Early release, published at www.cmaj.ca on May 31, 2010. Subject to revision.

CMAJ RESEARCH

Development of the AGREE II, part 1: performance, usefulness and areas for improvement

Melissa C. Brouwers PhD, Michelle E. Kho BHSc(PT) MSc, George P. Browman MD MSc, Jako S. Burgers MD PhD, Francoise Cluzeau PhD, Gene Feder MD, Béatrice Fervers MD PhD, Ian D. Graham PhD, Steven E. Hanna PhD, Julie Makarski BSc, for the AGREE Next Steps Consortium

@@ See related research article by Brouwers and colleagues

Abstract

Background: We undertook research to improve the AGREE instrument, a tool used to evaluate guidelines. We tested a new seven-point scale, evaluated the usefulness of the original items in the instrument, investigated evidence to support shorter, tailored versions of the tool, and identified areas for improvement.

Method: We report on one component of a larger study that used a mixed design with four factors (user type, clinical topic, guideline and condition). For the analysis reported in this article, we asked participants to read a guideline and use the AGREE items to evaluate it based on a seven-point scale, to complete three outcome measures related to adoption of the guideline, and to provide feedback on the instrument's usefulness and how to improve it.

Results: Guideline developers gave lower-quality ratings than did clinicians or policy-makers. Five of six domains were significant predictors of participants' outcome measures ($p < 0.05$). All domains and items were rated as useful by stakeholders (mean scores > 4.0) with no significant differences by user type ($p > 0.05$). Internal consistency ranged between 0.64 and 0.89. Inter-rater reliability was satisfactory. We received feedback on how to improve the instrument.

Interpretation: Quality ratings of the AGREE domains were significant predictors of outcome measures associated with guideline adoption: guideline endorsements, overall intentions to use guidelines, and overall quality of guidelines. All AGREE items were assessed as useful in determining whether a participant would use a guideline. No clusters of items were found more useful by some users than others. The measurement properties of the seven-point scale were promising. These data contributed to the refinements and release of the AGREE II.

Evidence-based guidelines are systematically developed statements aimed at assisting clinicians and patients in decisions about appropriate health care for specific clinical circumstances.[1] Guidelines assist decision-makers in solving system-level and population-level challenges.[2,3] The potential benefits of guidelines, however, are only as good as the quality of the guidelines themselves. Rigorous development and strategies for reporting are important precursors to successful implementation of the resulting recommendations.[4]

The quality of guidelines is variable, often falling short of basic standards.[5–7] To address this variability, an international team of guideline developers and researchers, the AGREE Collaboration (Appraisal of Guidelines, Research and Evaluation), created a generic instrument to assess the process of guideline development and reporting. The result of this work was the AGREE instrument, a 23-item tool targeting six quality-related domains.[8,9] It became accepted by many as the standard for guideline evaluation.[10]

As with any new assessment tool, ongoing development of the instrument is required. The AGREE Next Steps Consortium was established to conduct a program of research aimed at improving the AGREE and advancing the overall guideline enterprise. We report on the first of two studies designed to achieve these goals. This study focused on four key issues related to methodology and implementation (Figure 1).

First, the original four-point response scale was not in keeping with standards for test construction that are intended to maximize the reliability and discriminability of an instrument and minimize the number of appraisers required to evaluate a guideline.[11] To address this issue, we introduced a seven-point response scale, tested its performance and conducted a preliminary analysis of some of its measurement properties.

Second, to be of value, the AGREE instrument needs to be easy to apply and needs to generate information that is useful. To generate a reliable estimate of guideline quality, it is recommended that the 23 items of the AGREE instrument be applied by four independent reviewers.[8,9] This process can be cumbersome and resource-intensive. However, no systematic

From McMaster University (Brouwers, Kho, Hanna, Makarski); the Program in Evidence-based Care, Cancer Care Ontario (Brouwers), Hamilton, Ont.; British Columbia Cancer Agency (Browman), Victoria, BC; the Dutch Institute for Healthcare Improvement CBO and IQ Healthcare (Burgers), Radboud University Nijmegen Medical Centre, the Netherlands; St. George's University of London (Cluzeau), London, UK; the University of Bristol (Feder), Bristol, UK; Unité Cancer et Environement (Fervers), Université de Lyon – Centre Léon Bérard, Université Lyon 1, EA 4129, Lyon, France; and the Canadian Institutes of Health Research (Graham), Ottawa, Ont.

CMAJ 2010. DOI:10.1503/cmaj.091714

DOI:10.1503/cmaj.091714

analysis has been undertaken previously to determine if all items of the AGREE instrument generate information that is equally useful across different groups. This fact opens the possibility that fewer than 23 items, and varying combinations of items unique to different users, may be sufficient for evaluation purposes. Therefore, we explored whether evidence exists to inform the development of abridged versions of the AGREE that could be tailored to the unique priorities of different user groups.

Third, to be useful as well, the AGREE ratings should be associated with outcomes that are relevant to guideline usage. In keeping with previous findings,[4] guidelines of higher quality should be more attractive, endorsed or used than those of lower quality. We therefore explored whether these relationships existed and whether they were consistent across different types of users.

Finally, given that the AGREE had been in the field for some time, we systematically collected feedback from users on how the items and domains might be improved, updated and refined.

In combination with the results of the second study,[12] the data were used by the consortium to craft the AGREE II,[13] the next version of the AGREE instrument.

Methods

We report here on data collected from a larger study that contrasted the usefulness and performance of the AGREE instrument with a short generic tool, the Global Rating Scale. Results associated with components of the AGREE are presented here; those associated with the Global Rating Scale will be presented elsewhere.

Design

We used an unbalanced mixed factorial design incorporating four factors, which were user type (i.e., clinicians, guideline developers or researchers, and policy-makers), clinical topic (i.e., cancer, cardiovascular medicine and critical care), group (i.e., AGREE and Global Rating Scale versus Global Rating Scale only) and guideline (i.e., 10 reports). Guideline was not a factor of analytical interest. The group factor will not be discussed here. The study design and allocation of participants to the various factors are illustrated in Table 1. Participants who were randomized to group 1 completed guideline assessments with and questionnaires about both the AGREE and the Global Rating Scale, and group 2 did so with the Global Rating Scale only. We report here on the AGREE-specific data from group 1.

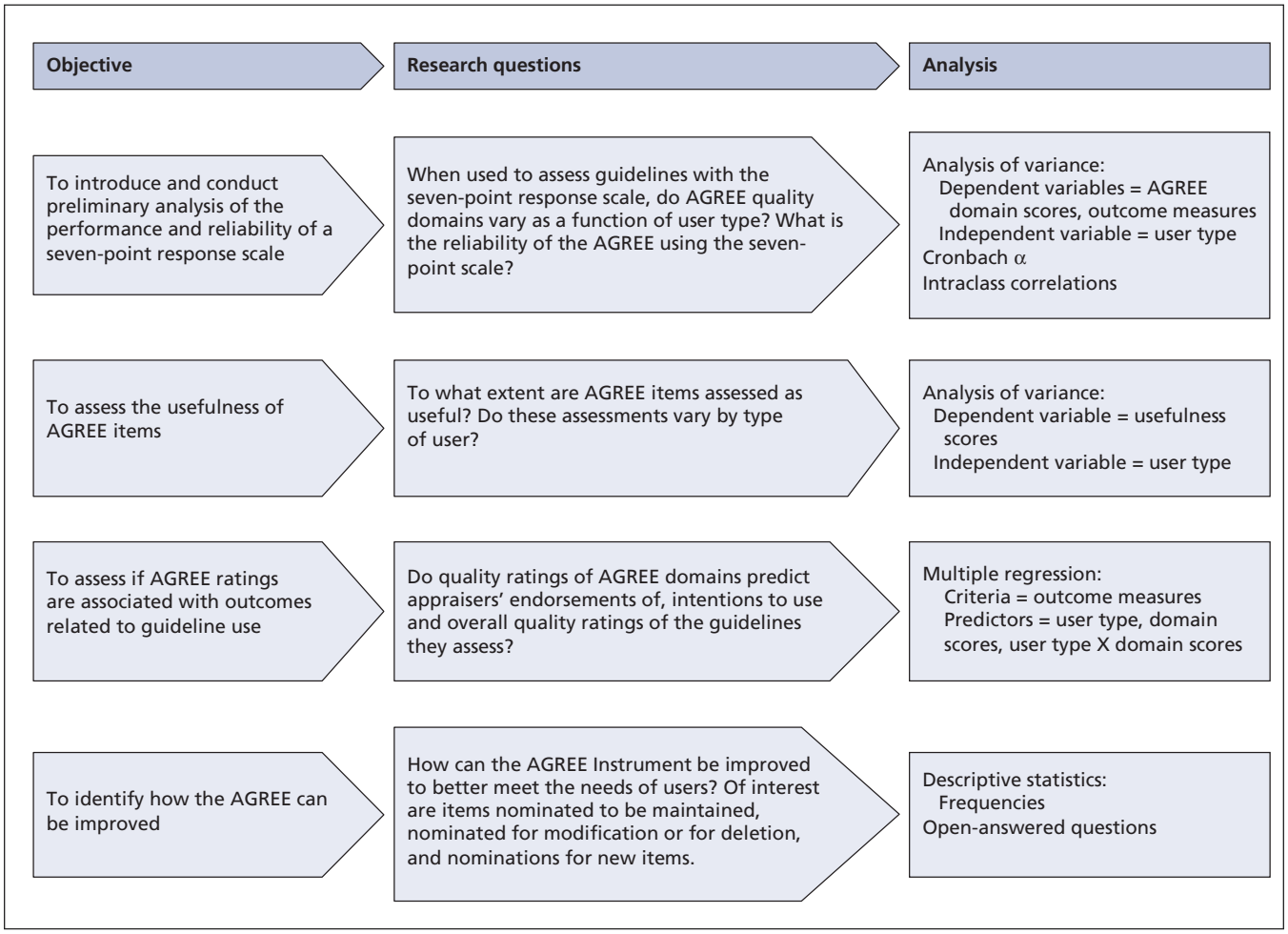

Figure 1: Program of research.

Participants

A sample of clinicians, policy-makers, and guideline developers or researchers was recruited to participate in the study (Table 1). For the whole study (i.e., including both the AGREE instrument and the Global Rating Scale), 503 clinicians, 174 policy-makers, and 164 developers or researchers were invited to participate. Three strategies that were unique to each user group were used to allocate participants to factors (Table 1).

Sample size

Our sample-size calculation was based on the interaction of user type x clinical topic x group, which was the primary analysis of interest of the whole study. After an a priori sample-size calculation for our primary outcome, which was domain score, our recruitment target for the total study was 192 participants. For the component of the study in this report, it was 96, comprising 40 clinicians (16 oncologic, 16 cardiovascular and 8 critical care clinicians), 16 policy-makers, and 40 developers or researchers.

Table 1: Study design

		User type, no.*		
Clinical topic	Guideline	Clinicians† *n = 28*	Researchers and developers‡ *n = 38*	Policy-makers§ *n = 17*
Cancer	A	15	38	17
	B			
	C			
	D			
Cardiovascular	E	7		–
	F			
	G			
	H			
Critical care	I	6		–
	J			

*Participants of each user type were randomly assigned to guidelines as indicated by clinical topic.
†The areas of expertise of clinicians were matched to clinical topic.
‡Researchers and developers were randomly assigned to clinical topics.
§Policy-makers were allocated only to the clinical topic of cancer.

Selection of guidelines

We used the US National Guidelines Clearinghouse database to search for eligible guidelines. To ensure a range of quality, a purposeful sample of 10 was chosen based on two independent assessments (M.K., J.M.) for rigour of development using the original AGREE instrument (i.e., the original seven items of this domain were assessed using the original four-point scale).

Administration

After obtaining ethics approval, we sent each participant-candidate a personalized letter of invitation in the mail and an email to ascertain interest. Two reminders were sent by email to each nonresponder. Candidates who did not complete their survey were categorized as non-responders. Consent to participate was implied with receipt of the participant's data.

Candidates who agreed to participate were assigned a unique identifier code and a confidential username and password to access the Web-based study platform. (We also accepted submissions by mail or fax.) Once logged on to the Web-based platform, participants read the guideline assigned to them and evaluated it. They then completed a series of questionnaires aimed at assessing the usefulness of the AGREE items and domains, ways in which the items could be improved and the feasibility of application.[1] An online portal for web surveys collected the data and saved it on a secure password protected data storage site.

Measures

Modified AGREE

A modified AGREE tool was used. It comprised the same 23 items within six domains (i.e., scope and purpose, stakeholder involvement, rigour of development, clarity of presentation, applicability, and editorial independence) as in the original instrument.[8,9] However, a new seven-point scale (i.e., from strongly disagree[1] to strongly agree[7]) replaced the original four-point scale.[11]

Outcome measures

Three items were designed as overall outcome measures and were also rated using a seven-point scale. These items were guideline endorsement, intention to use and overall quality.

Usefulness scale

For each AGREE item and domain, participants were asked to indicate their agreement on a seven-point scale (i.e., from strongly disagree[1] to strongly agree[7]) with the statement "rating this concept helps me determine whether or not to use a guideline." Participants also ranked domains, but these data are not presented here.

Improvement scale

Participants provided feedback on how to improve the AGREE items by considering the items clustered within each domain. Four options for feedback were offered: no changes required, modifications required (with explanation), delete item or concept, or include additional item or concept.

Analysis of data

A series of analysis of variance tests, multiple regressions, Chonbach α, and intraclass correlations were used to assess the data set (Figure 1) (Appendix 1, available at www.cmaj.ca/cgi/content/full/cmaj.091714/DC1).

Results

A total of 158 people participated in the whole study and 83 participated in the AGREE arm of the study reported here

Table 2: Demographic characteristics of participants

Characteristic	No. *n = 83*
Type of participant	
Developer or researcher (international)*	38
Clinician (Canadian)†	28
Policy- or decision-maker (Canadian)‡	17
Sex	
Male	54
Female	29
Age, yr	
25–34	3
35–44	25
45–54	35
55–64	20
≥ 65	0
Trained in methods of health research or evidence-based care	64
Academically appointed	61
Self-rated level of experience as developer of clinical practice guidelines	
Novice	28
Expert	44
Not applicable	11
Had experience as evaluator of clinical practice guidelines	65
Context of experience as user of clinical practice guidelines	
Clinical decisions or clinical practice	66
Policy-related decisions	51
Administrative decisions	51
Health system decisions	51
Had used AGREE in guideline development, no. of times	
0	42
1–5	22
6–10	6
11–15	3
16–20	0
> 20	10
Had used AGREE to evaluate guidelines, no. of times	
0	24
1–5	34
6–10	7
11–15	6
16–20	1
> 20	11

*Recruited from partner organizations, submitting authors of the Canadian Medical Association Guidelines Infobase, participants in the Canadian Partnership Against Cancer and the Conference on Guideline Standardization (COGS), members of the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) working group, and members of the Guidelines International Network.
†Recruited from publicly-available lists of websites of Canadian provincial colleges of physicians and surgeons (Alberta, British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island).
‡Recruited from members of the Canadian Agency for Drugs and Technologies in Health, Cancer Care Ontario Committee to Evaluate Drugs, heads of clinical programs of Cancer Care Ontario, Canadian Pharmacists' Association, Health Canada (Chronic and Continuing Care Division, Health Products and Food Branch, Pharmaceuticals Management Strategies, Therapeutic Effectiveness and Policy Bureau), and the Ontario Health Technology Advisory Committee.

(Table 2). Half of participants self-identified as experts in guideline development, 78% reported having experience in evaluating guidelines, and 61% have used guidelines to inform decisions. The AGREE instrument had been used by 50% of participants to inform guideline methods and by 71% for evaluation purposes.

Seven-point scale

When we tested the performance of the seven-point scale, we found significant differences in domain scores as a function of user type. With the exception of the applicability domain, where no differences were found, guideline developers and researchers gave lower domain-quality ratings, on average, than did clinicians or policy-makers (Table 3).

An exploratory analysis of the measurement properties of the scale showed that internal consistency ranged from 0.64 (for editorial independence) to 0.89 (for rigour of development). Inter-rater reliability was adequate. The number of appraisers required to reach a level of inter-rater reliability of 0.7 ranged from two to five across domains (Table 3).

Usefulness of the items

All items and domains were rated above the mid-point of the scale as useful by participants (Table 4). Among the items, "updating procedure" (item 14) received the lowest mean score (4.80) and "link between evidence and recommendations" (item 12) was rated highest (6.53). The domain of applicability was rated lowest (4.98) and that of scope and purpose was rated highest (6.32). No significant differences were found across ratings of usefulness of items or domains as a function of user type.

Predicting outcomes

With the exception of the editorial independence domain, each of the remaining five domains was a significant positive predictor for the three outcome measures. The magnitude of effect varied across domains and outcomes, and ranged from a change of 0.1 to 0.8 (on a seven-point scale) in the outcome measure for every 10% change in the domain score (Table 5). For example, every 10% change in the rigour of development domain predicted a 0.8 change in participants' endorsement score. User type and the interaction of domain and user type did not predict outcomes.

Improving items and domains

Participants provided recommendations for improvement of all items and domains. (Qualitative data is not reported.) Feedback was offered most frequently for the stakeholder involvement domain, with 39.8% of all participants suggesting modifications. Ten items were recommended by at least one user for deletion. The items nominated most frequently for deletion were "pilot testing" (item 7) (9.6% of participants) and "tools for application" (item 18) (10.6%).

Interpretation

We explored strategies to improve the measurement properties and usefulness of the AGREE instrument. We introduced

a new seven-point scale to align with standards for test construction.[11] We demonstrated that the instrument was able to detect significant differences in ratings of guideline quality. Differences of at least 10% were detected as a function of user type for five of the six domains, with the least positive assessments made by the user group that comprised guideline developers or researchers.

Although this study was not specifically designed or powered to be a reliability study, our exploratory analysis of the internal consistency of the domains aligned with ranges reported with the original AGREE instrument.[8] Given the sample size, inter-rater reliabilities were adequate. For some domains, acceptable reliability was achieved using an average of scores by two raters. However, ongoing analysis of reliability is required before definitive changes can be made to the currently recommended norm of four independent reviewers. Together, our data demonstrated a successful introduction of the seven-point response scale.

Next, we systematically explored the usefulness of the items as a function of different types of users. Our data show that all items were rated as useful in determining whether an appraiser would consider using a guideline. Although some variability was evident in absolute scores, ratings across all items and domains were above the mid-point of the response scale. In contrast to our expectations, no significant differences were evident as a function of type of user on any measure. Therefore, these data do not provide direction toward or show value in the development of abridged tools comprised of fewer concepts tailored to different users that align with group priorities.

Next, we investigated predictors of outcomes associated with uptake of guidelines. Domains of the AGREE significantly predicted participants' endorsements of guidelines, intention to use them, and overall ratings of the quality of guidelines. However, neither user type nor interaction between user type and domain were significant predictors of any of the three outcomes.

These findings are important for three reasons. First, although the instrument with the seven-point scale is sensitive to meaningful group differences in ratings of guideline quality, this sensitivity did not translate into group differences in outcomes associated with guideline adoption. Therefore, the relationships between quality and different outcomes appears to have been universal across potential users and not unique to each stakeholder group. These data, in combination with the universal endorsement of AGREE items by users, have led us to abandon our objective of developing abridged tailored versions of the AGREE.

Table 3: Performance of AGREE (as shown by scores for domain percentage and mean outcome measures by user type), and measurement properties of the seven-point scale (as shown by Cronbach α and inter-rater reliability)

	Performance of AGREE					Measurement properties of seven-point scale					
	AGREE domain percentage score, mean (SD), or outcome measure score, mean (SD), by user type				Significance (D v. C v. P)	Cronbach α	Inter-rater reliability (intraclass correlation average of *n* raters)				
Domain (D) or outcome measure (O)	Overall (83)	D (38)	C (28)	P (17)			1	2	3	4	R
D1: Scope and purpose	74 (22)	67 (23)	80 (19)	78 (21)	0.03	0.89	0.42	0.59	0.68	0.74	4
D2: Stakeholder involvement	52 (23)	45 (23)	55 (22)	62 (19)	0.03	0.73	0.40	0.57	0.67	0.73	4
D3: Rigour of development	69 (19)	62 (22)	74 (17)	75 (12)	0.01	0.75	0.30	0.46	0.56	0.63	6
D4: Clarity of presentation	68 (19)	62 (22)	72 (15)	74 (15)	0.04	0.68	0.35	0.52	0.62	0.68	5
D5: Applicability	45 (28)	44 (30)	49 (27)	41 (28)	0.62	0.80	0.57	0.73	0.80	0.84	2
D6: Editorial independence	62 (30)	54 (31)	67 (29)	75 (21)	0.03	0.64	0.47	0.64	0.73	0.78	3
O1: Endorse (I would recommend this guideline for use in practice)	5.11 (1.51)	4.89 (1.64)	5.25 (1.40)	5.35 (1.41)	0.49	NA	NA	NA	NA	NA	NA
O2: Intend to use (I would make use of a guideline of this quality in my professional decisions)	5.12 (1.57)	4.97 (1.62)	5.11 (1.73)	5.47 (1.18)	0.56	NA	NA	NA	NA	NA	NA
O3: Guideline quality (Rate the overall quality of this guideline)	5.11 (1.31)	4.82 (1.52)	5.32 (1.19)	5.41 (0.80)	0.17	NA	NA	NA	NA	NA	NA

Note: C = clinicians, D = developers or researchers, NA = not applicable, P = policy-makers, R = number of raters required to achieve inter-rater reliability of 0.7.

Table 4: Overall ratings of the usefulness of AGREE items

Domain	Items in domain	Overall rating of usefulness, score from 1–7, mean (SD)	
		Item	Domain
Scope and purpose	The overall objective(s) of the guideline is (are) specifically described	6.22 (0.96)	6.32 (0.73)
	The clinical question(s) covered by the guideline is (are) specifically described	6.25 (1.00)	
	The patients to whom the guideline is meant to apply are specifically described	6.49 (0.80)	
Stakeholder involvement	The guideline development group includes individuals from all relevant professional groups	6.05 (0.94)	5.41 (1.02)
	The patients' views and preferences have been sought	4.92 (1.56)	
	The target users of the guideline are clearly defined	5.86 (1.14)	
	The guideline has been piloted among end users	4.82 (1.74)	
Rigour of development	Systematic methods were used to search for evidence	6.48 (0.89)	6.05 (0.73)
	The criteria for selecting the evidence are clearly described	6.14 (1.06)	
	The methods for formulating the recommendations are clearly described	6.12 (1.14)	
	The health-related benefits, side effects and risks have been considered in formulating the recommendations	6.37 (0.95)	
	There is an explicit link between the recommendations and the supporting evidence	6.53 (0.69)	
	The guideline has been externally reviewed by experts prior to its publication	5.92 (1.12)	
	A procedure for updating the guideline is provided	4.80 (1.63)	
Clarity of presentation	The recommendations are specific and unambiguous	6.41 (0.70)	5.98 (0.76)
	The different options for management of the condition are clearly presented	6.00 (1.02)	
	Key recommendations are easily identifiable	6.35 (0.88)	
	The guideline is supported with tools for application	5.14 (1.58)	
Applicability	The potential organizational barriers in applying the recommendations have been discussed	4.81 (1.56)	4.98 (1.36)
	The potential cost-related implications of applying the recommendations have been considered	5.11 (1.53)	
	The guideline presents key review criteria for monitoring and/or audit purposes	5.01 (1.50)	
Editorial independence	The guideline is editorially independent from the funding body	5.77 (1.45)	5.76 (1.36)
	Conflicts of interest of members of the guideline development group have been recorded	5.75 (1.50)	

Second, intentions to use guidelines and, to a lesser extent, endorsement of guidelines, have been proposed (conceptually and empirically) as reasonable surrogates of measures of behaviour, albeit with methodologic limitations.[14–16] Eccles and colleagues found a moderately positive correlation between stated intention and actual behaviour in the health care literature.[16] Although few studies are available in the literature testing this notion, and despite some methodologic limitations, this correlation corresponds to findings in fields other than that of health care. Our study thus opens the door to future research to examine the use of the AGREE instrument as a key strategy for promoting the ultimate adoption of recommendations and for modelling the adoption by aligning the guidelines to the AGREE domains.

Finally, we received considerable feedback about how to improve and modify the instrument. Of particular importance was feedback on the underpinnings of several of the concepts, with suggested examples, wording changes, criteria or considerations. These suggestions were formally vetted and incorporated by the research team. Although some items were nominated for

Table 5: Prediction of outcome measures by AGREE domain scores

Outcome measure	AGREE domain no.*	Type III sum of squares F statistic *p* value	Parameter estimate (95% CI)
Endorse (I would recommend this guideline for use in practice)	1	< 0.001	0.01 (–0.02 to 0.05)
	2	< 0.001	0.01 (–0.02 to 0.05)
	3	< 0.001	0.08 (0.02 to 0.13)
	4	< 0.001	0.04 (0.0 to 0.08)
	5	0.014	0.01 (–0.02 to 0.04)
	6	0.155	NA
Intend to use (I would make use of a guideline of this quality in my professional decisions)	1	0.001	0.01 (–0.03 to 0.04)
	2	< 0.001	0.01 (–0.03 to 0.04)
	3	< 0.001	0.06 (0.0 to 0.12)
	4	< 0.001	0.05 (0.01 to 0.09)
	5	0.044	0.02 (–0.01 to 0.05)
	6	0.186	NA
Overall quality (Rate the overall quality of this guideline)	1	< 0.001	0.01 (–0.01 to 0.04)
	2	< 0.001	0.01 (–0.02 to 0.03)
	3	< 0.001	0.05 (0.01 to 0.08)
	4	< 0.001	0.03 (0.0 to 0.06)
	5	0.008	0.01 (–0.01 to 0.03)
	6	0.052	NA

Note: CI = confidence interval, NA = not applicable.
*Domain names, by number, are: 1 = Scope and purpose, 2 = Stakeholder involvement, 3 = Rigour of development, 4 = Clarity of presentation, 5 = Applicability, 6 = Editorial independence.

deletion, the proportion of participants who did so was small in each case. The support to keep the majority of items in the instrument and the favourable ratings the items received across stakeholder groups validate the pertinence of the original AGREE items and the underlying concepts they reflect.[8,9]

Limitations

Our study has limitations. First, although we oversampled by at least four times to achieve our target sample size, we were unable to meet our goal of 96 participants. While this shortfall did not have an impact on the power required for our primary analysis, it does attest to the challenges of conducting research related to health services.[17] Second, owing to the smaller pool from which to recruit participants, policy-makers evaluated cancer-related guidelines only (i.e., in contrast to developers or researchers and to clinicians) (Table 1). Therefore, differences found in performance as a function of type of user may have been confounded by guideline topic. Indeed, in a separate analysis (not shown), cancer-related guidelines tended to be evaluated as being of higher quality than other guideline topics. However, when policy-makers were excluded from the primary analyses, with the exception of one finding, the results continued to show that guideline developers and researchers gave statistically lower ratings of quality than did clinicians. This finding gives us confidence that we are seeing true differences between types of users.

Conclusion

The results presented here serve as an important component of the consortium's overall program of research. Our study shows a promising introduction of the new seven-point scale. It shows that items in the AGREE have universal value and can predict important outcomes associated with guideline adoption. Finally, we received considerable feedback on the original version of AGREE, which we used to improve and refine the tool. These results, in combination with the results from our second study reported in this series,[12] have led to the release of the AGREE II, the revised standard for guideline development, reporting and evaluation.[13]

This article has been peer reviewed.

Competing interests: Melissa Brouwers, Francoise Cluzeau and Jako Burgers are trustees of the AGREE Research Trust. No competing interests declared by the other authors.

Contributors: Melissa Brouwers conceived and designed the study, led the collection, analysis and interpretation of the data, and drafted the manuscript. All of the authors made substantial contributions to the study concept and the interpretation of the data, critically revised the article for important intellectual content and approved the final version of the manuscript to be published.

Acknowledgements: The AGREE Next Steps Consortium thanks the US National Guidelines Clearinghouse for its assistance in the identification of eligible practice guidelines used in the research program of the consortium. The consortium also thanks Ms. Ellen Rawski for her support on the project as research assistant from September 2007 to May 2008.

Funding: This research was supported by the Canadian Institutes of Health Research, which had no role in the design, analysis or interpretation of the data. Michelle Kho is supported by a CIHR Fellowship Award (Clinical Research Initiative).

REFERENCES

1. Committee to Advise the Public Health Service on Clinical Practice Guidelines, Institute of Medicine. In: Field MJ, Lohr KN, editors. *Clinical practice guidelines: directions for a new program*. Washington (DC): National Academy Press; 1990.
2. Whitworth JA. Best practices in use of research evidence to inform health decisions. *Health Res Policy Syst* 2006;4:11.
3. Browman GP, Brouwers M, Fervers B, et al. Population-based cancer control and the role of guidelines — towards a "systems" approach. In: Elwood JM, Sutcliffe SB, editors. *Cancer Control*. Oxford (UK): Oxford University Press; 2009.
4. Woolf SH, Grol R, Hutchinson A, et al. Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. *BMJ* 1999;318:527-30.
5. Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature [see comment]. *JAMA* 1999;281:1900-5.
6. Grilli R, Magrini N, Penna A, et al. Practice guidelines developed by specialty societies: the need for a critical appraisal. *Lancet* 2000;355:103-6.
7. Vigna-Taglianti F, Vineis P, Liberati A, et al. Quality of systematic reviews used in guidelines for oncology practice. *Ann Oncol* 2006;17:691-701.
8. The AGREE Collaboration. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. *Qual Saf Health Care* 2003;12:18-23.
9. AGREE Research Trust. Hamilton (ON): The Trust; 2004. Available: www.agreetrust.org (accessed 2009 Sept. 17).
10. Vlayen J, Aertgeerts B, Hannes K, et al. 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.
11. Streiner DL, Norman GR. *Health measurement scales: A practical guide to their development and use.* 3rd ed. Oxford (UK): Oxford University Press; 2003.
12. Brouwers MC, Kho ME, Browman GP, et al. for the AGREE Next Steps Consortium. Development of the AGREE II, part 2: assessment of validity of items and tools to support application. *CMAJ* 2010 May 31. [Epub ahead of print].
13. Brouwers M, Kho ME, Browman GP, et al. for the AGREE Next Steps Consortium. AGREE II: Advancing guideline development, reporting and evaluation in healthcare. *CMAJ.* In press.
14. Brouwers MC, Graham ID, Hanna SE, et al. Clinicians' assessments of practice guidelines in oncology: The CAPGO survey. *Int J Technol Assess Health Care* 2004;20:421-6.
15. Ajzen I, Albarracín D, Hornik R, editors. *Prediction and change of health behavior: Applying the reasoned action approach.* Mahwah (NJ): Lawrence Erlbaum Associates; 2007.
16. Eccles MP, Francis J, Foy R, et al. Improving professional practice in the disclosure of a diagnosis of dementia: A modeling experiment to evaluate a theory-based intervention. *Int J Behav Med* 2009;16:377-87.
17. Kho ME, Rawski E, Makarski J, et al. Recruitment of multiple stakeholders to health services research: lessons from the front line. *BMC Health Serv Res.* In press.

Correspondence to: *Dr. Melissa C. Brouwers, Department of Oncology, McMaster University, Henderson site, G wing, Rm. 207, 711 Concession St., Hamilton ON L8V 1C3; mbrouwer@mcmaster.ca*

Members of the AGREE Next Steps Consortium: *Dr. Melissa C. Brouwers, McMaster University and Cancer Care Ontario, Hamilton, Ont.; Dr. George P. Browman, British Columbia Cancer Agency, Vancouver Island, BC; Dr. Jako S. Burgers, Dutch Institute for Healthcare Improvement CBO, and Radboud University Nijmegen Medical Centre, IQ Healthcare, Netherlands; Dr. Francoise Cluzeau, Chair of AGREE Research Trust, St. George's University of London, London, UK; Dr. Dave Davis, Association of American Medical Colleges, Washington, USA; Prof. Gene Feder, University of Bristol, Bristol, UK; Dr. Béatrice Fervers, Unité Cancer et Environement, Université de Lyon – Centre Léon Bérard, Université Lyon 1, EA 4129, Lyon, France; Dr. Ian D. Graham, Canadian Institutes of Health Research, Ottawa, Ont.; Dr. Jeremy Grimshaw, Ottawa Hospital Research Institute, Ottawa, Ont.; Dr. Steven E. Hanna, McMaster University, Hamilton, Ont.; Ms. Michelle E. Kho, McMaster University, Hamilton, Ont.; Prof. Peter Littlejohns, National Institute for Health and Clinical Excellence, London, UK; Ms. Julie Makarski, McMaster University, Hamilton, Ont.; Dr. Louise Zitzelsberger, Canadian Partnership Against Cancer, Ottawa, Ont.*