Evaluation of Guideline Quality Using GEM-Q

Abha Agrawal, MD, Richard N. Shiffman, MD MCIS

Yale Center for Medical Informatics, New Haven, CT, USA

Abstract

A variety of rating instruments that evaluate the quality of practice guidelines have been published. Application of these instruments can be difficult and time-consuming. In a literature review, we identified two evaluation instruments that are comprehensive, have clearly defined constructs, and have undergone validation/testing—the Guidelines Quality Assessment Questionnaire (GQAQ) and the Appraisal Instrument for Clinical Guidelines (AICG). Overall, the AICG is more comprehensive. The AICG addresses the implementability of a guideline, which is not evaluated by the GQAQ. However, the GQAQ is more amenable to computerization. GEM-Q is a Guideline Elements Model (GEM)-derived application intended to facilitate automated evaluation of guideline quality using one of the published instruments. To develop GEM-Q, various items in the GQAQ were mapped to corresponding elements in the GEM hierarchy and a customized XSL stylesheet was designed based on this mapping. GEM-Q selectively extracts text components of the guideline relevant to quality evaluation and displays the results in HTML format. GEM-Q was applied to a set of six guidelines to test its reliability. It ranked two guidelines as of "good" quality, two as "intermediate", and two as "poor". In all six instances, GEM-Q ranked guidelines in the same order of quality as the experts who validated the GQAQ. This work demonstrates the feasibility of developing an application to facilitate automated guideline quality evaluation.

Keywords:

Practice guideline; Evaluation; Rating Instrument; Quality Control

Introduction

In the last decade, there has been a surge of interest in the use of guidelines in clinical practice and health policy. The Institute of Medicine has defined clinical practice guidelines as "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" [1]. If appropriately developed, disseminated and implemented, guidelines offer many potential benefits, including reduced variation in practice among physicians [2], improved health outcomes for patients [3], cost containment, and speedy translation of research into practice. However, there is growing concern that many are based on expert opinions that do not always measure up to contemporary standards of scientific evidence [2]. In response to this concern, a variety of rating instruments to evaluate the quality of practice guidelines have been published [5-9].

We describe GEM-Q, an XML-based application that facilitates evaluation of guideline quality based on published quality rating instruments. We also describe a systematic comparison of two of the more comprehensive rating instruments.

What is guideline quality?

Ideally, a guideline's quality should be measured by a prospective evaluation of its effectiveness in achieving intended health outcomes. The Institute of Medicine defines practice guidelines as *valid* if "when followed, they lead to the health gains and costs predicted for them" [5]. However, health outcome evaluation data are lacking for most guidelines. Therefore, a surrogate assessment of guideline quality most often involves an evaluation of the methodology used in developing a guideline and the contents of the resulting guideline document.

Three key principles underpinning the development of high quality guidelines have been consistently emphasized – (a) guidelines should be multidisciplinary, (b) they should be based on a systematic review of published work, and (c) they should explicitly link their recommendations to the supporting evidence [10].

The IOM Guideline Quality Rating Instrument

In 1992, the Institute of Medicine (IOM) developed a provisional assessment instrument to examine both the process used to develop a specific guideline document and the substantive content of the document and its recommendations [5]. It evaluated 8 attributes of guidelines—four concerning the substance of the guideline—clinical applicability or scope, clinical flexibility, reliability/reproducibility and validity, and four concerning the process of guideline development—clarity, multidisciplinary process, scheduled review, and documentation. Of a total of 46 descriptive questions in the instrument, validity was assessed in 22 questions; clarity in 8; multidisciplinary process in 4; clinical flexibility in 4; reliability and reproducibility in 4; clinical adaptability in 3; and scheduled review in 1. By this metric, evaluation of the validity of a guideline was accorded the major emphasis in the instrument. This instrument did not propose a numerical scoring. Responses to each question were typically “yes” or “no”. If the answer was “yes”, a follow-up question asked whether the quality of that information provided was “satisfactory”, “conditionally satisfactory” or “not satisfactory”. If the initial answer was “no”, the follow up question evaluated the significance of the absence of information. There was no cut point against which guidelines might be judged acceptable or unacceptable. If most responses to questions were “satisfactory” (or “unimportant omissions”), then a reasonable assumption was that a guidelines document would be sufficient for most clinical situations.

Although, quite comprehensive, this instrument has proven to be too complex and unwieldy to implement. It required experts in multiple domains, including both the clinical specialty of the guideline topic and guideline development methodology. Based on the IOM instrument, a variety of other rating instruments have been developed and tested to evaluate the quality of practice guidelines.

GEM-Q

GEM-Q is a tool intended to facilitate automated evaluation of guideline quality using a published guideline quality-rating instrument. GEM-Q is one of the many possible applications of the Guideline Elements Model (GEM) [11]. It uses the Extensible Stylesheet Language (XSL) technology for its implementation [12].

GEM is a guideline document model based on the Extensible Markup Language (XML) [13] that can store and organize the heterogeneous knowledge contained in practice guidelines. GEM is a multi-level hierarchy of more than 100 discrete elements in nine major braches – Identity, Developer, Purpose, Intended Audience, Target Population, Method of Development, Testing, Review Plan, and Knowledge Components. The elements are basic units of information that store data and define structure by virtue of their position in the tree structure of the document. GEM can accommodate information at both high and low levels of abstraction. GEM in intended to facilitate translation of natural language guideline documents into a format that can the processed by computers. Use of XML enables computer processing of guideline information, while the documents remain understandable to domain experts. The GEM-Q application utilizes this computer processability of GEM documents to facilitate automation of guideline quality evaluation. Use of XML also offers other compelling benefits – it is platform independent, it separates data or content from presentation, and it provides easy interoperability in transforming data between applications.

Methods

(a) Comparison of guideline quality rating instruments

To better understand the construct of guideline quality, we systematically reviewed the literature regarding guideline quality evaluation instruments. We searched the MEDLINE database from January 1966 through October 2000 using the following search terms: *guideline*, *practice guideline*, *evaluation studies*, *quality control*, *appraisal*, and *rating instrument*. In addition, bibliographies of all relevant retrieved articles were examined. We selected articles describing guideline-rating instruments that (a) appeared to be comprehensive in content, i.e. addressed the major features outlined by the IOM, (b) clearly defined component constructs, and (c) demonstrated validation and testing. From the instruments that were identified, we compared qualitatively the general approach used by the rating instrument, the contents and scope of the instrument, and the usability of the instrument.

(b) Development of GEM-Q

We selected the Guidelines Quality Assessment Questionnaire (GQAQ) to serve as the basis for our automated evaluation tool [7]. To develop GEM-Q, We mapped the 25 quality rating items in the GQAQ to corresponding elements in the GEM hierarchy**.** These GEM elements store specific text components from a guideline document that are used to evaluate whether a guideline meets the rating instrument’s criteria. Based on this mapping, a customized XSL stylesheet was designed which incorporated concepts from GEM and the rating instrument.

As a part of the ongoing GEM project, a variety of natural language guideline documents were marked up as XML files using the GEM structure (http://ycmi.med.yale.edu/GEM). GEM-Q takes a GEM-encoded guideline in XML format as its input. Using a variety of XSL methods, it selectively retrieves text components from the guideline that are relevant to quality evaluation [12]. The resultant output of the GEM-Q application is an HTML document that can be displayed in a web browser.

(c) Testing GEM-Q

We used a test set of six guidelines to evaluate the reliability of GEM-Q (the same guidelines that were used by Shaneyfelt et al. to evaluate the validity of the GQAQ) [7]. These guidelines were marked up as GEM documents using GEM Cutter (available at http://ycmi.med.yale.edu/GEM). GEM Cutter is an XML editing tool that has been developed by our group to improve the convenience, consistency, and efficiency of marking up guidelines. It functions as an XML editor with many GEM-specific enhancements. It applies the GEM document model and provides context-sensitive definition of various GEM elements in the editing window for the convenience of the end-user.

The six guidelines in GEM format were then evaluated for quality using the GEM-Q application. The guidelines were scored as “good”, “intermediate”, or “poor” using the same rating used by the authors of the GQAQ (personal communication, Terrence Shaneyfelt, May 2000). A guideline was ranked as being “good” if it scored >18 points, “intermediate” if 6-18, and poor if it scored 0-5. This ranking was compared with the rating assigned by the experts who validated the GQAQ.

Results

(a) Comparison of guideline quality rating instruments

From our literature review, only the Guidelines Quality Assessment Questionnaire (GQAQ) [7] and the Appraisal Instrument for Clinical Guidelines (AICG) [8] fulfilled all our inclusion criteria.

Shaneyfelt et al published the GQAQ, a 25-item rating instrument to evaluate methodological quality of clinical guidelines. These 25 items were broadly grouped into three categories: guideline format and development (10 items), identification and summary of evidence (10 items), and formulation of recommendations (5 items). Each question was answered using a yes/no format with each ‘yes’ answer adding one point to the overall score. Thus, each guideline could have a maximum score of 25 (and a minimum of zero).

Cluzeau et al published the AICG, a guideline quality evaluation tool containing 37 items. These items were categorized into three conceptual dimensions that could be mapped to the eight IOM attributes. The first dimension—rigor of development (20 items) —assessed the process of development including responsibility for guideline development, composition of the development group, identification and interpretation of evidence, formulation of recommendations, links between evidence and main recommendations, and peer review and updating. The second dimension—context and content (12 items)—evaluated the aims and objectives of the guidelines, the target group, the circumstances for applying the recommendations, the presentation and format of the guidelines, and the estimated outcomes benefits harms and costs. The third dimension—application (5 items) —addressed implementation and dissemination strategies and monitoring.

A comparison of the GQAQ and the AICG indicated the following:

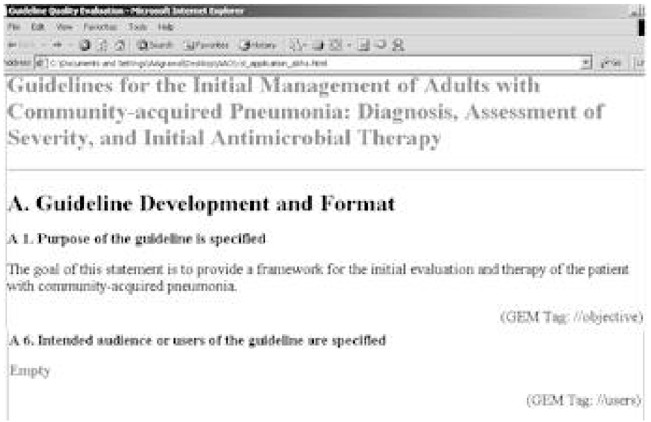
1. The instruments were different in the content and the scope of the evaluation process. Overall the AICG was broader in scope, especially in evaluating whether a guideline document explicitly addressed dissemination, implementation, and monitoring strategies (dimension 3). The GQAQ instrument did not address the implementability of a guideline.
2. The instruments differed in the need for subjective / qualitative assessment of information present in the guideline. The GQAQ evaluated only the presence or absence of information relevant to fulfilling a specific criterion. In the AICG, on the other hand, many items required subjective and qualitative evaluation, not just presence or absence of the information. For example, item 3 in the GQAQ—“The participants in the guideline development process and their area of expertise are specified” —evaluated only the presence of this information in the guideline. However, item 4 in the AICG evaluated— “Is there description of the individuals (e.g. professionals, interest groups-including development group?” This was followed by a related item to evaluate the quality of information— “If so, did the group contain representatives of all key elements? In the AICG, many questions were framed to ask if the information was ‘adequate’, ‘unambiguous’, ‘satisfactory’, or ‘measurable’ thus requiring a subjective evaluation of the information present in the guideline.
3. The instruments provided varying approaches to dealing with uncertain responses. In the AICG, the response to each question could be ‘yes’, ‘no’, ‘not sure’ (reflecting the circumstances for which there is uncertainty) or ‘not applicable’ (when the question may not be relevant). In the GQAQ, the only two possible answers were ‘yes’ and ‘no’ with no provision for uncertainty or relevance of the question.
4. Since the AICG required a qualitative evaluation of information present in the guideline, it was less amenable to automated implementation using a tool like GEM-Q.

(b) Development of GEM-Q

GEM-Q displays the results of a guideline’s quality evaluation in two browser-based formats. The first format displays the criterion under consideration, the text extracted from the guideline to fulfill the criterion, and the corresponding GEM tag used to process that criterion. If a criterion is not fulfilled within a guideline, the word “EMPTY” is displayed (figure 1). The second format is a summary report card of the results of the guideline quality evaluation in a tabular display. The left column indicates the various rating instrument criteria being tested and the right column displays a check mark if a guideline meets the criterion or a cross if the criterion is not fulfilled (figure 2).

(c) Testing GEM-Q

Out of the six test guidelines evaluated using GEM-Q, two ranked as “good”, two as “intermediate”, and two as “poor”. In all six instances, there was 100% concordance between this ranking and that obtained by the experts who evaluated the validity the GQAQ [7].


Guideline Quality Evaluation - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Search Favorites History

Guidelines for the Initial Management of Adults with Community-acquired Pneumonia: Diagnosis, Assessment of Severity, and Initial Antimicrobial Therapy

A. Guideline Development and Format

A 1. Purpose of the guideline is specified

The goal of this statement is to provide a framework for the initial evaluation and therapy of the patient with community-acquired pneumonia.

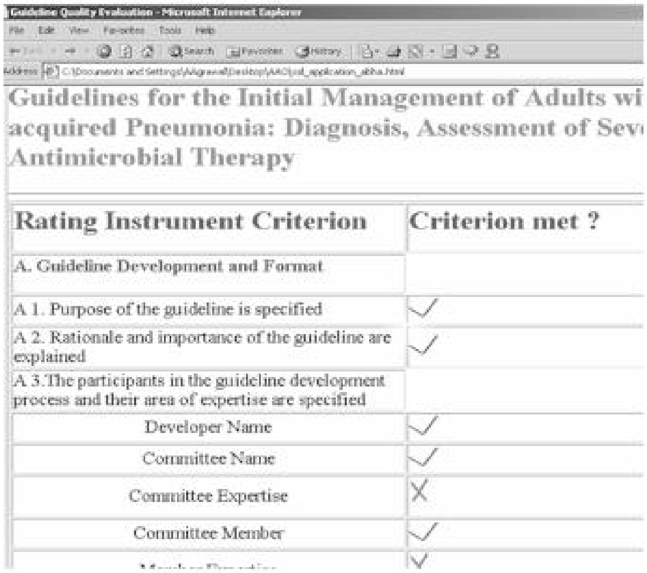
(GEM Tag: //objective)

A 6. Intended audience or users of the guideline are specified

Empty

(GEM Tag: //users)

Figure 1 – Criterion-based display of GEM-Q output


Guideline Quality Evaluation - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Search Favorites History
Address C:\Documents and Settings\AAgrawal\Desktop\AAO\xsl_application_abha.html

Guidelines for the Initial Management of Adults wi
acquired Pneumonia: Diagnosis, Assessment of Sev
Antimicrobial Therapy

Rating Instrument Criterion	Criterion met ?
A. Guideline Development and Format	
A 1. Purpose of the guideline is specified	✓
A 2. Rationale and importance of the guideline are explained	✓
A 3.The participants in the guideline development process and their area of expertise are specified	
Developer Name	✓
Committee Name	✓
Committee Expertise	✗
Committee Member	✓
Member Expertise	✓

Figure 2 – Tabular display of GEM-Q output

Discussion

This study demonstrates the feasibility of an application to facilitate automated guideline quality evaluation. Development, dissemination, and implementation of a guideline are resource-intense endeavors; ensuring that the product is of highest quality is of critical importance. We believe that GEM-Q may be used throughout the guideline life cycle by a variety of stakeholders to improve the likelihood of improving healthcare process and outcomes. Developers may use GEM-Q to evaluate guideline quality prior to widespread dissemination and implementers can evaluate a guideline before adapting it to their institutions.

There are several limitations to our study. The output of guideline quality evaluation using GEM-Q is highly dependent on the quality of markup of the guideline using GEM. Karras et al. [14] found that there was significant inter-rater variability in marking up a guideline using GEM. This may confound the output of GEM-Q evaluation and, consequently, the assessment of guideline quality.

Secondly, in several instances, more than one GEM element maps to a single GQAQ criterion. For example, item 10 in the GQAQ evaluates whether "an expiration date or date of scheduled review" is specified. This corresponds to two elements in the GEM structure—*expiration* and *scheduled.review*. This is because the architecture of GEM is more granular than the individual criterion in the GQAQ. Since there is no partial scoring system implicit in the rating instrument, we gave a full score for the criterion, even if it fulfilled requirements for only one of the corresponding GEM elements. Future enhancement of GEM-Q will address this issue.

Finally, GEM-Q inherits the weaknesses of the GQAQ (or whatever rating instrument is used to customize the XSL stylesheet). The GQAQ is based on a "composite quality the fact that different items may have different scientific and clinical importance. Also these items may have a differential relevance for different guidelines. Secondly, because of its "yes/no" format, the relative quality of a guideline's compliance with a given item can't be assessed.

One of the general limitations of all rating instruments is that they evaluate only the published versions of the guideline document report. Therefore, the result of the guideline quality evaluation is dependent not only on the quality of the guidelines themselves, but also on the quality of the reporting process. Published guidelines may inadequately document the process by which the guideline was produced, and this may affect the results of guideline quality evaluation unfavorably.

The GEM-Q application may be modified to implement rating instruments other than the GQAQ. We used the GQAQ in our quality evaluation application because we felt this instrument captured key features of methodological quality of guidelines in 25 items. The instrument was developed using both literature review and a careful and comprehensive process that included guideline developers, evaluators, implementers, and practicing clinicians. It was also piloted at workshops, pretested by the authors, and validated by experts, who have published articles in guideline methodology. We recognize the potential advantages of the AICG; however, the binary (yes/no) format of the GQAQ is more readily amenable to computerization. We are currently designing a GEM-Q module that applies the AICG.

Conclusion

By providing for customized and selective retrieval of text components from a prose guideline document, GEM-Q can facilitate automated evaluation of guideline quality.

Acknowledgment

This work was supported in part by the National Library of Medicine through grants 1-R29-LM 05552-01A1 and T-15-LM07056 and by the National Institute of Standards and Advanced Technology through grant 70 NANB 7H3035. Dr. Shiffman is a Robert Wood Johnson Foundation Generalist Physician Faculty Scholar.

References

[1] Institute of Medicine. Clinical Practice Guidelines: Directions of a New Program: National Academy Press, 1990.

[2] Woolf SH GR, Hutchinson A, Eccles A, Grimshaw J. Potential benefits, limitations, and harms of clinical guidelines. BMJ 1999;318:527-30.

[3] Grimshaw JM RI. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. Lancet 1993;342:1317-22.

[4] Sudlow M TR. Clinical guidelines: Quantity without

quality. Quality in Health Care 1997;6:60-61.

[5] Institute of Medicine. Guidelines for clinical practice: from development to use; National Academy Press, 1992.

[6] Grilli R MN, Penna A, Mura G, Liberati A. Practice guidelines developed by specialty societies: the need for a critical appraisal. Lancet 2000;355:103-6.

[7] Shaneyfelt TM M-SM, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. JAMA 1999;281:1900-1905.

[8] Cluzeau FA LP, Grimshaw JM, Feder G, Moran SE. Development and application of a generic methodology to assess the quality of clinical guidelines. International Journal for Quality in Health Care 1999;11:21-28.

[9] Scottish Intercollegiate Guidelines Network. SIGN guidelines: an introduction to SIGN methodology for the development of evidence-based clinical guidelines: http://www.sign.ac.uk. Last accessed Nov 27, 2000.

[10] Jackson R FG. Guidelines for clnical guidelines. BMJ 1998;317:427-28.

[11] Shiffman RN KB, Agrawal A, Chen R, Marenco L, Nath S. GEM: a proposal for a more comprehensive guideline document model using XML. JAMIA 2000;7:488-98.

[12] W3C. Extensible Stylesheet Language: http://www.w3c.org/Style/XSL. Last accessed November 28, 2000

[13] W3C. Extensible Markup Language: http://www.w3c.org/Style/XSL. Last accessed November 28, 2000

[14] Karras BT NS, Shiffman RN. A preliminary evaluation of guideline content mark-up using GEM - An XML guideline elements model. Proc AMIA Symp 2000:413-417.

Address for correspondence

40 Temple St, Suite 3D, New Haven, CT 06520, USA

E-mail: abha.agrawal@yale.edu

URL: http://ycmi.med.yale.edu/GEM