluation; HMW, high molecular weight.

© 2009 John Wiley & Sons A/S

tives (7). These objectives include identifying physiological derangements (e.g. increased concentration of immunoglobulin E in blood signifying atopic dermatitis), establishing prognosis, monitoring clinical course of illness or response to treatment, informing clinical management (e.g. identifying a single allergen responsible for symptoms of allergic rhinitis before commencing specific immunotherapy) and diagnosis. In this article, we will focus on using tests to establish a likely diagnosis, i.e. increase probability of presence or absence of a given condition.

Some tests provide only two results (e.g. pregnancy test), other tests report their results as a categorical (e.g. lung imaging) or a continuous variable (e.g. spirometry) with the likelihood of disease increasing or decreasing as the results become more extreme. For simplicity, in this article, we discuss a diagnostic approach that categorizes test results as positive or negative. This common approach always requires defining a more or less arbitrary threshold dichotomizing categorical or continuous test results into positive or negative. This simplified presentation classifies patients into four categories: those with the correct diagnosis – either true positives (TP; people with the target disease and a positive test result) or true negatives (TN; people without the target disease and a negative test result) and those misclassified – either false positives (FP; people without the target disease, but a positive test result) or false negatives (FN; people with the target disease, but a negative test result).

The etymology of the word 'dia-gnosis' [from Greek dia + gignōskein – to 'beknow apart' or distinguish (8)] suggests its naturally complex clinical context, where tests are used to differentiate among a number of conditions, rather than just checking for one (7). Despite this, the lion's share of current research methodology on medical tests relies on estimating the accuracy in predicting the presence of just one target condition.

Clinicians often use tests in combination – as a package or strategy – not in isolation. For example, in patients with symptoms suggesting asthma, clinicians may initially order spirometry with a flow-volume loop. When this test demonstrates airway obstruction, they may order post-bronchodilator spirometry or full body plethysmography. Clinicians can later opt to order bronchoprovocation testing if the results of previous tests were inconclusive, but one still suspects asthma. Furthermore, a testing sequence may use an initial sensitive but nonspecific test, which, only if positive, is followed by a more specific test (e.g. determining the presence of specific IgE with a mixture of allergens followed by testing for single allergens). Thus, one can often think of evaluating or recommending not a single test, but more complex diagnostic strategies.

Some might argue that given the above challenges, problems with linking test accuracy with patient-important outcomes, and limitations in current knowledge about diagnostic test research, applying a system of grading quality of evidence and strength of recommendations about using medical tests is premature. One might claim that we should concentrate on generating proper evidence, developing adequate research frameworks and improving statistical techniques instead. However, clinicians cannot stop performing diagnostic tests while waiting for researchers, methodologists and statisticians to develop better methods of assessing the usefulness of the tests they use. Clinicians need guidance on the best application of diagnostic methods now and the guideline panels need to provide this advice. Consequently, clinicians, guideline panels and decision makers need clear direction on how to approach this complex problem.

The GRADE approach

Ask specific clinical questions

When making recommendations about using new medical tests, guideline panels should begin by defining the current diagnostic pathway (currently used test, a sequence of tests, or no testing) and identifying its limitations. Identifying the limitations of the present diagnostic pathway helps determine alternative tests that could offer a remedy (e.g. eliminate a high proportion of FP or false negative results, enhance availability, reduce invasiveness or decrease costs). This initial assessment of status quo and its limitations aids to clarify the role of a new test in the context of current tests. A new test can play one of three roles: act as a triage (to minimize use of invasive or expensive tests), replace a current test (to replace tests with greater burden, invasiveness, or cost), or add-on (to enhance accuracy of a diagnosis beyond existing tests) (9). This process leads to formulation of a sensible clinical question (i.e. 'PICO'-question) that, as with other management problems, specifies the relevant population or patients (P) for whom the diagnostic testing is being considered, the diagnostic strategy or intervention (I), the comparison strategy (C) and the patient-important outcomes (O) related to the use of a test (10, 11).

Consider the following example of a question about a replacement test: in patients with asthma suspected to be related to exposure to high molecular weight (HMW) agents in the workplace (P), do specific skin prick tests (I) as a replacement for specific inhalation challenge tests (C) reduce complications of testing itself with an acceptable rate of false negative results (implying continued patient exposure to an offending agent) and FP results (leading to unnecessary removal from the workplace and its consequences, needless use of personal protection equipment, and futile resource use for reducing the exposure) (O)?

Identify the evidence for every clinical question

As for therapeutic interventions, ideally, every diagnostic clinical question would be answered based on a systematic

© 2009 John Wiley & Sons A/S

review of the relevant evidence. The results of the review serve to create a summary of evidence that a guideline panel can use to inform their judgements (12).

Guideline panels should make recommendations based on the outcomes that are important to patients. As we mentioned earlier, there seems to be no reason to recommend and use even the most accurate test, if it fails to improve important outcomes (e.g. mortality, morbidity, symptoms, complications of test procedure, resource utilization and quality of life). Thus, the most appropriate way to assess the usefulness of a test or test strategy is a randomized trial in which patients receive either the experimental or the control diagnostic approach and patient-important outcomes are measured (Fig. 1, left) (4).

When making recommendations about the use of a particular test, guideline panels should first look for diagnostic intervention studies (ideally randomized trials but also observational management studies) with direct assessment of patient-important outcomes. If such studies are available, guideline panels can use the GRADE approach as described for other interventions (3, 13). If, however, such studies are not available, guideline panels must rely on the information from studies of test accuracy and from this surrogate outcome make inferences about the presumed impact on patient-important outcomes (Fig. 1, right). Evidence from accuracy studies can be sufficient to make strong inferences about patient-important outcomes, when clinicians already have evidence from randomized trials showing that management of patients detected by a diagnostic test improves patient outcomes. However, this approach requires a clear understanding of the proposed place of a new test in a diagnostic pathway and its suggested benefits, as well as careful consideration of whether the patients detected by a new test are representative of the patients included in management trials (14).

Decide which outcomes to consider and what is their relative importance

According to the GRADE approach, results that are TP, TN, FP, FN, or inconclusive, as well as the complications of a test and its cost (resource utilization) constitute the outcomes of a diagnostic accuracy study (5).

For instance, consider the consequences of replacing inhalation challenge tests with more accessible and safer specific skin prick tests for immediate hypersensitivity for the diagnosis of work-related asthma (Table 2) (15). True positive results will lead to cessation of exposure (removal of the worker from the workplace, use of personal protection equipment, or instituting methods of eliminating or lowering allergen concentration); TN results will spare patients the possible adverse effects and burden as well as save the cost of the reference standard; FP results will cause unnecessary removal of the worker from the workplace and its likely financial consequences [many workers diagnosed with work-related asthma remain unemployed (15, 16)], needless use of personal protection equipment or futile use of resources to eliminate or lower allergen concentration at the workplace; and FN results will lead to delayed diagnosis, prolonged exposure and possible deterioration of asthma.

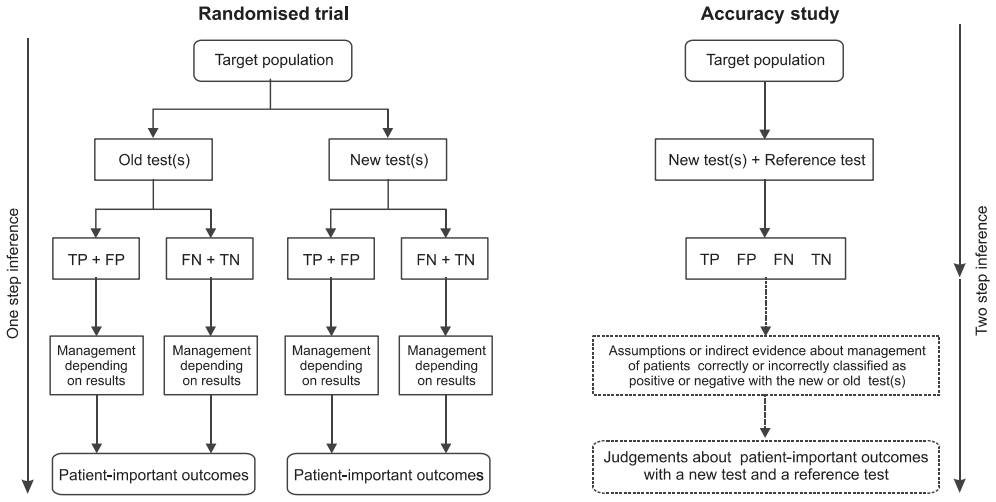

Figure 1. Two generic ways in which a test or diagnostic strategy can be evaluated. On the left, patients are randomized to a new test or to an old test and, depending on the results, receive the best available management (4). On the right, patients receive both (one or more) new test(s) and reference test. Investigators can then calculate the accuracy of the new test(s) compared with the reference test (first step). To make judgments about the relation of new test accuracy to patient-important outcomes, one needs to make additional assumptions (relying on the information from subsequently or previously done studies) about successive management and likely outcomes of patients categorized with a new test or a reference test as either having or not having a target condition (second step)(5). TP, true positive, FP, false positive, TN, true negative, FN, false negative.

© 2009 John Wiley & Sons A/S

The GRADE approach asks guideline developers to make judgements about the relative importance of each outcome for making a recommendation explicit (17). Outcomes such as removal from the workplace and its financial consequences to the worker diagnosed with work-related asthma might be considered critical for making a recommendation, whereas the burden of a skin prick test (other than allergic reactions) used to diagnose the allergy to a putative occupational agent might be considered important, but not critical (Table 1). In addition to their relative importance, the table shows the numbers of TP, FN, TN, and FP results, if specific skin prick tests are used instead of inhalation challenge tests. The key questions are whether the numbers of FP and FN results and their consequences will be acceptable.

Table 1. Key findings for diagnostic accuracy studies and their consequences (based on a systematic review of diagnostic tests for work-related asthma) (15). Illustrative question: should specific skin prick tests *vs* specific inhalation challenge tests for HMW agents be used in patients with asthma exposed to these agents with an average pretest probability of work-related asthma?

Test findings			
Pooled sensitivity	0.81 (95% CI: 0.70–0.88)	LR(+)*	2.03
Pooled specificity	0.60 (95% CI: 0.42–0.75)	LR(−)*	0.32

Consequences†		
	n per 1000‡	Importance§
TP	81	8
TN	540	8
FP	360	9
FN	19	7
Inconclusive results¶	–	4
Complications**	–	6
Cost††	–	5

LR, likelihood ratio;
TP, true positive results – important, because may improve patient's health by removal from workplace, use of personal protection equipment or induce cost associated with removing from or reducing the exposure;
TN, true negative results – important, because spare patients unnecessary removal from workplace and saves resources to reduce the exposure;
FP, false positive results – important, because patients may be unnecessarily removed from workplace and suffer financially or needlessly use personal protection equipment; resource use for reducing the exposure may be futile;
FN, false negative results – important, because lead to delayed diagnosis, prolonged exposure, and uncertain but possible deterioration of health.
*Likelihood ratios of positive and negative test results were estimated from pooled sensitivity and pooled specificity.
†Assuming a pre-test probability of work-related asthma of 10%, based on the estimate from a systematic review of studies assessing the proportion of adult asthma attributable to workplace factors (23).
‡All results are given per 1000 patients tested for a prevalence of 10% and pooled sensitivity and specificity shown above.
§Subjective rating on a 9-point scale; GRADE recommends a systematic classification of the importance of outcomes to patients as: 'not important' (score 1–3), 'important, but not critical' 1–6, or 'critical' 1–9to a decision 17.
¶Inconclusive results are either uninterpretable, indeterminate or intermediate test results – important because they 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, complications of both tests can be potentially serious.
††Values for these consequences are not reported because they are not exactly known on the basis of the available data, however, they are important.

Grade the quality of existing evidence

Study design. GRADE's four categories of quality of evidence reflect a gradient of confidence in estimates of the effect of a diagnostic test strategy on patient-important outcomes (13). In this system, a body of evidence obtained from randomized trials of alternative diagnostic strategies directly measuring patient-important outcomes is initially rated as high quality, and that obtained from observational management studies – as low quality. Quality of evidence from these trials or studies directly measuring patient-important outcomes is graded in the same way as for other interventions and the initial grading based on study design can decrease because of other factors (3, 13).

The GRADE system suggests different criteria when the evidence comes from the studies of diagnostic accuracy. Valid diagnostic accuracy studies – cross-sectional or cohort studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard – provide high quality evidence (5). However, they often are downgraded to lower quality evidence, because they are liable to limitations, particularly indirectness of outcomes, i.e. uncertainty about the link between the test accuracy and outcomes that are important to patients.

Thus, as for evidence about therapeutic interventions, the GRADE system provides additional quality criteria that can reduce the quality of evidence about using diagnostic tests.

Limitations in study design and/or execution (risk of bias). Valid studies of diagnostic test accuracy include representative and consecutive patients in whom legitimate diagnostic uncertainty exists (i.e. patients to whom clinicians would apply the test in the course of regular clinical practice), compare the test(s) under consideration with an appropriate reference or 'gold' standard, and assure that those who carry out or interpret the results of the test(s) under consideration are not aware of the results of the reference standard and vice versa. Serious limitations in study design or execution warrant downgrading of the quality of evidence. Guideline panels can use quality criteria from the existing instruments to assess the risk of bias in diagnostic accuracy studies (18–21).

In the example of using specific skin prick tests as a replacement for inhalation challenge tests for the diagnosis of work-related asthma, many studies of test accuracy suffered very serious methodological limitations (15) that warranted downgrading the quality of evidence by two levels – from high, through moderate, to low (Table 2).

Indirectness of evidence. In a previous article in this series, we described assessing the directness of evidence for therapeutic interventions (3). Guideline panels

Table 2. Quality assessment of diagnostic accuracy studies comparing specific skin prick tests to specific inhalation challenge tests [based on a systematic review by J. Beach et al. (15)]*

No. of studies	Design	Limitations	Indirectness	Inconsistency	Imprecision	Publication bias	Quality
True positives (patients with asthma correctly classified as allergic to occupational HMW agents)							
16 studies (434 patients)	Cross sectional studies	Very serious‡	Little or no	Serious¶	Little or no	No**	⊕○○○ very low
True negatives (patients with asthma correctly classified as not allergic to occupational HMW agents)							
16 studies (434 patients)	Cross sectional studies	Very serious‡	Little or no	Serious¶	Little or no	No**	⊕○○○ very low
False positives (patients with asthma incorrectly classified as allergic to occupational HMW agents)							
16 studies (434 patients)	Cross sectional studies	Very serious‡	Little or no	Serious¶	Little or no	No**	⊕○○○ very low
False negatives (patients with asthma incorrectly classified as not allergic to occupational HMW agents)							
16 studies (434 patients)	Cross sectional studies	Very serious‡	Serious§	Serious¶	Little or no	No**	⊕○○○ very low

*Full evidence profile should include a row for each of the outcomes important to patients associated with each possible test result (true positive, true negative, false positive, false negative, and inconclusive) as well as for complications and costs of test (Table 1); here, we present a simplified summary of quality of evidence only for outcomes judged to be critical.

‡In nine of 16 studies subjects were not selected consecutively, in 13 studies independent assessment of the tests was unclear, in eight studies there was likelihood of differential bias (those who tested negative or strongly positive were given a less or more thorough reference standard for verification), and in nine studies it was likely, that a decision to perform the reference standard was based upon the results of the test under consideration.

§Uncertainty about directness for false negatives relates to possible detrimental effects from delayed diagnosis, prolonged exposure, and uncertain but likely deterioration of health status.

¶Significant, unexplained heterogeneity of results for sensitivity (proportion of patients with positive inhalation challenge test who also tested positive with skin prick test) and specificity (proportion of patients with negative inhalation challenge test who also tested negative with skin prick test).

**Possibility of publication bias was not excluded, but considered insufficient to downgrade quality of evidence.

making recommendations about diagnostic tests face similar challenges regarding indirectness.

One type of indirectness can arise from indirect comparison of two or more alternative new tests. Indeed, two tests can be compared either directly in one study (every patient receives each of alternative tests and a reference standard) or indirectly in separate studies (in each study one of new tests is compared with a common reference standard). This indirect comparison can lower the quality of evidence supporting the choice among the two or more new tests.

Another type of indirectness arises from the differences in populations, tests, and outcomes of interest between the studies (existing evidence) and the scope of the recommendation (components of the PICO clinical question). Test accuracy may vary across populations of patients, because of a different spectrum of disease, differing competing diagnoses, etc. A guideline panel has to consider how similar the patients included in the studies are to patients for whom the recommendations are intended. Likewise, guideline panels and clinicians need to consider how comparable alternative tests used in the studies are to the tests used in the settings for which the recommendations are made. The differences may arise from different technical specifications (e.g. allergen extracts prepared by different manufacturers), the thresholds discriminating positive and negative results or categories of the results (e.g. different thresholds for 'positivity' when the fraction of nitric oxide in exhaled air is measured), how and when measurements were taken (e.g. during or outside the pollen season), as well as training and expertise of the individuals performing and interpreting the tests.

Finally, guideline panels making recommendations based on the results of diagnostic accuracy studies face a difficult challenge: of inferring from the information about test accuracy if applying a test will improve outcomes that are important to patients. Even if a new test is more accurate (reduces the number of FP and FN results), to what extent will it lead to improved patient-important outcomes? This judgement will depend on the role (triage, replacement or add-on) and the evidence about the relationship between the reference test and patient outcomes. Therefore, as we emphasize throughout this article, guideline panels may need to rate accuracy down because of uncertainty about impact on outcomes and thus conclude that only low quality evidence for making recommendations is available.

Recall again the example of specific skin prick tests for the diagnosis of work-related asthma. When defining the outcomes associated with particular test results and deciding on their importance, one can be quite certain about the likely consequences to patients. Thus, inferences that minimizing FP s and FN will benefit patients and that increasing them will have a negative impact on patient-important outcomes, are relatively strong. However, there is some uncertainty about directness for FN

results related to delayed diagnosis, prolonged exposure and uncertain but possible deterioration of health status. This uncertainty warrants further reduction in the quality of evidence for FN test results from low to very low (Table 2).

The impact of inconclusive test results is less clear; however, they will probably lead to performing the reference standard – specific inhalation challenge. The complications of specific inhalation challenge tests, although rare, are important. There is also higher burden and cost associated with specific inhalation challenge tests that are performed in specialized centres or sometimes at the workplace.

Inconsistency of results. GRADE suggests similar criteria as for therapeutic interventions to assess inconsistency of results. For diagnostic accuracy studies, unexplained inconsistency (heterogeneity) in the estimates of sensitivity, specificity, or likelihood ratios among individual studies can reduce quality of evidence. There was unexplained heterogeneity of results for sensitivity (proportion of patients with positive inhalation challenge tests that also had positive skin prick test results) and specificity (proportion of patients with negative inhalation challenge tests that also had negative skin prick test results) in a systematic review of specific skin prick tests with HMW occupational allergens as a replacement for the inhalation challenge tests in the diagnosis of work-related asthma (15). This heterogeneity (inconsistency) of results among studies justifies further reduction in the quality of evidence (defined as the extent to which our confidence in an estimate of the effect is adequate to support a recommendation) for TP, TN and FP results from low to very low (Table 2). It would also lower the quality of evidence of FN results were it not already downgraded to the very low category, because of limitations in design and indirectness.

Imprecision of results. As for therapeutic interventions, wide confidence intervals around estimates of test accuracy or true and FP and FN results can reduce quality of evidence from diagnostic accuracy studies. In the example of diagnosis of work-related asthma, the confidence intervals around sensitivity and specificity were narrow enough to infer that within the range of plausible results, the final recommendation would be the same.

Publication bias. High risk of publication bias can lower quality of evidence. Statistical methods for investigating publication bias in test accuracy studies differ from those used in intervention studies, because of the different nature of diagnostic and intervention questions (22). Publication bias may be suspected, however not proven, when published evidence is limited to few small studies, in particular, if they support a presumed hypothesis and were funded by a body with a vested interest in a particular diagnostic method.

Determine the overall quality of evidence across outcomes

Following the GRADE process, the overall quality of evidence across outcomes is determined by the lowest grade of quality for any of the outcomes deemed critical (their importance for the decision was rated as 7, 8, or 9 on a 9-point scale). However, in the example of diagnosis of work-related asthma, evidence for all critical outcomes was rated as very low quality (Table 2). Consequently, the overall quality of evidence supporting the recommendation for or against using skin prick tests with HMW allergens instead of specific inhalation challenge tests would be very low.

In this example, with an average probability of occupational asthma (23), skin prick test with HMW allergens resulted in a large number (36%) of FP results leading to unnecessary removal of the worker from the workplace and its consequences or futile use of resources to eliminate or lower allergen concentration at the workplace (Table 1). It also led to about 2% of patients who have occupational asthma being missed (FN).

Conclusions

The GRADE system provides a comprehensive and transparent approach for grading the quality of evidence and developing recommendations about diagnostic tests. GRADE stresses that test accuracy is a substitute for outcomes that are important to patients. Clinicians should always bear in mind that, whatever the test accuracy, application of any diagnostic test is of value only if it results in improved outcomes that are important for patients. Evaluation of diagnostic tests has recently accelerated, but the methodology of diagnostic research and number of performed studies lag far behind that for evaluating treatments. When looking for examples for this article, we were unable to identify any systematic review of diagnostic tests in allergy except for one health technology report on the diagnosis and management of work-related asthma (15). We encourage clinicians and researchers to engage actively in well-designed studies evaluating use of tests in allergy and to undertake systematic reviews of tests in allergy to identify areas where new studies are critically needed.

In the next article in this series, we will discuss how guideline panels and others making recommendations integrate the information about the quality of evidence and the estimated magnitude of the effects with controversial considerations of cost as well as patients' values and preferences to arrive at final clinical recommendations.

© 2009 John Wiley & Sons A/S

References

1. Diagnosis noun. The Oxford Dictionary of English (2nd edition revised), Soanes C, Stevenson A (Eds). Oxford University Press, 2005. Oxford Reference Online. Accessed 23 February 2009.
2. Diagnosis n. Concise Medical Dictionary. Oxford University Press, 2007. Oxford Reference Online. Oxford University Press. Accessed 23 February 2009.
3. Brozek JL, Akl EA, Alonso-Coello P, Lang D, Jaeschke RZ, Williams JW et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines. An overview of the GRADE approach and grading quality of evidence about interventions. Allergy 2008;**64**:669–677.
4. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;**356**:1844–1847.
5. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;**336**:1106–1110.
6. Burks AW. Peanut allergy. Lancet 2008;**371**:1538–1546.
7. Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. BMJ 2002;**324**:477–480.
8. Diagnosis. In: Hoad TF, editor. Concise Oxford Dictionary of English Etymology. Oxford University Press. Oxford Reference Online. Accessed 23 February 2009.
9. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;**332**:1089–1092.
10. Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. J Gen Intern Med 1989; **4**:288–295.
11. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. CMAJ 1988; **138**:697–703.
12. Guyatt GH, Jaeschke R, Prasad K, Cook DJ. Summarizing the evidence. In: Guyatt GH, Rennie D, Meade MO, Cook DJ, editors. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 2nd edn. McGraw-Hill Professional, 2008:523–542.
13. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is 'quality of evidence' and why is it important to clinicians? BMJ 2008;**336**:995–998.
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**:850–855.
15. Beach J, Rowe BH, Blitz S, Crumley E, Hooton N, Russell K et al. Diagnosis and management of work-related asthma. Evid Rep Technol Assess (Summ) 2005;**000**:1–8.
16. Ross DJ, McDonald JC. Health and employment after a diagnosis of occupational asthma: a descriptive study. Occup Med (Lond) 1998;**48**:219–225.
17. Schünemann HJ, Hill SR, Kakad M, Vist GE, Bellamy R, Stockman L et al. Transparent development of the WHO rapid advice guidelines. PLoS Med 2007;**4**:e119.
18. 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**:40–44.
19. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;**138**:W1–W12.
20. Simel DL, Rennie D, Bossuyt PM. The STARD statement for reporting diagnostic accuracy studies: application to the history and physical examination. J Gen Intern Med 2008;**23**:768–774.
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. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005;**58**:882–893.
23. Blanc PD, Toren K. How much adult asthma can be attributed to occupational factors? Am J Med 1999;**107**: 580–587.

© 2009 John Wiley & Sons A/S