Conflict between Guideline Methodologic Quality and Recommendation Validity: A Potential Problem for Practitioners

JOSEPH WATINE,1*† BRUNO FRIEDBERG,2 EVA NAGY,3 RITA ONSODY,3 WYTZE OOSTERHUIS,4‡ PETER S. BUNTING,5† JEAN-CHRISTOPHE CHARET,6 and ANDREA RITA HORVATH3§

Background: It is not clear if good methodologic quality in current practice guidelines necessarily leads to more valid recommendations, i.e., those that are supported with consistent research evidence or, when evidence is conflicting or lacking, with sufficient consensus among the guideline development team. To help clarify this issue, we assessed whether there is a link between methodologic quality and recommendation validity in practice guidelines for the use of laboratory tests in the management of patients with non-small cell lung cancer (NSCLC).

Methods: We conducted a systematic review of data on laboratory tests in NSCLC published in English or in French within the last 10 years and retrieved 11 practice guidelines for the use of these tests. The guidelines were critically appraised and scored for methodologic quality and recommendation validity based on the Appraisal of Guidelines Research and Evaluation (AGREE) criteria and on the systematic review.

Results: Overall, these 11 guidelines had considerable shortcomings in methodologic quality and, to a lesser extent, in recommendation validity. Practice guidelines with the best methodologic quality were not necessarily the most valid in their recommendations, and conversely.

Conclusions: Poor methodologic quality and lack of recommendation validity in laboratory medicine call for methodologic standards of guideline development and for international collaboration of guideline development agencies. We advise readers of guidelines to critically evaluate the methods used as well as the content of the recommendations before adopting them for use in practice.

© 2006 American Association for Clinical Chemistry The methodologic quality of current practice guidelines must be improved (1–20). For many reasons, the methodologic quality of diagnostic guidelines is poorer than that of therapeutic guidelines, particularly in the field of laboratory medicine (10, 21, 22). It is not clear, however, whether these shortcomings in the methodology of guideline development necessarily lead to invalid recommendations, i.e., those that are not supported with consistent research evidence or sufficient consensus among the guideline development team when evidence is conflicting or lacking. We assessed to what extent methodologic quality is linked to recommendation validity in practice guidelines for the use of laboratory tests in the management of patients with non-small cell lung cancer.
were selected for appraisal laboratory tests in the management of NSCLC patients the last 10 years that provided advice for the use of practice guidelines published in English or French within new drugs and new therapeutic schemes possible are needed to assess the potential contribution of fore, controlled trials that include as many patients as facilities, no standard therapeutic schemes exist; there- overwhelming majority of patients relapse. In many treatment mostly aim to prolong patient survival, and the over- therapy with radiotherapy are options, but these therapies to only a small minority of patients. In more advanced chance of long-term survival, but this option is available or IIIA to a lesser extent) offers patients a reasonable anatomic) pathology tests.

Materials and Methods
SEARCH FOR AND SELECTION OF GUIDELINES
The general principles of our manual and electronic search strategies have been described previously (21). All practice guidelines published in English or French within the last 10 years that provided advice for the use of laboratory tests in the management of NSCLC patients were selected for appraisal (29–39).

RECOMMENDATIONS IN PRACTICE GUIDELINES
Two of us (J.W. and B.F.) extracted all laboratory-related recommendations from the 11 guidelines selected for review (Table 1). In Table 1, the term “unclear recommenda- tion” indicates either that the clinical decisions to be made based on the results of the recommended laboratory investigations were not precisely specified or that the names of the recommended laboratory tests themselves were not specified (e.g., the general term “biochemistry tests” was used).

Four organizations, the CIGNA HealthCare Medicare Administration (CIGNA), the European Group on Tumor Markers (EGTM), the National Academy of Clinical Biochemistry (NACB), and the Société de Pneumologie de Langue Française (SPLF), focused on tumor markers only. SPLF classifies 2 tumor markers in NSCLC, carcinoembryonic antigen (CEA) and cyfra 21-1, according to their “levels of scientific evidence”. SPLF considers the CEA level of scientific evidence as “not sufficient” for prognos- sis, staging, or surveillan ce, whereas the cyfra 21-1 level of evidence is considered sufficient for prognosis but not for staging or surveillan ce. EGTM recommends the measurement of cyfra 21-1 before therapy and during posttherapy follow-up in NSCLC patients, and of CEA in cases of adenocarcinoma or large cell carcinoma. EGTM also stresses the indepen- dent prognostic value of cyfra 21-1, CEA, and CA 125 and of cyfra 21-1, CEA, and tissue-polypeptide antigen for monitoring therapy efficacy in NSCLC.

The NACB guidelines are very similar to those of EGTM except that they do not as clearly recommend that cyfra 21-1 or CEA be measured in NSCLC patients. CIGNA does not recommend the measurement of CEA, neuron-specific enolase (NSE), or cyfra 21-1.

The 7 other organizations, the American College of Chest Physicians (ACCP), Agence Nationale pour le Développement de l’Evaluation Médicale (ANDEM), American Society of Clinical Oncology (ASCO), American Thoracic Society, and European Respiratory Society (ATS-ERS), British Thoracic Society and Society of Cardiotho- racic Surgeons of Great Britain and Ireland (BTS-SCG), Fédération Nationale des Centres de Lutte Contre le Cancer (FNCLCC), and the Scottish Intercollegiate Guide- lines Network (SIGN) recommend the measurement of several different laboratory variables for the pretreatment evaluation of NSCLC patients (Table 1). ATS-ERS stresses the pretreatment prognostic significance of serum albu- min and, to a lesser extent, that of serum calcium, partic- ularly in case of advanced disease, whereas SIGN stresses the pretreatment prognostic significance of calcium, alkali- line phosphatase (ALP), and liver function tests, and ASCO stresses the importance of lactate dehydrogenase (LD), hemoglobin, and leukocyte counts. ANDEM, ATS- ERS, and FNCLCC do not recommend the routine mea- surement of any laboratory variables other than those mentioned in Table 1, including serum tumor markers, in NSCLC patients.

SYSTEMATIC REVIEW OF THE EVIDENCE
We previously carried out a systematic review of the evidence (40, 41), which we updated last year (42). Rec- ommendations in the 11 guidelines about the use of laboratory tests in NSCLC (Table 1) were compared with the findings of our systematic reviews (40–42).

In the management of NSCLC patients, laboratory tests can be useful in relation either to the disease itself or to the therapies administered. In relation to the therapies ad- ministered, if “routine chemistries and hematological tests” were to be taken into account, as summarized in

Table 1. Recommendations made in 11 clinical practice guidelines providing advice for the use of laboratory variables in the pretreatment management of NSCLC patients (29–39).

<table>
<thead>
<tr>
<th>Guidelines</th>
<th>Recommended</th>
<th>Unclear recommendation</th>
<th>Not recommended</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCP</td>
<td>Hematocrit, ALP, calcium, electrolytes, glucose, GGT, SGOT</td>
<td>Other routine laboratory tests</td>
<td>None</td>
</tr>
<tr>
<td>ANDEM</td>
<td>Leukocyte count, albumin, SR, calcium, ALP, LD</td>
<td>None</td>
<td>Tumor markers</td>
</tr>
<tr>
<td>ASCO</td>
<td>Hemoglobin, leukocyte counts, LD, ALP, calcium</td>
<td>Other routine chemistries; liver function tests</td>
<td>LASA, CA 19-9, DNA index, DNA flow cytometric proliferation analysis, p53 tumor suppressor gene, ras oncogene</td>
</tr>
<tr>
<td>ATS-ERS</td>
<td>Blood counts, electrolytes, albumin, calcium, ALP, transaminases, bilirubin, creatinine</td>
<td>None</td>
<td>Tumor markers</td>
</tr>
<tr>
<td>BTS-SCG</td>
<td>Albumin, creatinine, glucose</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>CIGNA</td>
<td>None</td>
<td>None</td>
<td>CEA, NSE, cyfra 21-1</td>
</tr>
<tr>
<td>EGTM</td>
<td>cyfra 21-1, CEA</td>
<td>CA 125, TPA</td>
<td>None</td>
</tr>
<tr>
<td>FNCLCC</td>
<td>Hemoglobin, leukocyte counts with differential, LD, albumin, calcium</td>
<td>None</td>
<td>Tumor markers</td>
</tr>
<tr>
<td>NACB</td>
<td>None</td>
<td>cyfra 21-1, CEA, NSE</td>
<td>None</td>
</tr>
<tr>
<td>SIGN</td>
<td>ALP, calcium</td>
<td>Other biochemistry and hematology tests; liver function tests</td>
<td>None</td>
</tr>
<tr>
<td>SPLF</td>
<td>None</td>
<td>cyfra 21-1</td>
<td>CEA</td>
</tr>
</tbody>
</table>

*a GGT, γ-glutamyl transferase; SGOT, glutamic-oxaloacetic transpeptidase; SR, sedimentation rate; LASA, lipid-associated sialic acid; TPA, tissue-polypeptide antigen.

b Guidelines intended for tumor markers only.

c Only in cases of adenocarcinoma or large cell carcinoma.

Table 1, virtually all 7 practice guidelines dealing with nontumor markers (ACCP, ANDEM, ASCO, ATS-ERS, BTS-SCG, FNCLCC, and SIGN) would probably agree that to evaluate toxicity or tolerance to the therapies administered to NSCLC patients, it may be necessary to measure hemoglobin, leukocyte counts, platelets, electrolytes, glucose, creatinine, transaminases, bilirubin, and albumin. On the basis of such a consensus, we have therefore considered that it is valid to recommend the measurements of these variables in NSCLC patients, particularly in patients suffering from advanced stages of disease.

In relation to the disease itself, almost all authors agree that laboratory tests (excluding pathology tests) currently have no clinical utility for NSCLC screening or diagnosis. In addition, 2 prognostic covariables are universally used in NSCLC patients: disease stage and performance status (23–25). Among other prognostic covariables that can be used for the stratification of NSCLC patients in trials, some authors use patient age and sex. Our systematic reviews of the evidence (40–42) indicated that the pretreatment prognostic values of blood hemoglobin, leukocyte counts with differential, serum LD, albumin, calcium, and, to a lesser extent, NSE are very likely to be independent of the aforementioned other covariables. There thus is sufficient evidence for recommending the measurement of at least all of these laboratory variables in all NSCLC patients participating in therapeutic trials (42). We also consider it valid, based on the evidence, to recommend the measurement of blood hemoglobin in patients treated with radiotherapy (either inside or outside therapeutic trials), because the outcomes of patients with low blood hemoglobin are very likely to improve if they receive erythropoietin before radiation therapy (42, 43).

These guidelines also suggest that laboratory tests might be useful for the staging (pretreatment prognostic evaluation) and surveillance (posttreatment prognostic evaluation) of NSCLC. Seven guidelines (ACCP, ANDEM, ASCO, ATS-ERS, BTS-SCG, FNCLCC, and SIGN) thus recommend the use of diverse biochemistry and/or hematologic tests for the staging of NSCLC. Abnormal result(s) in a patient with otherwise resectable (stage IIIA, or lower) NSCLC might indicate the presence of metastases and unresectable disease; therefore, all putative metastatic sites must be carefully investigated. All of the guidelines do not agree, however, on which laboratory tests are necessary under these circumstances. Some recommend the use of calcium and/or albumin, whereas others recommend ALP and/or LD (Table 1). Other guidelines (e.g., those of EGTM, NACB, or SPLF to a lesser extent) extend this recommendation to cyfra 21-1 (44). According to our own systematic review, however, there is more evidence to support the use of routine biochemical and hematologic tests (i.e., leukocyte counts with differential, LD, albumin, and calcium), or even to a lesser extent NSE, than cyfra 21-1, thus confirming that some new laboratory tests (e.g., cyfra 21-1 or other tumor markers for the management of NSCLC) may be introduced into routine practice before they are demonstrated to have greater validity than older, less expensive tests.
We therefore consider there to be sufficient evidence for recommending the measurement of leukocyte counts with differential, serum LD, albumin, and calcium, but that it is not valid to recommend the measurement of tumor markers in NSCLC patients, except perhaps for NSE in patients in chemotherapy trials.

In summary, available evidence suggests that the laboratory tests indicated in Table 2 should be performed for the pretreatment evaluation of NSCLC patients. Some laboratory tests recommended by one or several experts in the 11 practice guidelines are not mentioned in Table 2, e.g., ALP for the staging of NSCLC, because according to the systematic review of the evidence, it is quite clear that the prognostic value of ALP is inferior to, and is not independent of, that of LD, as summarized in Table 3.

APPRaisal OF GUIDELINES

Scores for methodologic quality were assigned to each of the 11 guidelines (29–39), based on their critical appraisal using the Appraisal of Guidelines Research and Evaluation (AGREE) Instrument. The AGREE Instrument comprises 23 criteria, arranged in 6 domains (as shown in Table 4), covering the key elements of the guideline development process (47). Among the published appraisal checklists for practice guidelines, we chose the AGREE Instrument because it has shown the greatest potential as a tool for assessing recommendations for clinical pathways (48). This instrument has been endorsed by the WHO and the European Commission (21). In accordance with the AGREE recommendations, we assigned to each guideline 1 of 4 possible overall final scores: “very good” (the equivalent of “strongly recommend” of the AGREE Instrument) if the guideline rated high on the majority of items and most domain scores were >60%, indicating that the guideline had a high overall quality and could be considered for use in practice without alterations; “good” (the equivalent of “recommend with provisos or alterations” of the AGREE Instrument) if the guideline rated high or low on a similar number of items and most domain scores were 30%–60%, indicating that the guideline had a moderate overall quality; “not so good” (the equivalent of “would not recommend” of the AGREE Instrument) if the guideline rated low on the majority of items and most domain scores were <30%, indicating that the guideline had a low overall quality and serious shortcomings and thus should not be recommended for use in practice; and finally, “dubious” (the equivalent of “unsure” of the AGREE Instrument) if the guideline did not give sufficient information to enable us to assess its quality. We have chosen to use the scores “very good”, “good”, “not so good”, or “dubious”, rather than the original terminology of the AGREE Instrument, as indicated above, because we thought that this would lead to an easier understanding of our review.

Scores for validity of recommendations were also assigned to each guideline, based on a systematic review of the evidence (40), which has been updated twice (41, 42), also taking into account the consensual opinions of the guideline development teams, as summarized above in the section Systematic Review of the Evidence. The scale of ratings that we used was the same as for methodologic quality, consisting of 4 possible scores (very good, good, not so good, or dubious), as explained in more detail in the Results section. Two scores for validity of recommendations were assigned to each guideline: one for recommendations regarding tumor markers and one for recommendations regarding other laboratory tests.

HOW DISAGREEMENTS WERE RESOLVED

During the whole process of the study described in the 4 sections above (Search for and Selection of Guidelines, Recommendations in Practice Guidelines, Systematic Review of the Evidence, and Appraisal of Guidelines), disagreements between the 2 assessors (J.W. and B.F.) were resolved by

### Table 2. Laboratory variables that should be measured for the pretreatment evaluation of NSCLC patients, taking into account both the systematic review of the evidence and the consensual opinions of the development teams of current practice guidelines (29–39).

<table>
<thead>
<tr>
<th>Purpose of test</th>
<th>Variables to be measured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation of toxicity (or tolerance) to treatments</td>
<td>In all patients: Hemoglobin,(^a) leukocyte counts with differential,(^a) platelets,(^a) electrolytes,(^a) glucose,(^a) creatinine,(^a) transaminases,(^a) bilirubin,(^a) albumin(^a)</td>
</tr>
<tr>
<td>Pretreatment prognostic evaluation</td>
<td>In all patients: Hemoglobin (if radiation therapy),(^b) leukocyte counts with differential,(^a) LD,(^a) albumin,(^a) calcium(^a)</td>
</tr>
<tr>
<td>Pretreatment prognostic evaluation</td>
<td>In patients participating in therapeutic trials: Hemoglobin,(^b) leukocyte counts with differential,(^b) LD,(^a) albumin,(^b) calcium,(^b) NSE(^b)</td>
</tr>
</tbody>
</table>

\(^a\) Recommendation validated by most experts and not contradicted by systematic review of the evidence.

\(^b\) Recommendation based on systematic review of the evidence.

### Table 3. Pretreatment prognostic value of LD vs that of ALP in NSCLC patients.

<table>
<thead>
<tr>
<th></th>
<th>Not significant</th>
<th>Unspecified significance(^a)</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD(^c)</td>
<td>16 (2450)</td>
<td>1 (189)</td>
<td>9 (4900)</td>
</tr>
<tr>
<td>ALP</td>
<td>18 (7252)</td>
<td>0</td>
<td>1 (207)(^d)</td>
</tr>
</tbody>
</table>

\(^a\) Number of studies (number of patients included in the studies).

\(^b\) Significant in multivariate analysis that did not take into account performance status or disease stage.

\(^c\) The prognostic value of LD seems to be more obvious in unselected patients than in resected patients.

\(^d\) In this study, LD was not included in the multivariate analysis, whereas all studies that compared the prognostic value of LD with that of ALP concluded that LD performed better than ALP [for explanations and more details, see Refs. (40–42)].
consensus, and if necessary, a third person (J.C.C.) was available as a referee (this was never necessary). For methodologic quality, the consensual scores thus obtained were validated by an independent set of assessors (E.N. and R.O.), and when necessary a third person (A.R.H.) was used as a referee (this was necessary only once).

Results

Scores for Methodologic Quality

Each guideline was scored for each of the 6 domains of the AGREE Instrument, as shown in Table 4. As can be seen in Table 4, some guidelines are better in some aspects and others in other aspects of guideline development methodology. The overall final scores obtained for each guideline are indicated in Table 5.

Scores for Validity of Recommendations

The only guideline that clearly recommended the use of tumor markers (EGTM) was scored as not so good because there is no evidence that measurement of tumor markers in routine practice would improve NSCLC patient outcomes. The EGTM guideline was not attributed the worst possible score (dubious) because, as already stressed, tumor markers might be useful in therapeutic trials. The 4 guidelines that clearly did not recommend the use of tumor markers in routine practice (ANDEM, ATS-ERS, CIGNA, and FNCLCC) were scored as good. These 4 guidelines were not attributed the best possible score (very good) because they did not allude to the needs of patients in therapeutic trials. Guidelines that gave unclear recommendations regarding the use of tumor markers (NACB and SPLF) or that did not mention tumor markers at all (ACCP, ASCO, BTS-SCG, and SIGN) were also scored as not so good (Table 5).

Regarding the other laboratory tests, the 5 guidelines (ACCP, ANDEM, ASCO, FNCLCC, and ATS-ERS) in which only a few laboratory tests were missing among those recommended (compared with the reference list of tests in Table 2) were scored as good. SIGN and BTS-SCG guidelines were scored as not so good because their lists of recommended laboratory tests were clearly less evidence based than those in the 5 other guidelines (as can be seen in Table 1).

Discussion and Conclusions

It is universally acknowledged that the methodologic quality of practice guidelines must improve (1–22). For example, in a study that was probably one of the largest surveys ever done on the subject, Grilli et al. (7) assessed the quality of 431 practice guidelines produced by specialty societies and published in English-language

<table>
<thead>
<tr>
<th>Guidelines</th>
<th>Scope and purpose</th>
<th>Stakeholder involvement</th>
<th>Rigor of development</th>
<th>Clarity and presentation</th>
<th>Applicability</th>
<th>Editorial independence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCP</td>
<td>61</td>
<td>46</td>
<td>60</td>
<td>46</td>
<td>6</td>
<td>75</td>
</tr>
<tr>
<td>ANDEM</td>
<td>89</td>
<td>25</td>
<td>10</td>
<td>71</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>ASCO</td>
<td>94</td>
<td>50</td>
<td>71</td>
<td>67</td>
<td>17</td>
<td>75</td>
</tr>
<tr>
<td>ATS-ERS</td>
<td>44</td>
<td>4</td>
<td>5</td>
<td>29</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>BTS-SCG</td>
<td>100</td>
<td>33</td>
<td>60</td>
<td>79</td>
<td>6</td>
<td>83</td>
</tr>
<tr>
<td>CIGNA</td>
<td>67</td>
<td>13</td>
<td>12</td>
<td>54</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>EGTM</td>
<td>44</td>
<td>4</td>
<td>2</td>
<td>29</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FNCLCC</td>
<td>94</td>
<td>54</td>
<td>57</td>
<td>79</td>
<td>17</td>
<td>33</td>
</tr>
<tr>
<td>NACB</td>
<td>50</td>
<td>17</td>
<td>29</td>
<td>54</td>
<td>11</td>
<td>25</td>
</tr>
<tr>
<td>SIGN</td>
<td>89</td>
<td>75</td>
<td>76</td>
<td>75</td>
<td>33</td>
<td>25</td>
</tr>
<tr>
<td>SPLF</td>
<td>61</td>
<td>46</td>
<td>48</td>
<td>38</td>
<td>17</td>
<td>8</td>
</tr>
</tbody>
</table>

* Maximum score for best quality = 100%.
* Guidelines intended for tumor markers only.

<table>
<thead>
<tr>
<th>Guidelines</th>
<th>Methodologic quality</th>
<th>Tumor markers</th>
<th>Other laboratory tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCP</td>
<td>Good</td>
<td>Not so good</td>
<td>Good</td>
</tr>
<tr>
<td>ANDEM</td>
<td>Not so good</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>ASCO</td>
<td>Good</td>
<td>Not so good</td>
<td>Good</td>
</tr>
<tr>
<td>ATS-ERS</td>
<td>Dubious</td>
<td>Not so good</td>
<td>Not so good</td>
</tr>
<tr>
<td>BTS-SCG</td>
<td>Good</td>
<td>Not so good</td>
<td>Not so good</td>
</tr>
<tr>
<td>CIGNA</td>
<td>Dubious</td>
<td>Not so good</td>
<td>Not so good</td>
</tr>
<tr>
<td>EGTM</td>
<td>Dubious</td>
<td>Not so good</td>
<td>Not so good</td>
</tr>
<tr>
<td>FNCLCC</td>
<td>Good</td>
<td>Not so good</td>
<td>Good</td>
</tr>
<tr>
<td>NACB</td>
<td>Not so good</td>
<td>Not so good</td>
<td>Not so good</td>
</tr>
<tr>
<td>SIGN</td>
<td>Good</td>
<td>Not so good</td>
<td>Not so good</td>
</tr>
<tr>
<td>SPLF</td>
<td>Not so good</td>
<td>Not so good</td>
<td>Not so good</td>
</tr>
</tbody>
</table>

* The best possible score (i.e., very good, the equivalent of strongly recommend of the AGREE instrument) was never attributed.
* Guidelines published less than 3 years ago.
* Guidelines published more than 5 years ago.
* Diagnostic guidelines (the 4 other guidelines provide both diagnostic and therapeutic recommendations).
The process of many factors that may influence the translation of research findings to the context of use. The strategy to identify primary evidence; and an explicit grading of recommendations according to the quality of supporting evidence (7). These 3 criteria were not met at all in 67%, 88%, and 82% of these 431 guidelines, respectively, and all 3 criteria for quality were met in only 22 guidelines (5%) (7). It is not clear whether this situation will improve in the near future (49).

When we used the AGREE Instrument to assess the methodologic quality of 11 practice guidelines providing advice for the use of laboratory tests for the management of NSCLC, many fell short of basic quality criteria, a result that confirms the aforementioned observations (1–22).

Regarding our judgment that some of the recommendations made in the 11 guidelines are not entirely valid, it could be argued that our judgment might be incorrect because even evidence-based guidelines can contain different recommendations. Scientific evidence is only one of many factors that may influence the translation of research findings to the context of use. The process of considered judgment is essential in guideline development and often requires extensive discussions and consensus among experts (50). Availability of services, resources, and cost-effectiveness are important considerations (22) but do not apply to our study (Table 1). Recommendations about tumor markers are conflicting, and availability of services, resources, and cost-effectiveness do not support the use of any test that lacks evidence of clinical utility. In addition, some recommendations made in these 11 guidelines are unclear. Regarding these unclear recommendations, the 3 guidelines (ACCP, ASCO, and SIGN) in which the measurement of “other (routine) laboratory tests” was recommended (quite a vague recommendation indeed) offered both therapeutic and diagnostic recommendations; it is therefore possible that guideline development teams followed the right guideline development methods regarding therapeutic recommendations but did less well when formulating diagnostic ones. The AGREE Instrument, as a generic appraisal toolbox, did not allow us to investigate this possibility in more depth, and it was not possible for us to conclude whether the methodologic quality or validity of recommendations was better or worse in the 4 guidelines that offer both therapeutic and diagnostic recommendations than in the 7 other, purely diagnostic, guidelines, although the 3 guidelines that obtained the lowest possible scores regarding methodologic quality (dubious) were purely diagnostic guidelines (Table 5).

In summary, the clinical validity of the recommendations regarding laboratory tests made in some of the 11 guidelines can be questioned (at least in those scoring not so good in Table 5). Some authors have shown that guidelines of poor methodologic quality are more likely to provide invalid diagnostic recommendations than guidelines of high methodologic quality (10, 18), but other authors have shown that guidelines of poor methodologic quality can provide diagnostic recommendations as valid as guidelines with high methodologic quality (2). In our study, practice guidelines with the best methodologic quality were not necessarily those that were the most valid in the content of their recommendations, and conversely. For example, ATS-ERS and CIGNA guidelines were valid in their recommendations, whereas their methodologic quality was poor, and SIGN and BTS-SCG guidelines were not valid in their recommendations, whereas their methodologic quality was good (Table 5). Because the AGREE Instrument does not involve the use of global scores to assess the methodologic quality of guidelines, we also looked at the individual scores obtained not only in each of the 6 domains (Table 4) but also in each of the 23 questions (data not shown), and again we were not able to establish any correlation between validity of content and methodologic quality in any of the 6 domains or in any of the 23 questions.

This result is worrisome because the busy practitioner confronted with conflicting guideline recommendations has no easy means to help in deciding which guideline should be trusted. Conflicting recommendations on the use of laboratory tests are likely to lead to a waste of laboratory resources and might even cause harm to patients (51). Effective treatment depends on the effective use of diagnostic tests, and if diagnostic recommendations are not evidence based, it is reasonable to assume that therapeutic interventions will sometimes be initiated and monitored inappropriately. Fortunately, the shortcomings in methodologic quality seemed to be more frequent than those of the content validity (see Table 5). The discrepancy between methodologic guideline quality and clinical validity of recommendations is perhaps less obvious in therapeutic recommendations, in which the quality of evidence from randomized trials is higher, than in diagnostic recommendations, in which the level of evidence is generally much poorer. Another possibility is that in other areas of medicine, more valid laboratory recommendations in practice guidelines are available than in NSCLC. Whatever the true situation is, the results of our study call for the critical appraisal of guidelines providing both diagnostic and therapeutic recommendations in various medical areas. Such work is in progress within our team in the field of diabetes mellitus (21). On the basis of our studies and reports from the literature, however, we strongly advise colleagues to do similar studies in other areas of medicine before guideline recommendations are used in local practice.

Because FNCLCC guidelines obtained the best scores in all items used for comparison (Table 5), one could argue that the French authors of the present review were biased toward guidelines in their own language. Taking into account the fact that we had the opportunity of expert discussions with some of the authors of the FNCLCC guidelines before their guidelines were published [these
discussions have partly been published (52), we rather believe that the authors of the FNCLCC guidelines are much more likely than the authors of the other guidelines [except perhaps for the authors of NACB guidelines who quoted us (37)] to have read and (partly) taken into account the results of systematic reviews available on this topic (40–42, 52).

According to Shekelle et al. (53), the point at which no more than 90% of the guidelines published by the US Agency for Healthcare Research and Quality are still valid is 3.6 years (95% confidence interval, 2.6–4.6 years). To assess this hypothesis, we checked whether there was better correlation between methodologic quality and validity of recommendations in the most recently published guidelines. We failed to find any difference (Table 5), which suggests that the hypothesis of Shekelle et al. (53) may be valid for US Agency for Healthcare Research and Quality guidelines and similar sorts of guidelines, particularly in those areas in which the evidence base of recommendations develops faster than in the field we investigated.

In conclusion, to overcome the methodologic shortcomings of current practice guidelines and to improve the validity of resulting recommendations, standardized methods for making evidence-based guideline recommendations in laboratory medicine must be disseminated. In particular, we need a unified system for grading diagnostic recommendations [such a work is in progress within the GRADE group (54)], as well as common standards for guideline reporting [such a work also seems to be in progress (55)], together with appropriate tools for guideline implementation. Finally, we need to educate our profession about the principles of evidence-based laboratory medicine and guideline development methods (22). We advise that guidelines be critically evaluated for methodology and content before recommendations are used in clinical practice.

This work was presented at the Edutratk session entitled “The Role of Evidence-Based Laboratory Medicine in Guidelines”, developed with AACC’s EBM Group (R. Christenson, Chairman) and IFCC C-EBLM (A.R. Horvath, Chairwoman), at the XIXth International Congress on Clinical Chemistry (July 25–27, 2005, Orlando, FL).

References

22. Oosterhuis WP, Bruns DE, Watine J, Sandberg S, Horvath AR.


