Clinical Chemistry 58:10
000–000 (2012)

Editorials

Room for Improvement in National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines

Patrick M.M. Bossuyt[1*]

Practice guidelines are everywhere in healthcare. They have been around for decades, but their development, dissemination, and—possibly—their use skyrocketed in the 1990s. At that time, guidelines were positioned at the crossroads of 2 developments. One was the arrival of an era of assessment and accountability in healthcare. Undesirable variation in practice, spiraling healthcare costs, and concerns about the overuse and underuse of services encouraged professionals to take action. The other was the increasing belief that clinical-practice and healthcare policy decisions should be guided by the best-available evidence, a goal increasingly more difficult to achieve at the level of the individual, given the daunting number of medical journals and the avalanche of studies reported in them.

Evidence-based practice guidelines were seen as helpful in tackling these challenges, and professional societies, government panels, and other groups increasingly began developing recommendations to assist healthcare professionals in delivering appropriate healthcare. Soon, guidelines were ubiquitous in practice and policy, and their rapid production and promulgation created a new problem: information overload.

Not only was the sheer number of guidelines problematic, but guideline documents also varied considerably in form and appearance—from brief lists of *do*s and *don't*s to lengthy documents of textbook-level completeness that summarized the state of the art in the field. This variability was a reason for concern, because the promotion of flawed guidelines could encourage or even institutionalize the delivery of ineffective, harmful, or wasteful interventions *(1)*.

In a typical "early days" process of guideline development, a group of invited experts would gather around a conference table to develop recommendations based on their understanding of the scientific evidence, their knowledge of the field, and their opinion. A formal methodology for evaluating the evidence was not always followed, and similarly absent were procedures for the consensus process. Given the vulnerability of the methods, concerns about susceptibility to bias and conflicts of interests were not far away *(2)*.

All researchers know that a good method poorly implemented can lead to bias. Inadequate concealment of allocation in randomized trials, for example, or failure to include all randomized patients in the analysis can lead to flawed results and erroneous conclusions. Inspired by successful examples in clinical research, several groups tried to identify similar methodological-quality criteria for guideline development. By 1992, the Institute of Medicine had already proposed "desirable attributes" of guidelines, but the movement gathered momentum through the work of Francoise Cluzeau and her UK colleagues. They developed an instrument to evaluate the extent to which published guidelines are systematically developed, which is, after all, their defining feature *(3)*. Their seminal work morphed into an international collaboration with the participation of more than 15 countries, predominantly European. The resulting AGREE[2] (Appraisal of Guidelines for Research and Evaluation) instrument was published in 2003; a revision was disseminated in 2009 *(4)*.

AGREE and AGREE II address 6 dimensions of guidelines: scope and purpose, stakeholder involvement, rigor of development, clarity and presentation, applicability, and editorial independence. Each domain is scored for several items, 23 in all, the criteria for which are presented in a separate manual. Standardized guideline domain scores are calculated by summing up all the scores of individual items in a domain and expressing them as a percentage of the maximum possible score for that domain. An AGREE-perfect guideline would score 100% on all 6 domains.

The AGREE Research Trust, an independent body established in 2004, now manages the interests of the AGREE project. The instrument, a manual, support tools, and a rich set of other information are available at http://www.agreetrust.org. Despite the existence of

[1] Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.
* Address correspondence to the author at: Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Room J2-127, P.O. Box 22700, 1100 DE Amsterdam, the Netherlands. Fax 0030-210-2107795553; e-mail p.m.bossuyt@amc.uva.nl.
Received July 25, 2012; accepted July 27, 2012.
Previously published online at DOI: 10.1373/clinchem.2012.192997

[2] Nonstandard abbreviations: AGREE, Appraisal of Guidelines for Research and Evaluation; NACB, National Academy of Clinical Biochemistry; LMPG, laboratory medicine practice guideline; GRADE, Grading of Recommendations Assessment, Development and Evaluation.

several other instruments, AGREE has become the worldwide standard for guideline evaluation *(5)*. Available in many languages, it is used in Europe, Canada, and Australia, but it is less used in the US *(6)*.

In this issue of *Clinical Chemistry*, a Canadian group reports on their application of the AGREE instrument to guidelines developed by the National Academy of Clinical Biochemistry (NACB) *(7)*. The NACB started producing laboratory medicine practice guidelines (LMPGs) in the early 1990s, with the first one, on nutrition, published in 1994. The guidelines are available on the NACB Web site (http://www.aacc.org/members/nacb/pages/default.aspx#), which lists 8 guidelines published in the last 5 years. The guidelines are supposed to be updated 5 years after the initial publication. Six older, archived NACB LMPGs remain in the published section of the Web site; 5 are available for download and purchase. Five other guidelines are listed as archived, of which 2 are still available for download.

The NACB guidelines are rather lengthy documents, and shorter versions have been published in specialty journals. The "Guidelines and Recommendations for Laboratory Analysis in the Diagnosis and Management of Diabetes Mellitus" is a 120-page document with key recommendations, which are summarized in a single 3-page table *(8)*.

Don-Wauchope and colleagues used AGREE II to evaluate 11 of these guidelines. Their conclusion is not very positive: "The quality of the guidelines . . . was generally poor" *(7)*. The latest guideline, on diabetes, scored >85% on all but 1 domain, but the other 10 guidelines failed to reach 50% on most, if not all, domains. The median score was 42%. For "rigor of development," only 3 of the 11 guidelines scored >50% (diabetes, cardiovascular biomarkers, point-of-care testing); 6 scored 20% or lower.

What do we conclude from these findings? The AGREE group has not revealed thresholds that would allow a classification of guidelines as "good" or "bad." The authors do not compare these scores with other laboratory medicine guidelines. To evaluate 11 practice guidelines on data of laboratory tests for non–small cell lung cancer, Watine and colleagues used AGREE (including the NACB one) and a separate systematic review of the literature *(9)*. Rigor scores varied from 2% to 76%, with a median of 48%. Nagy and colleagues used AGREE in assessing 26 guidelines for the diagnosis and monitoring of diabetes mellitus *(10)*. Yet, the NACB diabetes LMPG shows that it is possible to achieve high-quality scores in laboratory medicine practice guidelines.

The group from McMaster University that applied AGREE to the NACB guidelines (this institution now hosts the AGREE project office) points out several areas for improvement and offers concrete suggestions, such as more-explicit descriptions of the target population, the intended audience, and the identification and involvement of stakeholders. Because the rigor domain requires the most improvement, the authors suggest that all NACB LMPGs make every effort to present their methodology clearly and explicitly, with transparency and assessable external validity.

The NACB diabetes LMPG group adapted the GRADE (Grading of Recommendations Assessment, Development and Evaluation) methodology in developing their recommendations *(8, 11)*. In the last decade, the international GRADE group has invested considerable efforts in making the development of evidence-based guidelines more systematic and more transparent *(12)*. Starting from focused questions that include specification of all outcomes important for patients, the GRADE process separates a rating of the quality of the body of evidence from weighing the desirable and undesirable consequences, and the process grades recommendations as "strong" or "conditional" in favor of or against alternative management options. The GRADE process was initially developed for recommendations about therapeutic interventions, for which randomized trials often offer the evidence base. Extensions of GRADE for laboratory and other medical tests are still in development, which is why the diabetes group had to develop their own approach *(11)*. Some of the problematic areas are the absence of direct evidence for the effects of tests on patient outcomes and different concerns, such as those concerning the preanalytical, analytical, and postanalytical phases. In the absence of strong and widely accepted methods, it is not surprising that previously published guidelines have been shown to lack rigor.

The AGREE instrument has received criticism as well. Use of the instrument relies on the report, which may not reflect the process fully. AGREE evaluates the process of constructing a guideline, not the validity of its content *(9)*. It is possible to create a guideline document with solid evidence-based recommendations but with a poor process, or a well-developed guideline document with poorly supported recommendations. The AGREE definition of quality is many-layered. The validity of AGREE II was evaluated, for example, by evaluating the predictability of the AGREE items for distinguishing among guideline information of known varying quality (a process that could be viewed as a rather circular definition of quality) and through correlations with statements about endorsement and the intention to use them. Both of these judgments are affected by multiple factors *(13)*. In the Cochrane Collaboration, former quality-appraisal tools for clinical studies have been replaced by risk-of-bias tools *(14)*. The latter focus more precisely and more explicitly on

the validity of the study results. Maybe the guideline movement will see a similar shift from a multidimensional definition of quality to a specification of the validity of the recommendations in a guideline.

AGREE limitations aside, the recommendations of Don-Wauchope and colleagues to the NACB should be taken seriously. Practice guidelines can still be of help in areas of practice uncertainty or variability, where scientific evidence or a well-balanced judgmental process can indicate what is appropriate and what is not. Professional societies bear a special responsibility in developing such guidelines, through a rigorously systematic and explicitly transparent process.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *No authors declared any potential conflicts of interest.*

References

1. Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J. Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. BMJ 1999; 318:527–30.
2. Oosterhuis WP, Bruns DE, Watine J, Sandberg S, Horvath AR. Evidence-based guidelines in laboratory medicine: principles and methods. Clin Chem 2004; 50:806–18.
3. Cluzeau FA, Littlejohns P, Grimshaw JM, Feder G, Moran SE. Development and application of a generic methodology to assess the quality of clinical guidelines. Int J Qual Health Care 1999;11:21–8.
4. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. CMAJ 2010;182:E839–42.
5. Vlayen J, Aertgeerts B, Hannes K, Sermeus W, Ramaekers D. A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. Int J Qual Health Care 2005;17:235–42.
6. Cates JR, Young DN, Bowerman DS, Porter RC. An independent AGREE evaluation of the Occupational Medicine Practice Guidelines. Spine J 2006; 6:72–7.
7. Don-Wauchope AC, Sievenpiper JL, Hill SA, Iorio A. Applicability of the AGREE II instrument in evaluating the development process and quality of current National Academy of Clinical Biochemistry guidelines. Clin Chem 2012;58:XXX–XXX.
8. Sacks DB, Arnold M, Bakris GL, Bruns DE, Horvath AR, Kirkman MS, et al. Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. Clin Chem 2011;576:e1–47.
9. Watine J, Friedberg B, Nagy E, Onody R, Oosterhuis W, Bunting PS, et al. Conflict between guideline methodologic quality and recommendation validity: a potential problem for practitioners. Clin Chem 2006;52:65–72.
10. Nagy E, Watine J, Bunting PS, Onody R, Oosterhuis WP, Rogic D, et al. Do guidelines for the diagnosis and monitoring of diabetes mellitus fulfill the criteria of evidence-based guideline development? Clin Chem 2008;54:1872–82.
11. Horvath AR. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. Clin Chem 2009;55:853–5.
12. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–6.
13. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. Development of the AGREE II, part 2: assessment of validity of items and tools to support application. CMAJ 2010;182:E472–8.
14. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529–36.