COMMENTARIES

Appraising a tool for guideline appraisal (the AGREE II instrument)

Antonio L. Dansa,*, Leonila F. Dansb

a Professor, Department of Medicine, University of the Philippines, College of Medicine
b Professor, Department of Pediatrics, University of the Philippines, College of Medicine

The AGREE II instrument has 3 goals — to assess the quality of clinical practice guidelines (CPGs), to provide a methodologic strategy for the development of guidelines, and to recommend how and what information should be reported in guidelines. Of these, the main purpose seems to be to assess guideline quality (i.e., "confidence that the potential biases of guideline development have been addressed adequately and that the recommendations are both internally and externally valid, and are feasible for practice") [1]. There is no doubt that this instrument is badly needed by health care providers and policy-makers around the world, who deal with a deluge of clinical practice guidelines in their day-to-day decisions.

AGREE II assesses CPGs based on 6 domains: 1) scope and purpose, 2) stakeholder involvement, 3) rigor of development, 4) clarity and presentation, 5) applicability, and 6) editorial independence. Each domain is scored using several items. There are a total of 23 items in all - to be scored on a 7-point Likert scale by at least 2 (preferably 4) independent observers. This is not a run of the mill appraisal guide, as users may take an average of 90 minutes to complete an evaluation.

To learn more about the instrument, we decided to use it to evaluate Philippine Guidelines on Periodic Health Examinations [2]. Because we were personally involved in drafting these guidelines, the process took only 53 minutes. Below, we reflect on our experience in using AGREE II:

1. Overall, the 23 items in 6 domains were extremely useful in evaluating important attributes of our guidelines. Methodologic Rigor, Applicability, and Editorial Independence were particularly useful domains that helped us identify areas that need to be strengthened in subsequent revisions. Clinical specialists, for example, were involved in making guidelines for tests and examinations, without recognizing (or disclosing) that they had financial stakes from these recommendations.

2. Scoring of items was sometimes confusing. Many criteria seemed to merely require items to be described well, rather than actually be performed well. Domain 1 is a good example. The 3 items in this domain require descriptions of the general objectives, health questions covered, and population addressed. The items do not explicitly require that the questions be constructed in compliance with scientific norms, that is, addressing all relevant patient groups (like higher risk and lower risk), management options (for example, surgical, medical, and a “do nothing” option), and possible outcomes (including harms and benefits) [3]. This problem led to some confusion regarding the main purpose of the evaluation. If the goal were to evaluate adequacy of reporting, then indeed, a statement of objectives would be enough. If the goal were to evaluate quality of guidelines, however, then requirements should include more than just descriptions of what was done. Perhaps one problem of AGREE II is that it lays claim to too many functions. Indeed, inclusion of domain 4 (clarity and presentation) as a measure of quality goes against the authors’ own definition of quality, which has more to do with internal and external validity rather than clarity of language.

3. Using the instrument is a tedious process, especially for those who may be less familiar with the 23 items that have been identified. The effort it takes to score the 6 domains may be justified by the need to compare several CPGs. However, the value of the score is less clear when one is evaluating a single CPG (as we were doing). Our guidelines scored 100%, 61%, 61%, 83%, 46%, and 42%, in the 6 domains respectively, but the numbers do not help us assess acceptability of the recommendations themselves. In this situation, it may be more useful to come up with a qualitative estimate of the direction and magnitude of bias.

We also had minor disagreements with how the authors assigned some items to domains, and how some criteria for scoring were described. A major strength of AGREE II, however, is that the authors continue to subject it to a thorough and systematic evaluation, which ensures that it will evolve in a direction that is most useful for actual users. As the tool evolves, we propose 4 areas of improvement that the authors may wish to consider.

* Corresponding author.
E-mail address: tdans@skybroadband.com.ph (A.L. Dans).

0895-4356/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2010.06.005
1. Strength of recommendation — in the present tool, strength of recommendation is just listed under “other considerations” of domain 4 (clarity). Here, it is considered of equal importance with other items that merely have to do with clarity. However, classifying the strength of recommendation has evolved into a rigorous process in the past decade [4], becoming an important and useful part of many CPGs. We believe it deserves to be a separate item, probably under the domain of methodologic rigor (Domain 3).

2. Explicit statement of values and preferences used in weighing outcomes — while this is partially addressed by the stakeholder involvement (domain 2), there is no perfect method for assuring that patients values and preferences are used in making decisions. In many instances preferences are crucial (for example, when the quality of evidence is poor, or whenever there is close balance between benefits and harms). In these situations, the judgments made in weighing outcomes should be explicitly stated.

3. Applicability — While applicability is a separate domain in AGREE II, the main focus of this domain is availability of resources for implementation of recommendations. However, applicability may also be related to external validity of the evidence on effectiveness, i.e., what works in a selected group of patients in a clinical trial will not necessarily work in other groups of patients in the real world [5]. Applicability, in this sense, may not only be related to resource issues, but to sociologic issues (e.g., patient compliance) and biologic issues as well (e.g., differences between study patients and real life patients in terms of sex, co-morbidity, race, age, and pathology) [6].

4. Equity - CPG developers typically address issues on the effectiveness and efficiency of interventions they consider. Few address issues on fairness and justice, especially regarding the effect of guidelines on disadvantaged populations in society. Even when treatments are effective and efficient, a recommendation to use them may be unjust. For example, local guidelines in the Philippines mandate screening for metabolic neonatal abnormalities based on the best evidence available. Unfortunately, because of the state of the economy, implementation of these guidelines would drain resources that could otherwise be used to save lives for more common illnesses such as pneumonia and diarrhea in children. To avoid such scenarios, we have recommended that guideline developers should make sure that their recommendations do not compromise problems of higher priority among the disadvantaged. In addition, they should make sure that the evidence they use is applicable to the underprivileged, and that their values and preferences have been given special consideration. Finally, guideline developers should focus attention on planning implementation and monitoring impact among the less fortunate members of society [7].

In summary, CPGs have the potential for forming policy and even legislation. With this great power comes great responsibility. AGREE II is an important tool that puts guideline development under the microscope of public scrutiny, and makes developers take responsibility for their actions.

References