International Journal of Technology Assessment in Health Care, **19:1** (2003), 148–157.

CHARACTERISTICS OF HIGH-QUALITY GUIDELINES

Evaluation of 86 Clinical Guidelines Developed in Ten European Countries and Canada

Jako S. Burgers
University Medical Centre Nijmegen

Françoise A. Cluzeau
St. George's Hospital Medical School

Steven E. Hanna
McMaster University

Claire Hunt
Institute of Psychiatry

Richard Grol
University Medical Centre Nijmegen and The AGREE Collaboration

Abstract

Objectives: To identify predictors of high-quality clinical practice guidelines.
Methods: A total of 86 guidelines from 11 countries were assessed by four independent appraisers per guideline using the AGREE instrument (23 items). Six aspects of guideline development were considered to explain the variation in quality scores: care level (primary/secondary care), scope (diagnosis/treatment), type of guideline (new/update), year of publication, type of agency (governmental/professional), and whether the guideline was produced within a structured and coordinated program.
Results: Guidelines produced within a guideline program and by governmental agencies had higher scores than their counterparts. Differences in the applicability of the guidelines could not be explained by the variables studied.
Conclusion: To ensure high quality, clinical guidelines should be produced within a structured and coordinated program. Professional organizations or specialist societies that aim to develop guidelines may adopt quality criteria from leading guideline agencies.

Keywords: Practice guidelines, Quality of health care, Health policy

Within the last decade the body of available clinical guidelines has expanded enormously. Guidelines are increasingly used in healthcare systems throughout the world to improve the quality of patient care (20). To ensure good quality of care, the guidelines used should meet

The research was funded by a grant from the EU BIOMED2 Programme (BMH4-98-3669).

specific criteria for quality. Quality of guidelines can be defined as "the confidence that the potential biases inherent of guideline development have been addressed adequately and that the recommendations are both internally and externally valid, and are feasible for practice" (1). However, recent studies have reported that the methodologic quality of guidelines is often modest and varies among different guidelines and different agencies (5;7;15;18;19). Whereas variation in health care is a common reason for developing guidelines, variation in the quality of guidelines will be counterproductive. To address this issue, we should learn more about the characteristics of high-quality guidelines, aiming at ensuring improvement of clinical practice and patient care. This knowledge could help policy makers and healthcare providers in selecting the best guidelines and guideline developers in setting or refining their guideline development program.

There is little research regarding the characteristics of guidelines or guideline agencies predicting guideline quality. Studies conducted in the United Kingdom (5) and Finland (18) concluded that national guidelines had higher quality scores than local guidelines. In addition, Grilli et al. (7) suggested that guidelines produced by major technology assessment agencies are probably better than those developed by specialty societies. Other predictors of guideline quality are not known yet.

In this study we sought to identify predictors of guideline quality by analyzing data collected for validation of the Appraisal of Guidelines for Research and Evaluation (AGREE) instrument (Appendix 1) (1). This instrument was developed by an international group of researchers from 13 countries (the AGREE Collaboration) with the aim to create a common, valid, and transparent approach to the appraisal of clinical guidelines (2).[1] The instrument was the result of a multistaged process of item generation, selection and scaling, field testing, and refinement procedures. As part of the validation of the instrument, a study was conducted to assess the quality of a sample of clinical guidelines developed in 10 European countries and Canada. As part of this project, information about several possible predictors was collected. We examined which of these guideline and agency characteristics were predictive of scores on the quality domains of the AGREE instrument.

METHODS

Instrument Development

To set the framework of the instrument, six theoretical quality domains were considered: scope and purpose, stakeholder involvement, rigor of development, clarity and presentation, applicability, and editorial independence. An initial list of 82 items from existing instruments and checklists and relevant literature addressed these domains (5;9;11;12;15;16). This list was examined for coverage, overlap, and content validity and reduced to 34 items. The refined list was then circulated for external review, including all AGREE partners and 15 international experts. The feedback from the reviewers led to reformulation of ambiguous items and removal of overlapping and value-laden items. The final instrument included 23 items (Appendix 1). A four-point Likert scale was used to score each item (4 = strongly agree, 3 = agree, 2 = disagree, 1 = strongly disagree).

Selection of Guidelines

We defined a *guideline* as "a set of systematically developed statements to assist practitioner and patient decisions about appropriate health care for one specific clinical condition or disease area" (10). Documents that did not contain recommendations for clinical practice (e.g., systematic reviews, service documents) were excluded. All country coordinators were asked to select 7 to 10 guidelines, published between 1992 and 1999. Coordinators were

instructed to provide guidelines that they regarded as both high and low in quality in order to test the discriminative value of the instrument. In all, 86 guidelines developed by 62 different agencies and organizations from 11 countries were selected.

Selection of Appraisers

In each country four independent appraisers per guideline were recruited. Where possible, each appraiser assessed two guidelines. The appraisers included medical practitioners, clinical experts, clinical researchers, and methodologists. Members of the guideline development group, members of the secretariat that produced the guidelines, and external referees were excluded.

Variables

To explain the variation in the quality of the guidelines, the following six characteristics of guidelines were considered:

1. Care level (primary, secondary/tertiary care, all levels);
2. Scope (prevention/diagnosis, treatment, combination);
3. Type of guideline (new, update);
4. Year of publication (1992–94, 1995–97, 1998–99);
5. Type of agency (professional/specialist societies, government-funded agencies, other);
6. Guideline program (part of guideline program, not part of guideline program).

A *guideline program* was defined as "a structured and co-ordinated program designed with the specific aim of producing several clinical practice guidelines" (Burgers J S, Grol R, Klazinga N S, et al. Towards evidence-based clinical practice: an international survey of 18 clinical guideline programs. In press.). The country coordinators were asked to include information about these variables for each guideline on a standardized form.

Analysis

We analyzed the scores according to the six quality domains of the instrument. Standardized guideline domain scores were calculated by summing the scores across the four appraisers and standardizing them as a percentage of the maximum possible score. Each guideline variable was entered into a multilevel model in order to consider the clustering effect of the agency responsible for the guideline (14). The significance of differences in standardized domain scores between guidelines with different characteristics was studied using one-way analysis of variance (ANOVA) as part of the multilevel model. We identified the proportion of variance in scores between guidelines *between* agencies and guidelines *within* agencies. Multilevel modeling also provides tests to measure the extent to which each variable could explain the variance. Analyses were performed using SPSS 9.0 and NLME 3.2 library for S-PLUS 2000 (13).

RESULTS

The standardized guideline domain scores ranged from 31.3 ("applicability") to 66.1 ("scope and purpose") (Table 1). The range of scores was broad within all six domains.

One-way ANOVA results from the multilevel models indicated that most significant differences were found for "rigor of development." Three variables accounted for these differences (level of care, scope, and guideline program). Overall, guidelines developed by government-funded agencies had the highest scores on all domains. However, the scoring differences between these agencies and professional or specialist societies were only

Table 1. Domain Scores of Guidelines Clustered According to Six Variables of Guidelines

	Scope and purpose	Stakeholder involvement	Rigor of development	Clarity and presentation	Applicability	Editorial independence
Care level						
Primary care (n = 21)	65.7	34.2	22.4[a]	57.1	29.6	48.0
Secondary/tertiary care (n = 32)	64.5	37.3	45.5	60.2	27.5	49.7
All levels (n = 33)	68.0	29.7	37.8	54.9	36.0	45.7
Scope[b]						
Prevention/diagnosis (n = 9)	73.8	32.6	38.1[a]	61.3	35.2	56.5
Treatment (n = 27)	64.2	37.9	45.2[a]	61.0	26.5	49.1
Combination (n = 47)	64.4	30.3	32.5	55.2	31.6	44.1
Type of guideline[b]						
New (n = 60)	66.2	33.9	38.7	57.3	31.9	48.9
Update (n = 25)	65.9	32.8	32.7	58.3	29.4	45.0
Year of publication[b]						
1992–94 (n = 7)	60.7	30.1	19.4	54.5	32.1	38.1
1995–97 (n = 25)	61.0	32.8	34.4	53.8	32.3	47.0
1998–99 (n = 52)	70.2	34.9	41.2	60.4[a]	31.2	50.2
Authors						
Professional/specialist societies (n = 39)	64.2	29.9	26.5	51.3	28.2	35.3
Government-funded organizations (n = 35)	71.2	39.6	48.8	64.6	36.1	59.8[a]
Other (n = 12)	57.6	28.3	36.0	55.9	27.3	53.5
Guideline program						
Part of guideline program (n = 55)	67.7	35.6	43.7[a]	63.2[a]	32.2	49.8
Not part of guideline program (n = 31)	63.5	30.2	25.3	47.5	29.8	44.3
Total (n = 86)	66.1	33.6	36.9	57.4	31.3	47.8

[a] $p < .05$.
[b] Total number is not 86 due to missing values.

significant on the domain "editorial independence." Guidelines developed within a guideline program had higher scores than their counterparts on all domains, but these were only significant for "rigor of development" and "clarity and presentation." For the domains "scope and purpose," "stakeholder involvement," and "applicability," significant differences were absent for all variables.

Multilevel modeling provides separate estimates of the variance in quality scores among guideline agencies and among guidelines within agencies. These estimates are reported in Table 2 as percentages of total variance. There is more between-agency than within-agency variation in quality scores for "stakeholder involvement," "clarity and presentation," and especially, "rigor of development." Thus, variations in these aspects of quality of guidelines are primarily associated with characteristics of guideline agencies. By contrast, variation in "applicability" scores was more associated with differences among guidelines than differences among agencies.

For "rigor of development" and "clarity and presentation," the variance of scores could be partly explained by certain characteristics of guidelines (Table 3). The level of care and scope of the guideline significantly explained variance *within* agencies, whereas the author and guideline program particularly explained variance *between* agencies. For "clarity and presentation," the guideline program and year of publication accounted for most of the variance.

DISCUSSION

The main finding of this study is that high-quality clinical guidelines were particularly produced within established guideline programs and by government-funded agencies. This is consistent with the study of Grilli et al. (7), which showed that guidelines produced by specialist societies were lower in quality than guidelines produced by major agencies such as the Scottish Intercollegiate Guidelines Network (SIGN) and the *Agence Nationale d'Accréditation et d'Évaluation en Santé* (ANAES) in France. These agencies have a structured guideline program, providing a systematic procedure with key elements such as a multidisciplinary guideline development group, a systematic literature review, external peer review, and different products for dissemination (Burgers et al. In press). These elements ensure high scores on several domains, in particular on "rigor of development." On the other hand, our study also showed that the agency responsible for guideline development had less influence on "applicability" than on other domains (Table 2). This suggests that agency policies and procedures are more concerned with the methodology of producing guidelines than with the effectiveness of guidelines in daily practice.

Developing high-quality guidelines requires a sufficiently skilled team of people and sufficient budget. In general, governmental agencies have greater resources than professional organizations and specialist societies, which might explain why their guidelines have higher quality scores. Nevertheless, we still believe that professional organizations can develop high-quality guidelines, provided they develop their guidelines within a structured program and adopt quality criteria of other programs.

The influence of other characteristics on the quality scores was limited. Guidelines with a narrow scope (i.e., exclusively focusing on prevention/diagnosis or treatment) had higher scores on "rigor of development" than guidelines that covered both prevention/diagnosis and treatment. The quality of a guideline might be improved by providing recommendations on a few well-defined issues instead of covering the whole clinical area of the condition selected for guideline development. As a consequence, guidelines produced for primary care had lower scores on "rigor of development," because these were broader in scope than guidelines in secondary care that focus on an already established diagnosis.

Table 2. Standard Deviation (95% Confidence Limits) and Proportion of Variance of Domain Scores Occurring Between Agencies (Agency Level) and Within Agencies (Guideline Level)

	Scope and purpose	Stakeholder involvement	Rigor of development	Clarity and presentation	Applicability	Editorial independence
Standard deviation						
Agency (n = 62)	14.5 (9.3–22.4)	14.5 (10.9–19.3)	22.7 (18.3–28.2)	15.5 (11.9–20.3)	9.7 (5.5–17.2)	19.0 (12.7–28.6)
Guideline (n = 86)	14.7 (10.9–19.8)	10.1 (7.5–13.5)	10.9 (8.3–14.3)	10.5 (7.9–13.8)	15.5 (12.4–19.4)	21.1 (16.3–27.3)
% Variance						
Agency (n = 62)	49.1	67.4	81.3	68.6	28.1	44.8
Guideline (n = 86)	51.0	32.6	18.7	31.4	71.9	55.2

Table 3. Relative Reduction of Variance by Different Predictors for the Domains "Rigor and Development" and "Clarity and Presentation"

	Rigor and Development		Clarity and Presentation	
	Between agency	Within agency	Between agency	Within agency
Care level	NS	11.1	NS	NS
Scope	NS	10.2	NS	NS
Type of guideline	NS	NS	NS	NS
Year of publication	NS	NS	−16.4	29.5
Author	7.4	NS	NS	NS
Guideline program	7.6	NS	19.5	NS

NS = $p > .05$.

Surprisingly, the year of publication and the type of guideline (new versus updated) had little influence on the scores. However, there was a small trend of overall improvement over time.

Estimates of the variance *between* agencies are difficult due to the low number of guidelines per agency on average. This could explain the odd increase (i.e., the reduction of variance is negative) in the estimate when year is added to the domain "clarity and presentation" analysis (Table 3). In contrast, *within* agencies the clarity and presentation of their guidelines obviously improves over time.

The strength of our study is that we assessed the guidelines with a rigorously developed instrument created by a collaboration of international experts in guideline development. There is insufficient evidence for adopting any other existing guideline appraisal instrument (6). In contrast to other studies (7;15), our sample of guidelines was not restricted to guidelines included in MEDLINE, thus representing a broad range of guidelines that are not necessarily representative of the quality of guidelines produced by the agencies selected. Moreover, we did not aim to provide a general statement about "the quality of clinical guidelines." We aimed to explain the variance in quality by characteristics of the guidelines. Therefore, we collected additional information about the background and context of the guidelines (e.g., guideline program) that enabled us to explain differences in quality scores. So far, this is the first study to achieve this. However, it is uncertain whether the selection process is related to other variables that have not been studied.

Our study was limited by the lack of information on the ultimate adherence to the guidelines. Evidence-based guidelines do not guarantee that they will be followed (8). Other factors, such as attitudinal and organizational barriers, should be overcome to ensure any effect of the guideline in daily practice (3;4). It would be interesting for future research to study the relationship between the "quality" of guidelines and the effectiveness of guidelines.

POLICY IMPLICATIONS

Clinical guidelines should be produced within a structured and coordinated program to ensure that they are of high quality. Professional organizations or specialist societies that aim to develop guidelines may adopt quality criteria from leading guideline agencies. International collaboration is needed to set standards for guideline quality. As an example, the AGREE instrument for assessing the quality of clinical guidelines (1) is a recent product of international collaboration that can be used by policy makers to help them decide which guidelines could be recommended for use in practice and by guideline developers to follow a structured and rigorous development methodology. A collaborative network of guideline organizations will contribute to further improvement of guideline methodology and implementation and to avoiding duplication of efforts. Guideline clearinghouses

(e.g., the U.S. National Clearinghouse [17]) can contribute to this process by disseminating high-quality guidelines internationally that can be used by different organizations for local adaptation. The overall cost of developing guidelines could be reduced considerably if guideline developers used high-quality guidelines as a basis for producing their own guidelines.

NOTE

[1]The following individuals, participating in the AGREE Collaboration, contributed to this paper: José Asua, Basque Office for Health Technology Assessment, Spain; Anne Bataillard, Fédération Nationale des Centres de Lutte Contre le Cancer, Lyon, France; Melissa Brouwers, McMaster University, Hamilton, Ontario, Canada; George Browman, Hamilton Regional Cancer Centre, Hamilton, Canada; Bernard Burnand, Institut Universitaire de Médecine Sociale et Préventive, Lausanne, Switzerland; Isabelle Durand-Zaleski, Hôpital Henri Mondor, Cedez, France; Pierre Durieux, Hôpital Européen Georges Pompidou, Paris, France; Cindy Farquhar, New Zealand Guidelines Group, Auckland, New Zealand; Gene Feder, Barts and The London, Queen Mary's School of Medicine and Dentistry, University of London, UK; Béatrice Fervers, Fédération Nationale des Centres de Lutte Contre le Cancer, Lyon, France; Roberto Grilli, Agenzia Sanitaria Regionale, Bologna, Italy; Jeremy Grimshaw, Ottawa Health Services Research Institute, Ottawa, Canada; Pieter ten Have, Dutch Institute for Healthcare Improvement CBO, Utrecht, The Netherlands; Rod Jackson, Effective Practice Institute, University of Auckland, New Zealand; Albert Jovell, Fundacio Biblioteca Josep Laporte, Barcelona, Spain; Niek Klazinga, Academic Medical Centre, University of Amsterdam, The Netherlands; Finn Kristensen, Danish Institute for Health Technology Assessment, Copenhagen, Denmark; Peter Littlejohns, National Institute for Clinical Excellence, London, UK; Pia Bruun Madsen, Danish Institute for Health Technology Assessment, Copenhagen, Denmark; Juliet Miller, Scottish Intercollegiate Guidelines Network (SIGN), Edinburgh, UK; Günter Ollenschläger, Agency for Quality in Medicine, Cologne, Germany; Camilla Palmhøj-Nielsen, Danish Institute for Health Technology Assessment, Copenhagen, Denmark; Loes Pijnenborg, Dutch College of General Practitioners, Utrecht, The Netherlands; Safia Qureshi, Scottish Intercollegiate Guidelines Network (SIGN), Edinburgh, UK; Rosa Rico-Iturrioz, Basque Office for Health Technology Assessment, Spain; Kitty Rosenbrand, Dutch Institute for Healthcare Improvement CBO, Utrecht, The Netherlands; Jean Slutsky, Agency for Healthcare Research and Quality, Rockville, Maryland, USA; John-Paul Vader, Institut Universitaire de Médecine Sociale et Préventive, Lausanne, Switzerland; Joost Zaat, Centre for Quality of Care Research, University Medical Centre Nijmegen, The Netherlands.

REFERENCES

1. The AGREE Collaboration. *Appraisal of Guidelines for Research and Evaluation (AGREE) instrument*. Available at: www.agreecollaboration.org.
2. The Appraisal of Guidelines, Research, and Evaluation in Europe (AGREE) Collaborative Group. Guideline development in Europe: An international comparison. *Int J Technol Assess Health Care*. 2000;16:1029-1049.
3. Browman GP. Improving clinical practice guidelines for the 21st century. *Int J Technol Assess Health Care*. 2000;16:959-968.
4. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? *JAMA*. 1999;282:1458-1465.
5. Cluzeau F, Littlejohns P, Grimshaw J, Feder G, Moran S. Development and application of a generic methodology to assess the quality of clinical guidelines. *Int J Qual Health Care*. 1999;11:21-28.
6. Graham ID, Calder LA, Hébert PC, Carter AO, Tetroe JM. A comparison of clinical practice guideline appraisal instruments. *Int J Technol Assess Health Care*. 2000;16:1024-1038.
7. Grilli R, Magrini N, Penna A, Mura G, Liberati. Practice guidelines developed by specialty societies: The need for a critical appraisal. *Lancet*. 2000;355:103-105.
8. Grol R. Beliefs and evidence in changing clinical practice. *BMJ*. 1997:315:418-421.

9. Grol R, Dalhuijzen J, Thomas S, et al. Attributes of clinical guidelines that influence use of guidelines in general practice: Observational study. *BMJ*. 1998;317:858-861.
10. Institute of Medicine. Field MJ, Lohr KN (eds.). *Clinical practice guidelines: Directions for a new program*. Washington, DC: National Academy Press; 1990.
11. Lohr KN. The quality of practice guidelines and the quality of health care. In: Selbmann HK. *Guidelines in health care: Report of a WHO Conference*. Baden-Baden: Nomos Verlagsgesellschaft; 1998:42-52.
12. Lohr KN, Field MJ. A provisional instrument for assessing clinical practice guidelines. In: Institute of Medicine. Field MJ, Lohr KN (eds.). *Guidelines for clinical practice*. Washington, DC: National Academy Press; 1992.
13. Pinheiro JC, Bates DM. *Mixed-effects models in S and S-PLUS*. Heidelberg: Springer Verlag. 2001.
14. Rice N, Leyland A. Multilevel models: Applications to health data. *J Health Serv Res Policy*. 1996;1:154-164.
15. Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. *JAMA*. 1999;281:1900-1905.
16. Thomson R, Lavender M, Madhok R. How to ensure that guidelines are effective. *BMJ*. 1995;311:237-242.
17. U.S. National Clearing house website. Available at: http://www.guideline.gov/index.asp.
18. Varonen H, Mäkelä M. Practice guidelines in Finland: Availability and quality. *Qual Health Care*. 1996;6:75-79.
19. Ward JE, Grieco V. Why we need guidelines for guidelines: A study of the quality of clinical practice guidelines in Australia. *Med J Aust*. 1996;165:574-576.
20. Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J. Potential benefits, limitations, and harms of clinical guidelines. *BMJ*. 1999;318:527-530.

APPENDIX 1

The AGREE Instrument, September 2001

Scope and Purpose

1. The overall objective(s) of the guideline is(are) specifically described.
2. The clinical question(s) covered by the guideline is(are) specifically described.
3. The patients to whom the guideline is meant to apply are specifically described.

Stakeholder Involvement

4. The guideline development group includes individuals from all the relevant professional groups.
5. The patients' views and preferences have been sought.
6. The target users of the guideline are clearly defined.
7. The guideline has been piloted among target users.

Rigor of Development

8. Systematic methods were used to search for evidence.
9. The criteria for selecting the evidence are clearly described.
10. The methods used for formulating the recommendations are clearly described.
11. The health benefits, side effects, and risks have been considered in formulating the recommendations.
12. There is an explicit link between the recommendations and the supporting evidence.
13. The guideline has been externally reviewed by experts prior to its publication.
14. A procedure for updating the guideline is provided.

Clarity and Presentation

15. The recommendations are specific and unambiguous.
16. The different options for management of the condition are clearly presented.
17. Key recommendations are easily identifiable.
18. The guideline is supported with tools for application.

Applicability

19. The potential organizational barriers in applying the recommendations have been discussed.
20. The potential cost implications of applying the recommendations have been considered.
21. The guideline presents key review criteria for monitoring and/or audit purposes.

Editorial Independence

22. The guideline is editorially independent from the funding body.
23. Conflicts of interest of guideline development members have been recorded.

https://doi.org/10.1017/S026646230300014X Published online by Cambridge University Press