Clinical Chemistry 50:5
806–818 (2004)

Special Report

Evidence-Based Guidelines in Laboratory Medicine: Principles and Methods

Wytze P. Oosterhuis,[1†*] David E. Bruns,[2] Joseph Watine,[3†] Sverre Sandberg,[4†] and Andrea R. Horvath[5†]

Background: **Guidelines are commonly used tools for supporting medical decisions. Formulating evidence-based recommendations has become a leading principle in guideline development.**
Aim: **This narrative review integrates the most recent methods of evidence-based guideline development and adapts those to the field of laboratory medicine.**
Summary: **We present a 10-step process and a list of criteria for the development of laboratory guidelines. Laboratory guidelines should be outcome oriented, be developed by a multidisciplinary team, and begin with a clear statement of the clinical question(s) that the use of the test(s) is addressing. The clinical questions define the type of study designs that offer the best evidence to answer those questions. Guidelines should be based on the critical appraisal and systematic review of literature and explicitly state the strength of evidence supporting each recommendation. Pragmatic considerations dictate that priority is given to topics with the highest clinical or economic impact. Scientific evidence is necessary but insufficient for recommendations, as considered judgment is required about benefits, harms, costs, and local applicability of recommendations. Formal consensus methods are needed when the evidence base is lacking or controversial. Guidelines should be disseminated widely and their impact monitored regularly. Regular reviewing is needed because the lack of timely updates is a major cause of nonadherence to guidelines.**
Conclusions: **Guidelines should be developed in a transparent process by a multidisciplinary team, with graded recommendations based on critically appraised scientific studies. Systematic, standardized, and explicit methodology, adapted to laboratory medicine, should be followed when developing recommendations involving the use of laboratory tests.**

Clinical practice guidelines aim at improving health outcomes by recommending effective interventions and advising against unnecessary, ineffective or harmful ones. Guidelines combine scientific evidence with patients' choices, clinicians' experience, and the availability of resources *(1)*. Guidelines are developed for a variety of purposes:

- To disseminate best practice based on systematically appraised scientific evidence;
- To decrease practice variation;
- To improve the reliability of medical decisions by use of standardized criteria;
- To improve quality of care and outcomes of patients;
- To decrease harm to patients and the misfortune of professional misconduct and court cases;
- To increase explicitness, transparency, and patients' information and autonomy of choice, thus facilitating ethical practice;
- To facilitate training, education, and continual professional development;
- To help target research to areas of uncertainty;
- To inform policymakers, payers, and managers; and
- To decrease costs and improve cost-effectiveness.

Clinical guidelines are statements that help healthcare professionals to make decisions about the appropriate and effective care for their patients *(2)*. Guidelines are most valuable where there is practice variation that affects patient outcomes and when there is strong evidence for or against certain interventions *(2, 3)*. Guidelines provide

[1] Department of Clinical Chemistry, St. Elisabeth Hospital, PO Box 90151, 5000 LC Tilburg, The Netherlands.
[2] Department of Pathology, University of Virginia Medical School, Charlottesville, VA.
[3] Laboratoire de Biologie Polyvalente, Centre Hospitalier Général, Rodez, France.
[4] Laboratory of Clinical Biochemistry, Haukeland University Hospital, and Division of General Practice, University of Bergen, Bergen, Norway.
[5] Department of Clinical Chemistry, University of Szeged, Albert Szent-Gyorgyi Medical School, Szeged, Hungary.
†Member of the IFCC Committee on Evidence-Based Laboratory Medicine.
*Author for correspondence. Fax 31-13-5352390; e-mail w.oosterhuis@elisabeth.nl.
Received August 6, 2003; accepted February 5, 2004.
Previously published online at DOI: 10.1373/clinchem.2003.025528

recommendations for best practice and serve as an aid to clinical judgment without replacing it *(3)*. They must not be seen as rigid standards or protocols of practice, but rather as quality improvement tools, and noncompliance with best practices may be justifiable in individual situations *(4, 5)*. In laboratory medicine, guidelines provide recommendations on the use of a wide range of tests in detecting or predicting a target condition, for staging and monitoring a disease, and for decisions to initiate, modify, or terminate treatments *(6)*.

Guideline development methods vary *(7)* and mostly rely on approaches for making therapeutic recommendations. Such approaches cannot be fully applied to diagnostic guidelines, and special methodologic standards are required for guideline development in laboratory medicine.

AIMS AND OBJECTIVES

In view of the above, the Committee on Evidence-based Laboratory Medicine of the IFCC reviewed the methods of several guideline development agencies and the methods of the "state of the art" of evidence-based diagnosis and adapted those principles to the specialist field of laboratory medicine. It is beyond the scope of this narrative review to discuss in depth the numerous methodologic and practical issues related to the development of evidence-based diagnostic guidelines. Throughout the text, however, interested readers are referred to literature that covers these aspects in detail. This review is targeted at professionals involved in developing and using guideline recommendations in laboratory medicine. We hope it will contribute to the improvement of the quality of guidelines on the use of laboratory tests.

TYPES OF GUIDELINES

There are several methods for developing guidelines:

Opinion-based guidelines. One or more experts in the field publish their recommendations. This process is not considered evidence based, and other expert groups may oppose the presented views.

Consensus-based guidelines. These have the advantages that their development is relatively rapid and inexpensive and that there is a high acceptance among users. The limitations of this approach are that the selection of the literature is not systematic, which can lead to biased conclusions and even contradictory (or potentially harmful) recommendations *(8–10)*. This might lead to noncompliance with and distrust in guidelines *(11)*.

Evidence-based guidelines. There is growing recognition that guidelines should be based, where possible, on the systematic identification, critical appraisal, and synthesis of the evidence. These evidence-based guidelines explicitly state the strength of evidence underlying each recommendation. In the most advanced approach, guideline development methodology involves an outcome-oriented process that explicitly takes into account all advantages and disadvantages, including those perceived by the individual patients, of the implementation of recommendations.

A distinction has to be made between (inter)nationally produced guidelines that are rigorously developed and follow a strict methodologic process and guidelines in their local form, which are characterized by pragmatism and simplicity. International guidelines should, most of all, assess the "objective" evidence behind the recommendations, whereas "local" guidelines/protocols must also take into account adaptability, practicability, local conditions of care, costs, and other factors. In their local form, guidelines should be composed of simple algorithms that give a practical sequence of steps to follow, an explanation of the content of the algorithm, and a detailed summary of the evidence that underpins such advice *(12)*.

Principles of Developing Laboratory Guidelines

GUIDELINES CONCERNING DIAGNOSIS VS THERAPY

Therapeutic recommendations are based on the effects of treatment, balancing benefits on the one hand with the harms and costs on the other. Therapeutic research is preferably based on randomized controlled trials (RCTs) investigating the effect of a treatment, as measured by differences in patient outcome. Possible outcome measures include mortality, morbidity, length of hospital stay, and complications.

Diagnostic trials have characteristics of their own. As in therapeutic trials, the best evidence of the value of a test is improved outcome (e.g., increased survival rate) *(13)*. Among diagnostic trials, however, RCTs are a minority; RCTs also are not always the optimal or most practical designs for studying the effect of laboratory testing on outcomes. Diagnostic RCTs, which focus on differences in health outcome, will always study combinations of a test (or tests) and subsequent treatments, with decisions on treatment being guided by the results of testing. It may sometimes be difficult to attribute differences in outcome to the effect of testing because clinical outcome is determined by both test and treatment. Furthermore, as the number of test–treatment strategies increases, so does the number of trial arms, making it almost impossible to compare them all in a randomized trial.

In laboratory medicine, an alternative to the trial is a diagnostic accuracy study, in which sensitivity and specificity or other indices of diagnostic accuracy of the test are reported. The best design for these diagnostic accuracy studies is a prospective cohort study with a blinded comparison of the performance of the experimental test and that of an appropriate gold standard test in a spectrum of patients suspected of having the disease in question. Blinding is usually considered to be particularly important when subjective test result interpretation is part of the diagnostic procedure. Because diagnosis is only one aspect of the overall patient management proto-

Table 1. Common topics covered in laboratory-related guidelines.

Preanalytical information
- Prevalence of the condition
- When to request/not to request a test
- Diagnostic algorithm
- Patient preparation
- Timing and frequency of testing
- Sample type and handling of specimens
- Biological variation

Postanalytical information
- Medical decision limits
- Laboratory-related outcomes
- Diagnostic sensitivity, specificity
- Predictive values
- Likelihood ratio, ROC curve
- Post-test probability
- Reference values
- Interpretation of test

Analytical information
- Selection and validation of test methods
- Detection limit, sensitivity, specificity
- Imprecision, bias, quality goals
- Standardization
- Internal and external quality control
- Interferences

Other information
- Turnaround time
- Where test is done (e.g., accreditation)
- Qualification and competence
- Organizational and cost implications
- Areas for further research

col, diagnostic accuracy is an indirect or "proxy" outcome measure. In most diagnostic studies it remains to be demonstrated that the patients identified by a certain test result are the ones who will benefit the most from treatment and, consequently, have an improved outcome (for more details see *sections 3* and *5* below). An important goal of studies of diagnostic tests is to determine whether the new test adds information to that known from patient observation or other investigations.

Guidelines address not only the diagnostic value of a test, but also other clinically useful attributes of a test's performance *(14)*. The potential scope of topics in evidence-based guidelines in laboratory medicine is listed in Table 1. It may be argued that all these (i.e., preanalytical, analytical, postanalytical, and other technical or organizational) elements of diagnostic performance should be seen as part of evidence-based laboratory medicine, which by definition focuses on decisions concerning individual patients to improve their outcome. We might not always be able to identify in quantitative terms the impact of all elements of test performance on clinical outcomes in laboratory medicine. Despite this fact, there is no doubt that there is such a relationship and that these characteristics therefore must be considered when defining the scope of evidence-based laboratory guidelines (Table 1).

USE OF EVIDENCE IN GUIDELINES

To make informed decisions, evidence is necessary but often not sufficient *(3)*. The evidence must be supplemented with considered judgment of the potential clinical benefits and harms, patients' preferences, the organizational and economic impact of testing, and other factors *(3, 15)*. For this reason, it may happen that the highest level of evidence may not provide the strongest recommendations in the local context because some of these confounders are weighted higher than the scientific evidence per se (for details see *section 6c* below).

Evidence can more readily tell us what not to do (e.g., because one test is less useful than another) than what to do. For example, the low-dose (1 μg) corticotropin-stimulation test is superior to a high-dose (250 μg) test *(16)*. In this case the evidence is sufficient to suggest not using the high dose. One classic symptom of vitamin B_{12} deficiency is tenderness of tongue and mouth. In patients presenting with complaints related to the tongue and mouth, the prevalence of B_{12} deficiency was shown to be only 8% *(17)*. The relatively low cost of testing for B_{12} deficiency and the availability of effective treatment may counterbalance the low probability of this cause of the complaints and might lead to a recommendation in favor of testing for B_{12} deficiency in one setting. In another community, however, the recommendation may be different because the prevalence or the relative costs are significantly different. An example where patients' choices are considered is the triple test used for antenatal screening of Down syndrome because the consequence of a positive screening test is an amniocentesis, which may harm the fetus, and in positive cases an abortion might be considered.

HOW MUCH EVIDENCE IS ENOUGH?

Ideally, all recommendations need to be supported by evidence. In practice this may be difficult to achieve, and there is considerable variation among guidelines in the amount of evidence gathered and analyzed before recommendations are made. To perform a systematic review for every single recommendation or statement in a guideline may be too great an effort for the guideline development team. Thus it may be critical to identify the most important or controversial questions and those with the greatest potential for a positive effect on patient outcomes or costs. These topics should be the ones given the highest priority to be covered by systematic reviews. Other, less critical or controversial areas can be dealt with by a simpler, but well-defined methodology, e.g., a review of a limited number of studies. When identified studies are homogeneous in their results, limiting the number of studies to those that have a large enough sample size and are of high quality will not lead to bias. The methods section of the guideline, however, should clearly describe how the guideline development group has dealt with this prob-

lem. The evidence base of the guideline could be expanded in later updates.

Guideline Development

The methods used by the guideline development group should be clearly defined and agreed on at the very beginning of the development process. Many detailed descriptions of guideline development are available *(2, 3, 18–21)*. Although these are mostly devoted to therapeutic guidelines, their principles are applicable to diagnostic guidelines *(22)*. The classic process of guideline development is described in Fig. 1 [adapted from Refs. *(3)* and *(20)*] and includes the 10 phases described in the following sections.

1. selection of guideline topics

In general, the aims of a guideline are to standardize patient management and to improve outcomes. However, cost reductions and promotion of the rational use of laboratory tests and good laboratory practice are also legitimate reasons for guideline development. Guideline development is a tedious and costly process; therefore, its topic should be carefully selected. The following criteria have been proposed as aids in selecting topics *(3, 23)*:

- Guidelines are most effective when the appropriateness of healthcare processes is uncertain, when the uptake of new and worthwhile interventions is suboptimal, or when controversies exist.
- When a guideline is developed there must be a realistic expectation that implementation is possible and that, if the guidelines are followed, quality of care and/or patient outcomes will improve.
- Prevalence and severity of the health condition must be considered. For example, screening for colon cancer or hemochromatosis and diagnosis and monitoring of diabetes mellitus are priority topics because the implementation of guidelines for these conditions could have a major impact on public health.
- High-volume (e.g., preoperative testing) and/or high-cost diagnostic interventions (e.g., erythropoietin or molecular biology testing) can also be good targets.
- It is less worthwhile to develop guidelines if there is not enough evidence on which to base recommendations. It should be kept in mind, however, that the lack of evidence for the effect of a procedure is not equal to evidence that the procedure has no effect.

2. establishing a multidisciplinary guideline development group

In setting up the guideline development team, the target groups of the guideline should be carefully identified: e.g., are they targeted to patients, doctors, nurses, laboratory specialists, phlebotomists, or other groups? The working group should include all parties involved in the management of the target condition, such as doctors, patients, nurses, and other support staff, as well as managers, payers, and other policymakers *(24, 25)*. Guidelines should not be developed exclusively by high-profile professionals, especially if they are insulated from the day-to-day pressures of medical care. If a guideline differs too much from the routine working practices of most health professionals, it will act only as a gold standard to be admired *(26)*. Among guidelines on the use of laboratory tests, those that are developed exclusively by either clinicians or representatives of laboratory medicine are of limited value in terms of both recognition and clinical implementation. They can, however, have the status of a position statement of the discipline and thereby could very well serve as a starting point for further (or local) multidisciplinary guideline development.

In laboratory medicine, effective guideline development groups need to have members with the following skills: expertise in the field of laboratory medicine or subspecialty; clinical expertise; methodologic expertise in statistics, literature searching, critical appraisal, guideline development, health economics, and bioethics; practical understanding of problems faced in the delivery of diagnostic services and care (e.g., laboratory manager, phlebotomist, and nurse; in self-monitoring of glucose, a patient and a diabetes nurse or social worker may also be very useful) *(3)*. The multidisciplinarity of the guideline development team is key to successful implementation of recommendations.

Some organizations that undertake guideline development take precautions that group members have no conflict of interest. Group members should be able to act independently of undue outside influences whether internal (e.g., academic pressure), external (e.g., granting or sponsoring bodies, politics), or commercial (e.g., drug companies, the diagnostics industry, or health insurance organizations). It is advisable to make a declaration of interest, to be completed by group members, before the process of guideline development starts. Again this poses a special problem for diagnostic guidelines because many suppliers are prepared to support guideline development groups. Because guideline development is a time-consuming and expensive task, such support is, in principle, acceptable provided it does not influence the objectivity of the final recommendations. For transparency, sponsorship, or the lack of it, together with the names and the precise roles of contributors to the guideline development process should be made explicit.

3. identification of the scope of the guideline and of outcomes to be addressed

The guideline development process should start with the definition of the scope and the most relevant outcomes of the guideline. The scope of the guideline could cover areas listed in Table 1. Usually the scope of the guideline needs to be refined further during the collection of the literature. If the members of the guideline development group fail to carry out this refinement, they run the risk of having too broad a scope, which makes systematic re-

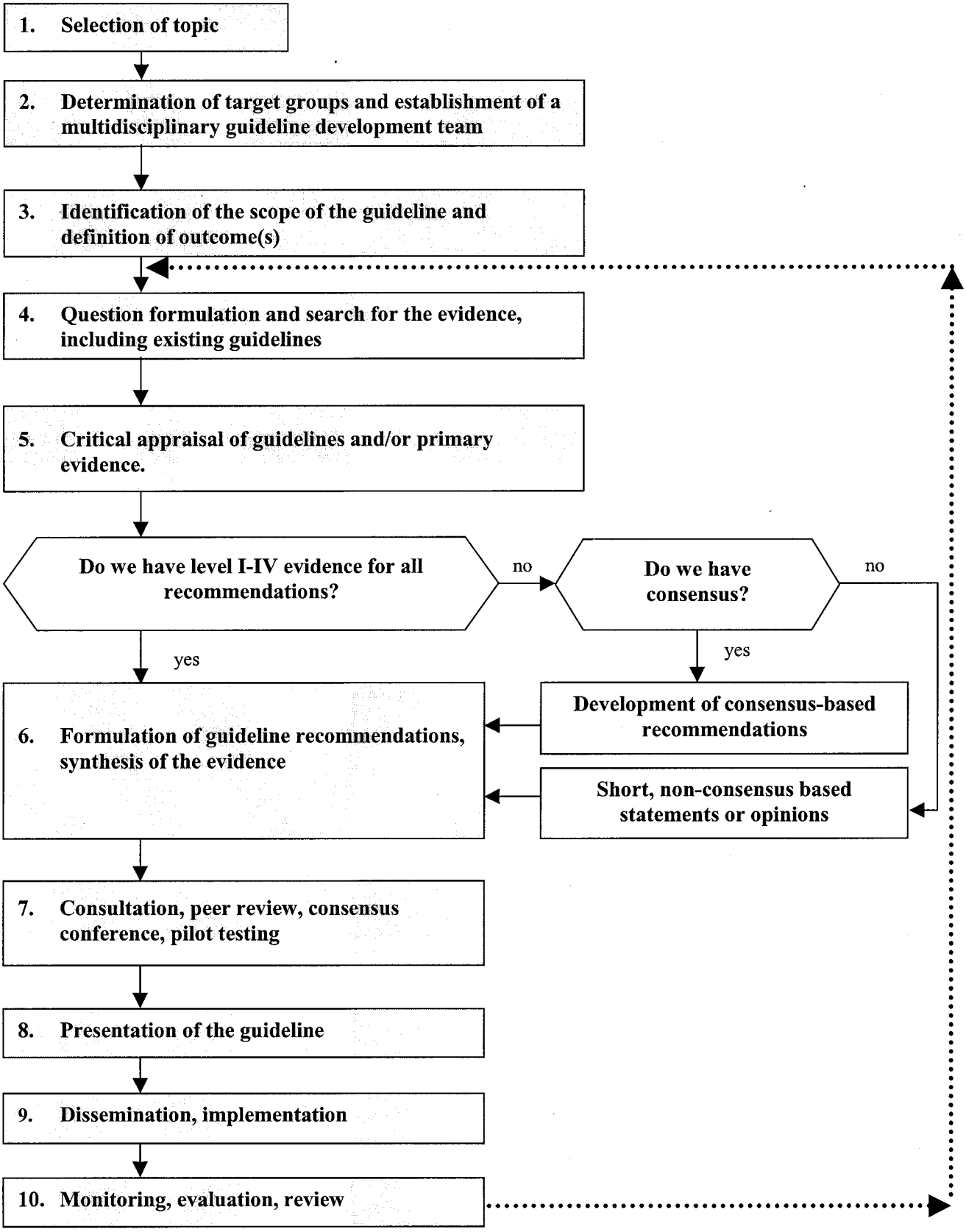

Fig. 1. Process of guideline development.

viewing of the literature far too complicated or even unmanageable.

In considering which outcomes are to be addressed, it is worthwhile to recall that outcomes are results of medical interventions in terms of health or cost *(27)*. All relevant outcomes should be considered when guideline recommendations are developed because their identification is key to informed diagnostic decisions. It also helps to formulate questions and to search for the evidence (see *section 4* below). Outcomes related to laboratory testing may be difficult to measure and may be perceived differently by patients and those who sell, carry out, or interpret the tests *(28)*. Outcomes related to laboratory testing can be clinically, operationally, and economically relevant *(15, 28)*. Examples of laboratory variables and potentially related outcomes include the effects of

- Availability (or lack) of testing on mortality, morbidity, incidence of disease, admission, readmission or discharge rate, or on quality of life;
- Method or the turnaround time of testing on the length of hospital stay or on patient satisfaction;
- Accuracy of the test on diagnosis or prognosis *(28)*; and
- Test availability on cost per life-year *(29)*.

A guideline, e.g., on the use of laboratory tests related to monitoring therapy with erythropoietin, may consider the following outcomes: a better prediction of patients who would benefit from this therapy and an improved recognition of nonresponders, better choice of dose (as measured by improved hemoglobin concentration), a decrease in blood transfusions in certain subpopulations, and thus reduced risks and improved cost-effectiveness.

4. QUESTION FORMULATION AND SEARCH FOR EVIDENCE

Asking the right question and asking it right. Question formulation is at the heart of guideline development. Questions should identify the reason(s) the test is being used and whether its result will influence clinical decisions and outcomes *(15)*. A clear question is essential for locating and selecting studies and for critically appraising their validity and relevance. In evidence-based laboratory medicine, a structured question consists of three or four parts: (1) the patient with a problem; (2) the diagnostic intervention; (3) its comparison with another test or a reference standard; and (4) the outcome. Using two examples, we illustrate below two types of questions that can be formulated differently, depending on whether the question is related to the diagnostic accuracy or the diagnostic utility of a test. These two questions are formulated according to the four- or three-part models indicated above:

Type 1 question (regarding the diagnostic accuracy of a test): In (1) patients presenting to the emergency department with shortness of breath, how well does (2) N-terminal pro-B-type natriuretic peptide (4) predict heart failure as assessed by (3) the cardiac ejection fraction measured by echocardiography?

Type 2 (related to the value of a test in improving patient outcomes): In (1) patients admitted to the hospital for treatment of heart failure, how well does (2) the use of N-terminal pro-B-type natriuretic peptide as a guide to therapy (4) improve the length of hospital stay and the rate of subsequent readmission for heart failure?

Searching for the evidence. The search strategy should guarantee that all relevant literature, or at least a representative or unbiased sample of the literature, is entered into the review process. It is best to start the search with looking for external evidence-based guidelines that can be updated and adapted. There are numerous websites where such guidelines can be found (Table 2). If evidence-based guidelines are available, it is important to check their expiry date and the last date of references, which will guide the updating process and the subsequent search of the most recent literature. All external guidelines must go through an appraisal process, which will be discussed in the next section.

If no recent external guidelines are available, methods of a systematic literature search include searching in electronic databases (Table 2) with predefined keywords; screening the reference list of retrieved studies, (systematic) reviews, and handbooks; hand-searching; and contacting experts in the subject. Restrictions in publication date, language, and other areas should be stated. The search for evidence usually starts in databases such as the Cochrane Library, which contains high-quality systematic reviews or metaanalyses. The Cochrane Library does not yet contain systematic reviews of diagnostic studies, but these will be included in the future. Several databases have been established for diagnostic studies, such as the MEDION database in Belgium, which contains >1000 references to reviews and methodologic reports on diagnostic studies. The DARE database and the periodicals *ACP Journal Club* and *Evidence-Based Medicine* include structured abstracts and commentaries of diagnostic reviews that meet methodologic criteria. The DARE database in March 2003 contained 572 diagnostic overviews when searching with key words of "sensitivity AND specificity", of which 103 systematic reviews were relevant to laboratory medicine. The database of the Committee on Evidence-Based Laboratory Medicine of the IFCC consists of ~80 systematic reviews in laboratory medicine. Several other references, such as the journals of *Bandolier* and the "Evidence-based Laboratory Medicine and Test Utilization" section of *Clinical Chemistry* are also good sources of critically appraised secondary publications or guidelines that synthesize the primary literature in a digestible format. Other sources of high-quality diagnostic overviews are health technology assessment databases. For an updated collection of evidence-based

Table 2. Evidence-based database resources.

Guideline databases
- AGREE (Appraisal of Guidelines, Research and Evaluation for Europe): http://www.agreecollaboration.org
- American Association of Clinical Endocrinologists: www.aace.com/clin/
- American College of Chest Physicians: www.chestnet.org
- American College of Physicians: http://www.acponline.org
- Australian National Health and Medical Research Council (NHMRC): http://www.health.gov.au/nhmrc/publications/
- Centre for Health Services Research (CHSR): www.ncl.ac.uk/pahs/research/services/
- Clinical Efficacy Assessment Project (CEAP): www.acponline.org/sci-policy/guidelines/ceap.htm
- CDC Task Force on Community Preventive Services: http://www.thecommunityguide.org
- Clinical Practice Guidelines: http://www.kurucz.ca/sue/clinicalpracticeguidelines.htm
- European Society of Cardiology: http://www.escardio.org/
- Finnish guidelines (in English): http://www.ebm-guidelines.com/
- German Agency for Quality in Medicine (www.aezq.de) Guidelines Information Service: www.leitlinien.de
- Guidelines-International-Network (G.I.N.): http://www.g-i-n.net
- Health Services Technology Assessment: http://hstat.nlm.nih.gov/hq/Hquest
- Health Technology Assessment databases: http://www.hta.nhsweb.nhs.uk/htapubs.htm; http://www.inahta.org/; and http://www.shef.ac.uk/~scharr/ir/htaorg.html
- New Zealand Guidelines Group: http://www.nzgg.org.nz
- NHS National Institute for Clinical Excellence: http://www.nice.org.uk
- SCHARR database: http://www.shef.ac.uk/~scharr/ir/guidelin.html
- Scottish Intercollegiate Guidelines Network (SIGN): http://www.sign.ac.uk/
- US Agency for Healthcare Research and Quality: www.ahrq.com
- US National Guideline Clearing House: http://www.guideline.gov
- US Preventive Services Task Force: http://www.ahcpr.gov/clinic/cps3dix.htm#Background

Systematic review and evidence-based resources
- ACP Journal Club: http://hiru.mcmaster.ca/acpjc/default.htm
- Cochrane Library: http://www.cochrane.org/
- Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests: www.cochrane.org/cochrane/sadtdoc1.htm
- DARE (Database of Abstracts of Reviews of Effectiveness): http://agatha.york.ac.uk
- Evidence-Based Medicine: http://www.bmjpg.com/template.cfm?name=specjou_be
- IFCC Committee on Evidence-Based Laboratory Medicine (C-EBLM) database: www.ckchl-mb.nl/ifcc
- MEDION database: www.mediondatabase.nl
- National Library for Medicine: http://text.nlm.nih.gov
- National Library for Health: http://www.nelh.nhs.uk/guidelinesdb/html/glframes.htm
- National Health Service (NHS) Centre for Evidence-Based Medicine: http://www.cebm.net/; Levels of evidence and grades of recommendations: http://www.cebm.net/levels_of_evidence.asp
- NHS Centre for Reviews and Dissemination: http://www.york.ac.uk/inst/crd/
- SCHARR database: http://www.shef.ac.uk/~scharr/ir/netting/

databases, we recommend the ScHARR database (Table 2).

If a search is not successful in the secondary literature, one can look for primary reports on Medline or Embase (preferably both because the overlap is often incomplete). We recommend using PubMed (http://www.ncbi.nlm.nih.gov) for searching Medline [see also Fielding and Powell *(30)*]. The most frequently used search strategy to locate diagnostic studies in Medline is probably the one developed by Haynes, which is available as a "clinical query" on PubMed *(31)* There is also a special feature in PubMed to retrieve systematic reviews and metaanalyses. For laboratory tests, the best single search term is "sensitivity" in the title or abstract or as a MeSH term. However, a more sensitive search can be achieved with the combination of "sensitivity-AND-specificity", and the words "diagnostic test", "diagnosis", or "diagnostic use" combined with the corresponding clinical condition (e.g., chronic renal failure), and finally the name of the test (e.g., soluble transferrin receptor). Names of diagnostic tests used in simple searches will be sensitive but not specific. The frequency with which the term sensitivity-AND-specificity occurs in a database indicates the frequency of diagnostic test evaluation studies. The search strategies must be customized to each database to take the different indexing systems into account. Regardless of indexing, it is recommended to combine controlled index terms (e.g., MeSH terms in Medline) with free text terms. Narrowing the search to a specific condition will generally trim down search results to a manageable number of studies.

5. APPRAISAL OF EVIDENCE

Adapting external guidelines. Guidelines and recommendations developed by others can be adapted, but only after critical appraisal of whether they fulfill certain quality criteria. Quality criteria have been developed specifically for the evaluation of guidelines, e.g., the Appraisal of Guidelines for Research & Evaluation in Europe Instrument *(32)*.

Critical appraisal of literature. After systematic searching of the literature, relevant studies of appropriate study design for the questions of the guideline *(33–35)* should be selected according to predefined criteria. Individual studies should be critically appraised to assess whether they are of sufficiently high quality that recommendations to be based on them with confidence. The careful appraisal of individual studies is usually done by two independent reviewers with the help of standardized checklists. There are different checklists for different study designs, e.g., at the Centre for Evidence-Based Medicine (www.cebm.net), at Casp International Network (www.caspinternational.org.uk), and at the Centre for Health Evidence (www.cche.net). Potential disagreements in the assessment process are resolved by consultation and consensus.

Factors to be assessed during the critical appraisal process relate both to the internal and external validity of

studies. The internal validity of the study refers to the extent to which its design and conduct are likely to be free from random and systematic errors. A high-quality diagnostic study should be free from biases *(36, 37)*, and seven essential categories should be considered in the critical appraisal process *(38)*: (*a*) the patient population should cover the whole spectrum of the disease to avoid spectrum bias (i.e., biased selection of study patients); (*b*) there should be an adequate description of the test; (*c*) an appropriate reference standard should be applied to all study participants; (*d*) the result should be interpreted blindly without the knowledge of the experimental test and vice versa; (*e*) study results should be reported unambiguously, with accurate description of all determinate and indeterminate results and with confidence intervals for estimates of measures of diagnostic accuracy (e.g., sensitivity, specificity, likelihood ratios, and area under the ROC curve); (*f*) prevalence of the target condition in all patients and in relevant subgroups should be given; and (*g*) there should be an assessment of the study's external validity, or applicability (whether the findings of a diagnostic study are generalizable to a particular population or setting).

The QUADAS tool has recently been proposed to help assess the quality of studies of diagnostic accuracy *(39)*. Standards on reporting of trials, such as the STARD statement for diagnostic studies, are not designed for use as critical appraisal checklists, but they can be used as a guide *(3, 40)*. A practical checklist was also constructed by the Bayes Library Working Group *(41)*. The evaluation and scoring of the quality of primary studies is subject to debate, and there is much heterogeneity in the way authors of systematic reviews deal with this issue *(38, 42, 43)*. A committee of the Agency for Healthcare Research and Quality evaluated the many different grading systems that have been developed *(38)*. In this systematic analysis of the available checklists on diagnostic studies, three were considered methodologically acceptable *(37, 44, 45)*. Two others did not address test description, but this omission is easily remedied *(46, 47)*. The committee concluded that "the information base for understanding how best to rate the quality of studies remains incomplete".

6. FORMULATION OF GUIDELINE RECOMMENDATIONS: SYNTHESIS OF THE EVIDENCE

Evidence-based guidelines contain graded recommendations, which are based on synthesis of the evidence. The grading system of guideline recommendations is influenced not only by the strength of the evidence, but also by other elements, as discussed before (see section on the "Use of Evidence in Guidelines") *(48)*. High-grade evidence is not always available to support each recommendation, but for transparency and explicitness and for informing readers it is also important to acknowledge whether the evidence base is poor or lacking. Furthermore, identification of knowledge gaps helps in designing better studies in the future. If a recommendation is not supported by high-grade evidence, it can be supported by the consensus of experts. Formal consensus development methods, such as the nominal group technique and the Delphi or consensus conference methods, can be used to minimize the influence of individual opinions and, thus, bias *(49)*.

There is no international consensus on a grading system for determining the strength of evidence supporting recommendations. Several grading systems exist, mainly for therapeutic interventions *(38)*. The guideline development group must decide which system is used, and a clear definition of the grading method needs to be given in the guideline. Irrespective of the grading system used, four key steps in the process can be identified *(3, 48)*:

(*a*) Determine the level of evidence of the primary studies or reviews included in the guideline on the basis of study type combined with the assessment of methodologic quality (Table 3);
(*b*) Compile an evidence table of studies of acceptable quality, identified as relevant to each of the key clinical questions addressed by the guideline;
(*c*) Make considered judgment as to the generalizability and applicability (i.e., external validity) of the evidence to the target patient group of the guideline, the consistency of the evidence base, and the likely clinical impact of the intervention; and
(*d*) Assign a grade to the recommendation by integrating the strength of the evidence and the degree of extrapolation required to form the recommendation.

Table 3. System to rate the level of evidence and the strength of guideline recommendations in diagnosis [example, adapted from Ref. *(51)*].

Quality of primary studies and reviews: rating the level of evidence of individual articles

Ia	Metaanalysis or systematic review based on at least several level Ib studies
Ib	Diagnostic trial or outcome study of good quality
II	Diagnostic trial or outcome study of medium quality, insufficient patients, or other trials (case–control, other designs)
III	Descriptive studies, case reports, other studies
IV	Statements of committees, opinions of experts, and so forth (reviews, not systematic)

Rating of the strength of the evidence supporting guideline recommendations:

A	Supported by at least two independent studies of level Ib or one review of level Ia ("it was shown/demonstrated")
B	Supported by at least two independent studies of level II or extrapolations[a] from level I studies ("it is plausible")
C	Not supported by sufficient studies of level I or II ("indications")
D	Advise of experts ("there is no proof")

[a] When data are used in recommendations for situations different from the original setting of the study.

The above steps are explained in more detail below:

(a) Determine the level of evidence of the primary studies and reviews. To score the level of evidence of the collected primary studies and reviews, define the type of the study (i.e., study design) and critically appraise the methodologic quality of the studies using set criteria. Give each study a quality rating according to a standard scale. A study can be characterized as having good, fair, or limited quality. Studies with limited quality should not be used to support recommendations.

There are many ways to determine the level of evidence. It is generally accepted that the highest level of evidence is a "good-quality", well-conducted systematic review or metaanalysis of RCTs (level Ia). In diagnostic guidelines, depending on whether the recommendation addresses, for example, diagnosis of or screening for a disease, diagnostic trials with differing designs and outcomes will represent the highest level of evidence *(34, 35)*. For example, in recommendations on the use of prostate-specific antigen in screening of prostate carcinoma, randomized controlled studies with a hard outcome measure (e.g., survival) are necessary. In recommendations on test accuracy, e.g., total prostate-specific antigen vs the free/total prostate-specific antigen ratio in diagnosis of prostate cancer, prospective cohort studies with diagnostic accuracy data, such as sensitivity, specificity, or likelihood ratios of the test in detecting the disease, are sufficient to produce the highest level of evidence. This again stresses the importance of formulating clear questions when developing guidelines in laboratory medicine. Statements of committees and expert opinion are graded level IV. A simple system is presented in Table 3. A more extensive system for grading the level of evidence is available on the website of the National Health Service Centre for Evidence-Based Medicine (Table 2). However, as yet no agreed system exists for diagnostic studies.

(b) Compile an evidence table. Once the assessment of the level of evidence is complete, compile an evidence table that summarizes all relevant studies. Evidence tables can be used for checking the consistency of data obtained in different studies or for subgrouping studies of similar study designs, patient populations, and validity or quality criteria. Evidence tables should include the following information [adapted from the Scottish Intercollegiate Guidelines Network *(3, 48)*]:

- Publication details of the individual studies;
- Study design;
- Size of study (i.e., number of patients included);
- Spectrum of patients and patient setting;
- Prevalence of the condition;
- Diagnostic tests used or compared (i.e., index and reference tests);
- Outcomes measured (i.e., patient- or laboratory-related outcomes);
- Effects measured, including measures of diagnostic accuracy (e.g., sensitivity, specificity, likelihood ratio, and ROC curve) with the level uncertainty (e.g., confidence intervals; *P* values);
- Comments on specific issues raised by the study [e.g., potential bias(es) in the study]; and
- Quality rating and level of evidence of the study.

(c) Make considered judgment. On the basis of the facts in the evidence table, the development group needs to formulate the final recommendations. This is perhaps the most difficult step and requires several skills, methodologic knowledge, and experience in decision analysis. The way this exercise is carried out varies from guideline to guideline; transparency of this process is, therefore, essential.

This task usually starts with applying a grade to the strength of evidence. The strength of evidence is a broader term than the level of evidence and refers to the total body of evidence supporting the recommendation. It incorporates the elements used in defining the level of evidence of single studies (i.e., study design and quality), but adds the investigation of more components to its determination. West et al. *(38)* identified only 8 grading systems of 40 that fully addressed the following three domains of strength of evidence:

1. Quality of the evidence: The quality of all relevant studies for a given topic, where "quality" is defined as the extent to which a study's design, conduct, and analysis have minimized selection, measurement, and confounding biases.
2. Quantity of the evidence: "Quantity" is defined as the magnitude of the effect, the number of studies that have evaluated the given topic, or the overall sample size across all included studies.
3. Consistency of the evidence: For any given topic, the extent to which similar findings are reported from work using similar or different study designs.

In other words, the strength assigned to the evidence supporting a recommendation indicates to users the likelihood that, if that recommendation is implemented, the predicted outcome will be achieved. A major challenge arises in evaluating the consistency of a body of knowledge consisting of different types of studies (e.g., two randomized trials, five cohort studies, and eight case-control studies) with conflicting conclusions (e.g., the two highest level randomized studies in 500 patients show negative results, whereas 13 lower level studies in 1500 patients show a positive effect). No grading system, no matter how good it is, can solve this problem easily. In principle, when such heterogeneity arises, greater weight should be given to findings from better-conducted studies *(3)*. However, the generalizability of the evidence should also be considered in these situations because it is possible that the results of the well-conducted RCTs cannot be generalized to the clinical setting in which the recommen-

dations are meant to be used, whereas those of the less well-designed studies can.

When studies of mixed validity are being combined it is important that guideline developers investigate the potential sources of heterogeneity by carrying out, for example, subgroup analysis or sensitivity analysis. In subgroup analysis, the results for groups with similar characteristics within the whole study population of the review are synthesized separately. Sensitivity analysis means repeating the analysis by excluding studies, or a subgroup of data, that are shown to introduce heterogeneity or are of low quality and suffer from several forms of biases. The results of reanalysis after excluding some studies are compared with the original results of the overview to assess the degree to which these excluded data affect the final conclusions. Statistical models, such as fixed-effect, random-effect, and multivariate models, are also commonly used to investigate heterogeneity in this context. Guideline development groups need to ask for methodologic help from biostatisticians in these situations.

(d) Assign a grade to the recommendation. The final step of formulating recommendations is to assign a grade to statements within the guideline. Most grading systems score only the strength of evidence rather than the strength of recommendations (Table 3). It should be emphasized that the strength of recommendation is a broader term, and the strength of evidence is only one element in its establishment. Not only the strength of the evidence, but also several other factors, such as clinical impact, benefits and harms, generalizability, applicability, practicality, the costs and burden to patients or society, and cultural or ethical considerations may influence the final decision when the group formulates and grades the strength of recommendations *(38, 48)*. Even when many high-quality studies consistently reach the same conclusion, judgment must be made about the applicability of the studies to the target population of the guideline. In some systems a lower grade recommendation is made when results from a study must be extrapolated (e.g., when the study population and the population addressed in the guideline are not identical). Guideline users often misinterpret the implications of the grading system. It is important to emphasize that grading of the evidence supporting the recommendation does not relate to the importance of the recommendation.

The grading system for recommendations varies among guideline agencies. There are no agreed systems, and, in particular, there is no consensus on which grading system should be used in diagnostic guidelines. Table 3 represents an example of the systems currently in use in the field of laboratory medicine. There is a move, however, initiated by the GRADE working group, toward harmonization of the grading system of different guideline development organizations around the world.

7. consultation and peer review, consensus conference, pilot testing

Depending on the authority (international, national, local) and the target groups of the recommendations, draft guidelines should go through a peer-review and consultation process involving recognized experts in the field(s), professional and patient representatives potentially involved in the use of the guideline, and relevant national/international societies so that important disagreements can be resolved before the guideline is launched. It is preferred to present the guideline at a multidisciplinary consensus meeting before final approval *(49)*. The Internet opens new possibilities for guideline presentation, consultation, and review. Opinions obtained in the consultation process should be built into the guideline, as appropriate. Individual opinions of key experts or bodies, if different from the conclusions of the core group, should also be explicitly referred to. Pilot testing before dissemination helps the successful implementation of guidelines. Results of pilot testing, if relevant, should also be incorporated into the final guideline *(3)*.

8. presentation of the guideline

The published guideline should explicitly state the methodology used for its development, including the grading system, and how recommendations were linked to the evidence. The transparency of the whole process facilitates the acceptance of the guideline by users.

This list, adapted from the Scottish Intercollegiate Guidelines Network, contains the most essential elements of a good diagnostic guideline *(3)*:

- Title and publication details:
  - —Authority responsible for issuing the guideline (e.g., international or national professional organization);
  - —Date and place of issue;
  - —Expiry date, planned update;
  - —Legal considerations.
- Definitions and abbreviations.
- Introduction:
  - —Background and rationale: Epidemiology, prevalence, incidence, demographics, summary of clinical problem and outcomes, clinical context, description of current practice, variations in practice, prognosis, and so forth. Definition of the clinical condition or aspects of management that the proposed guideline will address. Definition of the diagnostic investigation(s). Indication of the benefits likely to arise from the implementation of the guideline.
  - —Aims of the guideline: description of the main aims, objectives, intent, and target groups. Definition of the patient group and setting to which the guideline will apply.
- Guideline development methodology:
  - —Guideline development group, including peer reviewers, consultants, and observers and their affilia-

tions, sponsors, and declaration of conflicts of interest;
—Guideline development methods;
—Description of the grading system(s).
- Summary:
  —Brief summary of key recommendations;
  —Algorithms, preferably following the logical and procedural framework of diagnostic decisions.
- Detailed guideline:
  —Detailed recommendations with reference to the underlying evidence;
  —Indication of the level and strength of the evidence supporting recommendations;
  —Prevalence/pre-test probability in the given patient spectrum;
  —Size of effect (e.g., measures of diagnostic accuracy and their level of uncertainty).
- Strategy for implementation and monitoring of guideline (e.g., standards and indicators).
- References.
- Appendix:
  —Key questions, search strategies, and inclusion/exclusion criteria for studies;
  —Data extraction and critical appraisal forms, evidence tables, balance sheets, and other pertinent documentation;
  —Additional information: preanalytical information, patient information, materials supporting implementation (e.g., reminders, clinical audit data collection forms, and questionnaires);
  —Consensus conference reports, summary of differing opinions, comments.

9. DISSEMINATION AND IMPLEMENTATION

Systematic research has shown that passive dissemination of guidelines is not satisfactory *(50)*. Implementation requires active dissemination through training and continual medical education or by outreach visits and facilitation. Also useful are reminders put in patient records, electronic warning systems, diagnostic algorithms, checklists, or standardized questionnaires. Rapid access to recommendations can be provided by clinical intranet or decision support systems linked to patient records as well as by evidence-based databases on palm-top computers.

Respected clinical or diagnostic experts should be involved in the dissemination process. Insurance policies often use strict protocols and financial incentives, which if based on evidence could be a powerful vehicle for encouraging the use of rational diagnostic pathways. Patient information leaflets (e.g., for self-monitoring of glucose, collection of timed samples, and instructions for patient preparation before testing) should accompany the guideline. Implementation, monitoring, and feedback of the impact of guidelines are usually linked to local quality-improvement and clinical audit programs.

10. MONITORING, EVALUATION, AND REVIEW

A leading reason for physicians' failure to comply with guidelines is that the guidelines become obsolete. Some clinicians change their practice in response to new evidence faster than guidelines are updated. It is recommended to include in the guideline an expiry and revision date. It is also important that the impact of the guideline be monitored regularly, for example, in clinical audit cycles, which can focus on the following aspects:

- Demography: spectrum and severity of patients and diseases.
- Diagnosis: Which interventions are used and how does actual test utilization compare with recommendations?
- Outcome of care: Does the use of the diagnostic intervention influence therapy or the health status of patients or the rate of complications?
- Effectiveness of care: do the quality and effectiveness of care improve?
- Costs of care: do the costs of care decrease?
- Implementation: do professionals and patients collaborate and comply with the recommendations?

Conclusions

We have described the general principles and processes of developing evidence-based laboratory guidelines. Most guidelines, however, contain recommendations on both diagnostic procedures and therapeutic interventions. The methods described here also apply to the formulation of recommendations on diagnostic procedures within broader "clinical" guidelines. Laboratory professionals should be involved in guideline development if laboratory tests are included in clinical recommendations. On the other hand, guidelines developed exclusively by laboratory professionals are equally of limited value in terms of both recognition and implementation. They can, however, serve as a position statement of the discipline and as a starting point for further (or local) multidisciplinary development of clinical guidelines.

Following the principles of evidence-based medicine, all recommendations of an ideal guideline need to be supported by the best available evidence. Performing a systematic review for every single recommendation in the guideline may be too great an effort for the development group, especially in laboratory diagnosis where the evidence base is often poor. Thus it may be necessary to prioritize and cover by systematic reviews only those areas that are controversial or have the greatest impact on patient outcomes or on costs. Other, less critical, areas can be dealt with by a simpler but well-defined and clearly described methodology, e.g., inclusion of a few homogeneous studies that have a large enough sample size and are of high quality.

High-quality scientific data are not sufficient to make practice recommendations: social, ethical, financial, and other considerations and balancing of harms and benefits play an important part when guideline recommendations

are used as decision support tools in practice. For the above-mentioned reasons, there will, and should always, be a consensus element in evidence-based guideline development. This also highlights the need for transparency, explicitness, and the multidisciplinary nature of the guideline development process, which further contribute to the acceptance and uptake of recommendations. Local recommendations of best practice are usually much simpler than (inter)nationally produced guidelines. It is the responsibility of professionals to critically evaluate, adapt, update, and pilot test externally produced guidelines and translate them into practical recommendations at the local level. Guidelines should be widely reviewed and pilot tested before release. Publication of guidelines is insufficient for successful implementation. Active dissemination, supported by regular monitoring and measuring of the effect of guidelines, is essential for achieving improved outcomes. Regular revision and timely updates should be "part and parcel" of the guideline development process.

Evidence-based guideline development is a hard, time-consuming, and costly exercise. Producing high-quality, scientifically sound, practical, clinically and socially acceptable, and continuously updated recommendations requires special skills and standardized procedures. Because of the size and complexity of the task, guideline development is best coordinated by national or international organizations specialized in the field. There are numerous unresolved methodologic issues—such as the need for designing and conducting better outcome and diagnostic accuracy studies, heterogeneity and bias in systematic reviews, and lack of a unified evidence-grading system—that limit the practice of evidence-based laboratory medicine and thus the development of evidence-based guidelines. To overcome these shortcomings, international collaboration and harmonization of the numerous approaches of the many national guideline development organizations are needed.

We are grateful to Profs. P.M. Bossuyt and T. Kawai for comments during the preparation of the manuscript. This work was supported by the Education and Management Division of the IFCC

References

1. Haynes RB, Devereaux PJ, Guyatt GH. Physicians' and patients' choices and evidence-based practice. BMJ 2002;324:1350.
2. Field MJ, Lohr KN, eds. Guidelines for clinical practice. From development to use. Washington, DC: National Academy Press, 1992:426pp.
3. Scottish Intercollegiate Guidelines Network. SIGN 50: a guideline developers' handbook (February 2001; last updated October 2002). www.show.scot.nhs.uk/sign/guidelines/fulltext/50/index.html (accessed August 2003).
4. Larkin M. Noncompliance with "best practices" may be justifiable. Lancet 2001;358:1433.
5. McQueen MJ. Overview of evidence-based medicine: challenges for evidence-based laboratory medicine. Clin Chem 2001;47:1536–46.
6. 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. Clin Chem 2003;49:7–18.
7. AGREE: guideline development in Europe: an international comparison. Int J Technol Assess Health Care 2000;16:1039–49.
8. Hurwitz B. Legal and political considerations of clinical practice guidelines. BMJ 1999;318:661–4.
9. Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J. Potential benefits, limitations, and harms of clinical guidelines. BMJ 1999;318:527–30.
10. Watine J. Are laboratory investigations recommended in current medical practice guidelines supported by available evidence? Clin Chem Lab Med 2002;40:252–5.
11. Lewis SJ. Further disquiet on the guidelines front. Can Med Assoc J 2001;165:180–1.
12. Williams JG. Guidelines for clinical guidelines should distinguish between national and local production. BMJ 1999;318:942.
13. Bossuyt PMM, Lijmer JG, Mol BWJ. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844–7.
14. Dufour DR, ed. Laboratory guidelines for screening, diagnosis and monitoring of hepatic injury. Washington: National Academy of Clinical Biochemistry, 2002:75pp. (http://www.nacb.org/lmpg/hepatic_lmpg.stm).
15. Price CP. Evidence-based laboratory medicine. Clin Biochem Rev 2001;22:99–109.
16. Beishuizen A, Van Lijf JH, Lekkerkerker JFF, Vermes I. The low dose (1 μg) ACTH stimulation test for assessment of the hypothalamo-pituitary-adrenal axis. Neth J Med 2000;56:91–9.
17. Lamey PJ, Lamb AB. Prospective study of aetiological factors in burning mouth syndrome. BMJ 1988;296:1243–6.
18. Woolf SH. Practice guidelines, a new reality in medicine. II: methods of developing guidelines. Arch Intern Med 1992;152:946–52.
19. Jackson R, Feder G. Guidelines for clinical guidelines. BMJ 1998;317:427–8.
20. National Health and Medical Research Council (NHMRC). A guide to the development, implementation and evaluation of clinical practice guidelines. Canberra, Australia. NHMRC; 1998. http://www.health.gov.au/nhmrc/ (accessed January 2004).
21. Eccles M, Mason J. How to develop cost-conscious guidelines. Health Technol Assess 2001;5:1–69.
22. Bruns DE, Oosterhuis WP. From evidence to guidelines. In: Price CP, Christenson R, eds. Evidence-based laboratory medicine—from principles to outcomes. Washington: AACC Press, 2003:187–207.
23. Field MJ, ed. Setting priorities for clinical practice guidelines. Washington, DC: National Academy Press, 1995:176pp.
24. Black N, Donald A. Evidence based policy: proceed with care. BMJ 2001;323:275–9.
25. Petticrew M. Systematic reviews from astronomy to zoology: myths and misconceptions BMJ 2001;322:98–101.
26. Farmer A. Medical practice guidelines: lessons from the United States. BMJ 1993;307:313–7.
27. Bissell MG. Laboratory-related measures of patient outcomes: an introduction. Washington: AACC Press, 2000:193pp.
28. Bruns DE. Laboratory-related outcomes in health care. Clin Chem 2001;47:1547–52.
29. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. Can Med Assoc J 1986;134:587–94.
30. Fielding AM, Powell A. Using Medline to achieve an evidence-

based approach to diagnostic clinical biochemistry. Ann Clin Biochem 2002;39:345–50.
31. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1994;1:447–58.
32. AGREE Collaboration. Appraisal of Guidelines for Research & Evaluation (AGREE) instrument. September 2001. http://www.agreecollaboration.org (accessed January 2004).
33. Knottnerus JA, van Weel C, Muris JWM. Evaluation of diagnostic procedures. BMJ 2002;324:477–80.
34. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Books, 2002: 39–59.
35. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002;324:539–41.
36. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645–51.
37. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Meulen van der JHP et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061–6.
38. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SE, et al. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment Number 47. AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality, 2002. http://www.ahrq.gov (accessed January 2004).
39. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews [Abstract]. BMC Med Res Methodol 2003;3:25 (http://www.biomedcentral.com/1471-2288/3/25/abstract).
40. 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. Clin Chem 2003;49:1–6.
41. Battaglia M, Bucher H, Egger M, Mbattagkia H, Bucher M, Egger F, et al. (writing committee). The Bayes Library of Diagnostic Studies and Reviews, 2nd ed. Basel, Switzerland: Division of Clinical Epidemiology and Biostatistics, Institute of Social and Preventive Medicine, University of Berne and Basel Institute for Clinical Epidemiology, University of Basel, 2002. (Available from daniel.pewsner@bluewin.ch).
42. Lohr KN, Carey TS. Assessing "best evidence": issues in grading the quality of studies for systematic reviews. Jt Comm J Qual Improv 1999;25:470–9.
43. Oosterhuis WP, Niessen RW, Bossuyt PM. The science of systematic reviewing studies of diagnostic tests. Clin Chem Lab Med 2000;38:577–88.
44. Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests. Recommended methods, updated 6 June 1996. www.cochrane.org/cochrane/sadtdoc1.htm (accessed January 2004).
45. National Health and Medical Research Council (NHMRC). How to review the evidence: systematic identification and review of the scientific literature. Canberra, Australia: NHMRC, 2000. http://www.health.gov.au/nhmrc/ (accessed January 2004).
46. Khan KS, Ter Riet G, Glanville J, Eowden AJ, Kleijnen J. Undertaking systematic reviews of research on effectiveness. CRD's guidance for carrying out or commissioning reviews: York, England: University of York, NHS Centre for Reviews and Dissemination, 2000.
47. Harbour R, Miller J. A new system [Scottish Intercollegiate Guidelines Network (SIGN)] for grading recommendations in evidence based guidelines. BMJ 2001;323:334–6.
48. Scottish Intercollegiate Guidelines Network (SIGN). Grading system for recommendations in evidence-based clinical guidelines. Report of a review of the system for grading recommendations in SIGN guidelines. Edinburgh: Scottish Intercollegiate Guidelines Network, 2000. www.sign.ac.uk/ (accessed January 2004).
49. Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CF, Askham J, et al. Consensus development methods, and their use in clinical guideline development. Health Technol Assess 1998; 2:1–65.
50. Anonymous. Getting evidence into practice. Eff Health Care Bull 1999;5:1–16.
51. Anonymous. [Guideline development within the Institute for Quality for Healthcare CBO. Guideline for working group members]. Utrecht: Institute for Quality for Healthcare CBO, 2000 [In Dutch].