

Practice guidelines developed by specialty societies: the need for a critical appraisal

Roberto Grilli, Nicola Magrini, Angelo Penna, Giorgio Mura, Alessandro Liberati

Summary

Background There is increasing concern about the quality, reliability, and independence of practice guidelines. Because no information is available on the methodological quality of the guidelines developed by specialty societies, we undertook a survey on those published in peer-reviewed journals.

Methods Practice guidelines produced by specialty societies and published in English between January, 1988, and July, 1998, were identified through MEDLINE. Their quality was assessed in terms of whether they reported: the type of professionals and stakeholders involved in the development process; the strategy to identify primary evidence; and an explicit grading of recommendations according to the quality of supporting evidence.

Findings Overall, 431 guidelines were eligible for the study. Most did not meet the criteria: 67% did not report any description of the type of stakeholders, 88% gave no information on searches for published studies, and 82% did not give any explicit grading of the strength of recommendations. There was improvement over time for searches (from 2% to 18%, $p < 0.001$) and explicit grading of evidence (from 6% to 27%, $p < 0.001$). All three criteria for quality were met in only 22 (5%) guidelines.

Interpretation Despite improvement over time, the quality of practice guidelines developed by specialty societies is unsatisfactory. Explicit methodological criteria for the production of guidelines shared among public agencies, scientific societies, and patients' associations need to be set up. Common standards of reporting, following the same principles that led to the CONSORT statement for randomised clinical trials, should be promoted.

Lancet 2000; **355**: 103–06
See *Commentary page* ???

Istituto di Ricerche Farmacologiche "Mario Negri", Milan (R Grilli MD, A Penna MD, G Mura BA, A Liberati MD); **Universita degli Studi di Modena e Reggio Emilia** (A Liberati); **Centro per la Valutazione della Efficacia della Assistenza Sanitaria (CeVEAS), Modena, Italy** (N Magrini MD)

Correspondence to: Dr Roberto Grilli, Agenzia Servizi Sanitari Regionali, Piazzale Marconi 25, 00144 Rome, Italy (e-mail: webmaster@assr.it)

Introduction

Over the past 20 years practice guidelines have become an increasingly popular tool for synthesis of clinical information so as to change clinical practice and improve quality of health care. Medical specialty societies have been particularly active in producing such guidelines together with agencies whose remit includes technology assessment and health care evaluation.

Such a quantitative growth in the number of guidelines available in different specialties is, however, a source of concern since there is evidence that recommendations produced by different groups can be conflicting^{1,2} and some researchers go so far as to say they are invalid, unreliable, and irrelevant.³

Thus, growth in the numbers of guidelines without application of rigorous criteria for their production^{4,5} could undermine their credibility and lead to harm to the patient if the wrong recommendations were put into practice.^{1,2}

To see whether these concerns about the quality of existing guidelines have any foundation, we undertook a survey of practice guidelines officially issued by specialty societies over the past 10 years.

Methods

Practice guidelines developed by specialty societies were identified through MEDLINE, from January, 1988, to July, 1998. The search strategy is outlined in panel 1. The list of references identified through this process was examined by three independent assessors from our team, and papers and documents including in the title words such as "guidelines", "parameters", "standards", "consensus", explicitly written by a specialty society (or on behalf of a specialty society), were eligible as long as they reported recommendations on the courses of action to be undertaken in specific clinical circumstances. Editorials and reviews (narrative or systematic) articles were excluded. Disagreements on the eligibility of individual papers were rare and were resolved by discussion.

Each eligible paper was seen by at least two independent assessors who applied a standard checklist with three items (panel 2) to explore: whether the published document reported the type of professionals involved in the guideline development; the strategy used in searching for the primary evidence; and an explicit grading of recommendations according to the quality of supporting evidence.

We chose these three items because they can be expected to be, in various ways, related to the scientific quality of guidelines,

Panel 1: MEDLINE search strategy for identification of practice guidelines developed by specialty societies

Query	Search terms
1	Recommendation(s) or consensus, in title or abstract
2	Society(s) or college(s) or association(s), in title or abstract
3	Query 1 and 2
4	Practice guideline or consensus development conference (medical subject headings or publication type)
5	Query 3 or 4

Panel 2: Checklist to assess the quality of guidelines endorsed by specialty societies

Description of the type of professionals involved in developing the guideline

Yes—if there was a description of the type of professionals and other stakeholders involved in the development process

Partially—if only a list of names with institutional affiliation was provided

No—if only names were reported, without further information

Description of the sources of information used to retrieve the relevant evidence

Yes—if it was explicitly stated that searches were undertaken, at least through MEDLINE

No—if no information was reported.

Explicit grading of the evidence in support of the main recommendations

Yes—if any form of explicit grading of the quality of the supporting evidence was reported

No—if otherwise

for the following reasons: multispecialty panels have been shown to provide a more conservative view when interpreting the advantages and limitations of a given technology,⁶ the validity of systematic reviews depends heavily on thoroughness of the search, and consultation of at least MEDLINE and EMBASE databases is regarded as mandatory to achieve acceptable comprehensiveness in the identification of primary studies;^{7,8} grading of the evidence is essential to distinguish between evidence-based and consensus-based recommendations.⁹ The reliability of the three items was tested by three of us on a random sample of 20 guidelines with the κ statistic, the values of which ranged from 0.61 to 1.0.

Results

Overall, of 3129 abstracts and titles retrieved from the MEDLINE search, 576 (18%) papers were identified as potentially eligible. 145 (25%) were excluded after closer scrutiny of the published reports because: 90 were narrative or systematic reviews or position papers; 23 were duplicate publications of the same guideline or short summaries of full reports published elsewhere; and 32 were documents about professional curricula or organisational standards.

A total of 431 guidelines were eligible for our study (table 1). Most (289/431, 67%) of the guidelines did not describe the type of professionals involved in their

	n
Specialty area	
Cardiology	120 (28%)
Oncology	65 (15%)
Neurology	43 (10%)
Gynaecology	24 (6%)
Internal medicine	24 (6%)
Anaesthesia	19 (4%)
Other*	136 (31%)
Aspect of care	
Prevention	56 (13%)
Diagnosis	99 (23%)
Treatment	165 (38%)
Overall management	111 (26%)
Time of publication	
1988–91	48 (11%)
1992–93	81 (19%)
1994–95	125 (29%)
1996–98	177 (41%)

*Includes some specialty areas, individually representing less than 3% of the overall sample.

Table 1: General characteristics of practice guidelines issued by specialty societies and identified on MEDLINE between January, 1988, and July, 1998

	1988–91 (n=48)	1992–93 (n=81)	1994–95 (n=125)	1996–98 (n=177)	p for trend
Full description of professionals	6 (12%)	9 (11%)	11 (9%)	27 (15%)	0.99
Search undertaken	1 (2%)	4 (5%)	14 (11%)	32 (18%)	<0.001
Grading of recommendation	3 (6%)	5 (6%)	21 (17%)	48 (27%)	<0.001

Table 2: Number of guidelines that met the three quality criteria according to year of publication

development; this information was explicitly reported in just 12%, and the remaining 21% reported only the names and the institutional affiliation of those involved. When completeness of reporting on this item was assessed according to the year of publication there was no evidence of improvement over time (table 2).

In 118 (28%) guidelines there was evidence of inclusion of at least one professional or representative whose specialty differed from the prevailing specialty. When there was information relating to the involvement of non-clinicians (144), epidemiologists or methodologists were the professionals most commonly encountered (37/144, 26%), then primary-care physicians (20/144, 14%), health-care administrators (13/144, 9%), and patients' or consumers' representatives (12/144, 8%). Most guidelines (377 of 431, 87%) did not report any information on whether a systematic search for published studies was done. Among those reporting attempts at searching (54), MEDLINE was the only database searched in 28 (52%), and in 26 (48%) searches through EMBASE or other electronic sources were combined with MEDLINE. The proportion of guidelines reporting some form of search increased over time, from 2% to 18% ($p < 0.001$; table 2).

Just 77 (18%) guidelines used explicit criteria to grade the strength of the scientific evidence in support of their recommendations. As shown in table 2, the number of guidelines satisfying this criterion increased from 6% in 1988–91 to 27% in 1996–98 ($p < 0.001$). Altogether, the three quality criteria were met only in 22 (5%) of the identified guidelines, and 231 (54%) did not meet any criterion. 149 (34%) met one criterion, 29 (7%) met two criteria.

Discussion

Our survey shows that the quality of reporting of practice guidelines produced by specialty societies fell short of acceptable methodology up to mid 1998. If practice guidelines are to be widely accepted as an improvement tool for quality, greater attention needs to be paid to the methods used to develop them.^{9–12}

In the USA the Institute of Medicine's reference definition of practice guidelines appropriately underscores that they are "systematically developed statements",¹¹ thus highlighting that the recommendations should be the outcome of the methodological process. Panel composition, thoroughness of the search for published papers, and explicit definition of evidence are essential components of this process.^{9–12} Therefore, although efforts to develop a fully validated assessment tool for practice guidelines are still underway¹³ we did our survey limiting ourselves to these three characteristics. The issue of quality of guidelines is attracting much attention and Shaneyfelt and colleagues¹⁴ have, since the completion of our work, published their own findings based on a "composite quality score" that assigns equal weight to different items. We disagree with such an approach because it ignores the fact that the different items may

have very different relevance to validity and applicability of recommendations.

Lacking a validated and internationally accepted tool for assessing the quality of guidelines, we restricted our assessment to the three specific items mentioned above because they are easy to assess in published reports and have sufficient face value. By following this different approach we came to different conclusions and implications. Shaneyfelt and colleagues did not find differences in quality between guidelines developed by specialty societies and other organisations.¹⁴ We believe this is a misleading conclusion, because in our experience guidelines produced by major technology assessment agencies (such as Agence Nationale d'Accreditation et d'Evaluation en Santé in France, Agency for Health Care Policy and Research in the USA, and Scottish Intercollegiate Guidelines Network Initiative in Scotland, among others) all fulfil our three basic quality criteria. We suspect that the results of Shaneyfelt and colleagues were also due to the way in which the statistical analysis was done (ANOVA to compare the mean number of items satisfied by each category of developers). We thus disagree with the conclusion of their paper, that there is a general "quality problem" with published guidelines.¹⁴ Such a view does, we believe, obscure the main problem highlighted in our paper, namely that there is a major quality problem (and possibly a validity one, as well) in the guidelines developed by specialty societies

Our study too has its limitations. One may argue that we cannot discriminate between poor quality of reporting and limitation in the production of guidelines. In other words, our results could depend on the report itself and they might have been different if we had access to what was actually done. However, the evidence on this point is sparse, and the only study that we know for sure systematically explored the difference between reporting and actual conduct in the context of clinical trials did not show any difference.¹⁵ Either way, from a practical point of view our paper shows that guidelines developed by specialty societies fall short of the desirable informativeness since practice guidelines are meant to inform and guide choices in health care and, as such, should be explicit and transparent. Only in this way will readers be in a position to assess whether recommendations are valid (ie, based on evidence)¹⁶ reproducible,^{1,2} and free of conflicts of interest,¹⁷ an issue that the discussion after the publication of WHO guidelines on hypertension has clearly highlighted.¹⁸

Completeness of information aside, the approach adopted by specialty societies to develop their clinical policies seems to be questionable in most cases—because only about a quarter of the guidelines reviewed in this paper were developed by multispecialty panels, an approach that has been repeatedly suggested as a way to avoid a biased view in formulation of recommendations.⁶ Another reason for concern is the paucity of guidelines developed involving patients and consumer representatives, which raises the concern that the value of their inputs is not properly recognised.

Besides efforts to improve the quality of guidelines, actions should also be taken at the level of publication and peer review of guidelines published in medical journals. The example of what has been done on standard reporting of publication of randomised clinical trials (see the CONSORT statement¹⁹), should be followed, whereby authors of papers are requested to include the essential details about the methods used.

As in all review articles, appropriate identification of relevant primary studies can affect the validity of the conclusions. We do not know whether our search for published papers captured all guidelines published by specialty societies because no independent electronic or hand searches have been done. We used only MEDLINE as the reference database, and the search strategy could be improved in terms of specificity and sensitivity. In any case, we believe that guidelines published in journals not indexed on MEDLINE are unlikely to be systematically better. Moreover, the proportion of practice guidelines that achieve publication in peer-reviewed journals is unknown.

Since no differences in quality of reporting related to language of publication have been demonstrated,²⁰ we think that our focus only on guidelines published in English is unlikely to have caused any substantial sampling bias.

Contributors

Roberto Grilli, Nicola Magrini, Angelo Penna, and Alessandro Liberati were responsible for study design, for the development of the survey, and for quality assessment of individual guidelines. The reliability assessment was done by Roberto Grilli, Angelo Penna, and Alessandro Liberati. Selection of eligible guidelines was done by Angelo Penna, Roberto Grilli, and Nicola Magrini. Giorgio Mura was responsible for data input, quality control, and analysis. Roberto Grilli and Alessandro Liberati wrote the first version of the paper on which Angelo Penna and Nicola Magrini provided comments and suggestions.

Acknowledgments

This study was funded by the network of local health authorities and hospital trusts participating in the TRIPSS project, aimed at promoting the use of research information in the organisation and delivery of health care.

References

- Thomson R, McElroy H, Sudlow M. Guidelines on anticoagulant treatment in atrial fibrillation in Great Britain: variation in content and implications for treatment. *BMJ* 1998; **316**: 509–13.
- Unwin N, Thomson R, O'Byrne AM, Laker M, Armstrong H. Implications of applying widely accepted cholesterol screening and management guidelines to British adult population: cross-sectional study of cardiovascular disease and risk factors. *BMJ* 1998; **317**: 1125–30.
- Varonen H, Makela M. Practice guidelines in Finland: availability and quality. *Qual Health Care* 1997; **6**: 75–79.
- Hibble A, Kanka D, Pencheon D, Pooles F. Guidelines in general practice: the new Tower of Babel? *BMJ* 1998; **317**: 862–63.
- Feder G. Guidelines for clinical guidelines. *BMJ* 1998; **317**: 427–28.
- Murphy MK, Black NA, Lamping DL, et al. Consensus development methods and their use in clinical guideline development. *Health Technol Assessment* 1998; **2**.
- Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ* 1994; **309**: 1286–91.
- Smith BJ, Darzings PJ, Queen M, Heller RF. Modern methods of searching the medical literature. *Med J Aust* 1992; **157**: 603–11.
- Grimshaw J, Russell I. Achieving health gain through clinical guidelines: I—developing scientifically valid guidelines. *Qual Health Care* 1993; **2**: 243–48.
- 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.
- Institute of Medicine. Guidelines for clinical practice: from development to use. Washington DC: National Academy Press, 1992.
- Eddy DM. Practice policies: guidelines for methods. *JAMA* 1990; **263**: 1839–41.
- Littlejohns P, Cluzeau F. Promoting the rigorous development of clinical guidelines in Europe through the creation of a common appraisal instrument. Amsterdam: Scientific Basis for Health Services, 1997.
- 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**: 1900–05.
- Liberati A, Himel H, Chalmers TC. A quality assessment of

- randomized controlled trials of primary treatment of breast cancer. *J Clin Oncol* 1986; **4**: 942–51.
- 16 Cluzeau F, Littlejohns P, Grimshaw JM. Appraising clinical guidelines: towards a “which” guide for purchasers. *Qual Health Care* 1994; **3**: 121–22.
- 17 Stelfox HT, Chua G, O'Rourke K, Detsky AS. Conflict of interest in the debate over calcium-channel antagonists. *N Engl J Med* 1998; **338**: 101–06.
- 18 Woodman R. Open letter disputes WHO hypertension guidelines. *BMJ* 1999; **318**: 893.
- 19 Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996; **276**: 637–39.
- 20 Moher D, Fortin P, Jadad AR, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *Lancet* 1996; **347**:

Classification of thyroid size by palpation and ultrasonography in field surveys

S Peterson, A Sanga, H Eklöf, B Bunga, A Taube, M Gebre-Medhin, H Rosling

Summary

Background Goitre surveys are used to assess the degree of iodine deficiency in a population. The change of goitre classification made by WHO in 1994 implied that a smaller thyroid size should be regarded as goitre. Furthermore, the acceptable goitre prevalence was lowered from 10% to 5%, and ultrasonography was recommended as a more precise method for diagnosis of goitre. We studied the effects of the change of palpation system, and compared the precision of the old and new systems with that of ultrasonographic examination.

Methods We studied 225 schoolchildren (aged 7–14 years) in a highland village in Tanzania. The size of the thyroid was assessed in duplicate by ultrasonography and by WHO's 1960 and 1994 palpation systems. The latter were done by three examiners. Variations within and between examination methods and examiners were assessed, and measurement errors by ultrasonography were assessed from duplicate examinations. The sensitivity and specificity of the two palpation systems were calculated, with diagnosis by ultrasonography as the gold standard. Apparent palpation prevalences were calculated at a “true” 5% prevalence.

Findings The lowered criterion for goitre resulted in an extra 20–33% of children being diagnosed as having goitre by palpation. The variation between repeat examinations was only slightly smaller by ultrasonography ($\kappa=0.63$) than by experienced examiners ($\kappa=0.57$ – 0.58). The variation between thyroid volume estimation by ultrasonography and the true volume was about 50% due to both measurement error and variation in the shape of thyroid lobes. The new goitre criterion decreased specificity from 76% to 29%, whereas sensitivity rose from 56% to 80%. In contrast, a suggested sharpening of the old criterion increased specificity to 90%.

Departments of Women's and Children's Health (S Peterson MD, Prof M Gebre-Medhin MD), **Radiology** (H Eklöf MD), and **Statistics** (A Taube PhD), **Uppsala University, Sweden; Tanzania Food and Nutrition Centre, Dar es Salaam, Tanzania** (A Sanga MD, B Bunga Dipl Comm Hlth); and **Department of Public Health Sciences, Karolinska Institute, Stockholm, Sweden** (Prof H Rosling MD)

Correspondence to: Prof Hans Rosling, Division of International Health, Department of Public Health Sciences, Karolinska Institutet, SE-171 76 Stockholm, Sweden (e-mail: hans.rosling@phs.ki.se)

Interpretation A return to the old (1960) palpation criterion for goitre: “lobes larger than the terminal phalanges of thumbs” and to an accepted palpation goitre prevalence of 10% can allow affordable monitoring of thyroid size through palpation in field surveys.

Lancet 2000; **355**: 106–10

See Commentary page xxx

Introduction

The size of the thyroid gland reflects the severity of iodine deficiency in school-aged children. The goitre rate in school-aged children has therefore been a frequently used indicator of the degree of iodine deficiency in a population.^{1,2} The outcome of such a judgment depends on the cut-off size at which the thyroid is classified as a goitre, the precision with which the thyroid size is estimated around that cut-off, and the prevalence of goitre in school-aged children at which iodine-deficiency disorders are regarded as a public-health problem.

In 1960, WHO defined a goitre as a thyroid gland the lateral lobes of which have a volume greater than the terminal phalanges of the thumbs of the person examined.^{3–5} Iodine deficiency was defined as a public-health problem if more than 10% of schoolchildren in a population had goitre.^{3,5–7} In 1994, WHO decreased the number of grades in the classification of goitres from two to four (table 1), and effectively defined an enlarged, palpable thyroid as a goitre. This criterion meant that smaller thyroids than before were regarded as goitres. The goitre rate at which iodine-deficiency disorders are regarded as a public-health problem was simultaneously lowered from 10% to 5%.² No comparative study has been published on the effect of these changes, nor on the precision obtained in goitre surveys by WHO's 1960 and 1994 criteria for goitre diagnosis. Furthermore, ultrasonographic estimation of thyroid size has been advocated as being more precise than palpation.^{2,8} However, such examinations are cumbersome and costly to carry out in remote parts of low-income countries where goitre surveys are most needed. Therefore we have studied the effect of the changed goitre criteria on the precision of thyroid-size estimation by palpation and ultrasonography in a rural Tanzanian population.