Evaluation of the Practice Guidelines of Finnish Institute of Occupational Health with AGREE Instrument

Maritta KINNUNEN-AMOROSO1*, Iris PASTERNACK1, Sirpa MATTILA2 and Anu PARANTAINEN3

1Finnish Institute of Occupational Health, Topeliuksenkatu 41 a A, 00250 Helsinki, Finland
2Occupational Health Service of Municipality of Espoo, Kirkkojärventie 6 B, Espoo, Finland
3Mehiläinen Health Care Center, Siltasaarenkatu 18 A 00530 Helsinki, Finland

Received May 27, 2008 and accepted April 2, 2009

Abstract: Guidelines for occupational physicians are increasing in number. Their quality and content is varied and they may even provide conflicting recommendations. Earlier studies show that guidelines directed at professionals in occupational health use scientific evidence unsystematically or inadequately. This article assesses the guidelines of the Finnish Institute of Occupational Health (FIOH). We selected a random sample of 29 guidelines from all those published by FIOH, which were assessed by four people individually using the AGREE instrument. The items were scored in six domains: scope and purpose of the guideline, stakeholder involvement, rigour of guideline development, clarity and presentation, application, and editorial independence. Mean domain scores were calculated according to AGREE instructions. The guidelines presented their scope and purpose well; the mean domain score was 62%. Their clarity and presentation was fairly good, mean domain score 47%. The stakeholder involvement’s mean domain score was 33%. The other domains scored low: applicability domain, 15%, rigour of guideline development, 9%, and editorial independence, 7% only. The rigour and reporting of guideline development seems to be the main challenge for future guideline production in FIOH. A common structure for guideline preparation is needed.

Key words: Estimation, Occupational health, Evidence-based medicine, Guideline, Agree instrument

Introduction

Research is constantly discovering new knowledge, which is applicable in daily practice, and this data is usually published in scientific journals. Physicians seldom have time to read scientific journals, or may not be interested in them1–3). It is often not easy to find the required information in the article, or the results in different articles may be conflicting, making interpretation difficult. Therefore, it is important to have reliable summaries of evidence, in the form of evidence-based guidelines, for instance4). There may be several guidelines on the same topic, and if the recommendations differ, it is especially difficult to assess which to use5–7).

Guidelines are published by different institutes and in different formats. The 1990’s saw a great deal of discussion on the quality of the practice guidelines8–10). They should reflect the best recent knowledge, be easy to use, report benefits and drawbacks in a balanced way, and use economic evaluations. Good quality guidelines have the potential to improve clinical practice11). It is not only in the interest of patients to get the best health care currently available, but also for decision-makers, so as to know how best to invest the funds12).

A group of researchers from thirteen countries developed the Appraisal of Guidelines Research and Evaluation (AGREE) instrument in 1998. It was designed to assess medical guidelines prepared by local, national and international groups or public organizations. It estimates the
quality of the preparation process and reporting of guidelines, not the correctness of the content itself\textsuperscript{11, 13, 14}.

The AGREE instrument is generic, and can thus be used for different kinds of guidelines. It has been used in various evaluations with great success. A wide range of people from different cultural and professional backgrounds have found it useful and easy to use\textsuperscript{15}. The AGREE instrument can be downloaded free of charge in several languages from the Internet: (http://www.agreecollaboration.org/)\textsuperscript{13, 16, 17}.

The AGREE instrument contains 23 items categorized in six domains. Every domain defines a separate dimension of guideline quality. Domain one (items 1–3) looks at whether the overall aim of the guideline, relevant clinical questions and target patient population are sufficiently defined. The second domain (items 4–7) assesses stakeholder involvement and usability aspects. The third domain (items 8–14) is about the methods of guideline preparation: how research information was retrieved and selected, whether recommendations reflect evidence, and whether updating procedure is described. The fourth domain (items 15–18) evaluates the language and format of the guideline. The fifth domain (items 19–21) assesses whether the consequences of the guideline application have been considered. The sixth domain (items 22–23) examines possible conflicts of interest among the guideline authors. Each item is scored on a four-point scale where 1 stands for “strongly disagree”, 2 “disagree”, 3 “agree” and 4 for “strongly agree”.

At least two persons should be involved in guideline assessment with AGREE. Each one independently assesses and scores all 23 items. Individual item scores are summed up to form domain scores. The resulting domain score is a percentage of the maximum possible score for that domain. A hundred percent would mean that all assessors gave full scores for every single item in that particular domain. A domain score which lies between 60% and 100% is defined as good, a domain score between 30% and 60% as moderate, and scores lower than 30% are poor.

In the end, each appraiser should give an overall estimate of the guideline; they should state whether or not the guideline is recommended for use in practice. According to the AGREE manual, the appraiser should make the judgement based on the guideline as a whole. It is unlikely that a guideline will rate high on all of the items and domains. The number of highly rated items and the balance between the higher and lower rated domains should determine the overall assessment\textsuperscript{16}.

A good quality guideline, which could be used in practice without alteration, would receive scores greater than 60% in most domains. Ratings between 30–60% in most domains are considered to belong to a guideline that needs modification. If the scores remain under 30% in the majority of the domains, the guidelines are not recommended for use in practice.

We selected the AGREE instrument, as it has been reported to be a reliable and simple way to assess the quality of guidelines\textsuperscript{11, 13, 15, 17, 18}. It is also important to be able to compare the result with others made using the same tool. Several evaluations of practice guidelines, some in the field of occupational health, have been carried out\textsuperscript{19–25}. The purpose of this study is to assess the quality of the occupational medicine practice guidelines of FIOH.

**Subjects and Methods**

All FIOH publications, either printed (books, booklets, leaflets) or on the website, were screened for guidelines in October 2005. The AGREE instrument defines a guideline as a document that includes a set of systematically developed statements (recommendations) to assist practitioner and patient decisions regarding appropriate health care for specific clinical circumstances. We found it difficult to draw a clear line between a guideline and non-guideline document. Until now, FIOH has produced expert-based recommendations and only a few evidence-based (EB) guidelines. We excluded documents with more than a hundred pages, risk assessment instruments, case histories, statistics, project plans, and other documents with no clear guideline properties.

We identified 176 guidelines in October 2005. We listed them in the order they had appeared and chose every sixth to form a set of thirty guidelines. Although we had predefined inclusion and exclusion criteria, one of the thirty had to be excluded after a closer look at the content, while it included no recommendations.

We ended up with 29 guideline documents in the Finnish language. Those selected were published during 1997–2004. Some of them were targeted at occupational professionals, others to workers and/or employers. Certain guidelines were specifically targeted at occupational professionals, workers or employers. Most were suitable for anyone interested in the topic. Thirteen were books (14–100 pages), two were two-page leaflets, and fourteen were websites.

Four appraisers individually assessed the guidelines using the latest version of the AGREE instrument\textsuperscript{13, 16, 26}. Three were medical doctors, specialists (or in the process of specializing) in occupational health. One of them had a research background of critical appraisal of research literature, and was a physiotherapist with experience of guideline preparation. The assessors read through the AGREE training manual and practised with some test appraisals in order to reveal discrepancies and then dis-
cuss them. The appraisers declared no personal conflicts of interest.

**Results**

**Domain 1-Scope and purpose**

All appraisers agreed that the objectives and target groups of the guideline were well documented. Most guidelines (18/29) scored over 60% in this domain. Only two guidelines remained under 30%; one of which was a leaflet. The mean domain score of all guidelines was 61% (95% confidence interval (CI) 55–67%).

**Domain 2-Stakeholder involvement**

Target users were defined well in most guidelines. It was difficult to define whether all relevant professional groups were represented in guideline preparation when their background was rarely stated. There was no indication that the target population’s views and preferences had been examined, and there was rarely any mention of piloting the guideline before use. Twelve guidelines out of the twenty-nine scored moderately (domain scores 30–60%). Fourteen scored low and were assessed as poorly representing the views of intended users. Only three guidelines were classed as good (domain scores more than 60%). The average domain score of all guidelines was 35% (95%CI 28–42%).

**Domain 3-Rigour of development**

The majority of the guidelines had no description of the search methods or selection criteria of the evidence. The methods used for formulating recommendations were left without clarification and there was no explicit link between recommendations and supporting evidence. The balance between benefits and side effects or risks were rarely considered. Only a few guidelines had undergone external expert review before publishing. Updating the procedure was not considered at all. All guidelines scored low in this domain, and the mean domain score was 9% (95%CI 7–11%).

**Domain 4-Clarity and presentation**

Key recommendations were often difficult to identify. In most cases recommendations were considered clear and specific but here the appraisers’ ratings differed the most. Only a few guidelines had tools for application. Five guidelines had good scores, twenty-two moderates, and two guidelines scored poorly in clarity and presentation. The average score in this domain was 47% (95%CI 42–52%).

**Domain 5-Application**

The appraisers could not find a single definition regarding potential cost implications or any details of the administrative impact of applying the guideline. Performance monitoring indicators were presented in some guidelines. Only one guideline scored moderately, all the rest were poor. The mean score for applicability was 15% (95%CI 11–19%).

**Domain 6-Editorial independence**

There was no definition of the editorial independence or the conflicts of interest of the guideline authors. All guidelines scored poorly in this domain. The mean score was 7% (95%CI 3–11%).

**Overall assessment**

All in all, half of the guidelines (14/29) received moderate scores in the majority of their domains. This indicates that they could be suitable for use after some alterations. None of the guidelines had good scores in most domains (Fig. 1). Although half of them scored poorly in most domains, the appraisers still considered them applicable after modification, especially if no better guideline is available.

**Discussion**

This study evaluated samples of guidelines published by the Finnish Institute of Occupational Health (FIOH). The assessment was targeted at the preparation process of the guidelines, rather than the content. FIOH guidelines seem to have most deficiencies in the rigour of development, editorial independence and applicability. One reason for this finding may be that these properties were not reported properly or mentioned at all in the guidelines or their background documents, and the assessors could therefore not consider them correctly. The guideline pro-
duction in FIOH has been mainly unstructured and expert-based; only few have been evidence-based. Until now there has been no common structure for guideline preparation.

There are some limitations in this study. Firstly, we had difficulties in applying the guideline definition to the documents we were screening. A minimum criterion for a document to be included in the guideline pool was that it had to include sentences of a proposing or suggesting nature, or direct recommendations that aim at improved management choice in occupational health care.

Another problem encountered was defining the target group of the guidelines. Only clinical guidelines should be included, i.e. those targeted at professionals. But, without explicit descriptions it was difficult to separate public-orientated guidelines from those targeted at guiding management decisions in occupational health. We thought that the same kind of quality principles should apply to the guidelines of both professionals and lay people and decided not to attempt to divide these groups.

The format of the documents that we defined as guidelines was diverse. We considered whether the format of publishing a document made a difference when defining a guideline and concluded that it did not. A guideline, in its broad sense, is not necessarily a paper or electronic document with a certain amount of pages. It may also look like a leaflet, for instance. We considered that even a leaflet type of document, if it contains recommendations for practice, should be subject to the same quality requirements as traditional guidelines. We recognize that this interpretation made our sample more heterogeneous, and thus the results less applicable.

There are also other assessment tools for guidelines, and some would perhaps have been more appropriate for the evaluation of non-professional guideline documents. We chose to use the AGREE instrument because it is easily available, easy to use and has several references to other evaluation studies in the occupational health discipline.

Some other studies have assessed occupational health guidelines and obtained very similar results to ours\(^{17, 19–21, 24, 29}\). Information on data search, evidence selection criteria, and the method used to form the recommendations from the evidence seem to be lacking in FIOH’s guidelines, as it is in others. Guidelines could be made more applicable by involving relevant users in their preparation. Benefits and limitations of practice guidelines could be considered in pilot projects before publishing. Possible conflicts of interests should always be recorded, and cost-effectiveness should be considered more often.

In the future, FIOH should encourage its guideline creators to adopt a common structure for their preparation, in order to improve consistency and quality\(^{27–31}\).

References

3) Schaafsma F, Verbeek J, Hulshof C, van Dijk F (2005) Caution required when relying on a colleague’s advice; a comparison between professional advice and evidence from the literature. BMC Health Serv Res 5, 59.
14) AGREE Collaboration (Appraisal of Guidelines R, and


