A Preliminary Evaluation of Guideline Content Mark-up Using GEM—An XML Guideline Elements Model

Bryant T. Karras, MD, Sujai D. Nath, MD, Richard N. Shiffman, MD, MCIS
Yale Center for Medical Informatics, New Haven, Connecticut

Objective*: To describe application of GEM to analysis and categorization of guideline content.*
Method*: We examined the application of GEM constructs to the AAP guideline on neurodiagnostic evaluation of febrile seizures. Subjects at 4 sites marked-up the guideline content using a hierarchical template that includes branches for identity, developer, purpose, intended audience, method of development, knowledge components, testing, and review. The types of elements used were tabulated. Subjects were surveyed regarding the usability of the model.*
Results*: Eight subjects analyzed the guideline, using between 46 and 149 elements to model its content. There was considerable variation in the application of elements. The number of elements used correlated with time to complete the task. Subjects found application of GEM to be straightforward in 6 of 8 categories and sufficiently comprehensive to model the guideline's information content.*
Conclusions*: Subjects found GEM constructs were able to model the content of the guideline. Improved editing tools will facilitate translation.*

BACKGROUND

Built upon a careful analysis and understanding of research evidence combined with expert consensus, clinical practice guidelines have become increasingly important repositories of knowledge about ideal practice. Extracting and applying that knowledge in tools that support guideline development, dissemination, implementation, and maintenance have proven to be arduous tasks.

In addition to knowledge regarding recommendations for clinical care, guidelines also contain important metadata including the reasons why and methods by which the guidelines were developed, the intended audience of clinicians and the target population of patients toward whom they pertain, and benefits and harms that may be anticipated when they are applied. Informatics [1-5] and health services researchers [6-8] have devised a wide variety of guideline knowledge models that reflect their individual interests and perspectives. Not surprisingly, these divergent models often lack components to comprehensively model guideline content [9].

The cognitive task of translating guideline knowledge into representations that can be processed by computer has been beset with difficulties. Tierney recommended that guideline *developers* structure recommendations "as 'if-then-else' statements with all parameters strictly defined" to assist translation of guidelines into electronic format [10]. Ohno-Machado et al. found "substantial variability" in the encoding of practice guidelines into the Guideline Interchange Format (GLIF) [11]. Patel and coworkers noted that the variability in encoding in GLIF correlated with an individual's prior experience and knowledge of the domain [12].

We hypothesized that text mark-up would be a simpler method for encoding guideline knowledge than the programming task required for creating other computer-processable formats. In this pilot study, we investigate the comprehensiveness and usability of the Guideline Elements Model and the variability of categorization when the model is used.

GEM Overview

GEM (the Guideline Elements Model) is a guideline document model that can store and organize the heterogeneous information contained in practice guidelines [9]. It is intended to facilitate translation of natural language guideline documents into a format that can be processed by computers. GEM is constructed as a hierarchy with 8 major branches—Identity, Developer, Purpose, Intended Audience, Method of Development, Testing, Review Plan, and Knowledge Components*. Each of these headings further arborizes into smaller branches. For example, the Knowledge Components branch can store information about the guideline's Recommendations, Definitions, and Algorithms. GEM is defined using an XML Document Type Definition [13].

In XML terminology, each of the nodes in the model tree is called an *element*. In marking-up a specific document, some elements are used repeatedly (e.g., most guidelines have more than one recommendation) whereas others are used only once (e.g., title), and many elements may not be used at all in a particular guideline.

Each of the elements has a name that is intended to clearly indicate the kind of information it stores. Elements can store information derived verbatim from the guideline or metadata that is inferred about the guideline. Some elements are designed to store

specific terms from a controlled vocabulary from the National Guidelines Clearinghouse [14]. Each element has an XML attribute that designates its source as Explicit, Inferred, or Controlled Vocabulary.

The most complex branch of the model represents Knowledge Components—the guideline's Recommendations, Definitions, and Algorithms. This branch accounts for nearly half of the GEM elements.

A Recommendation element can be composed of Conditional and/or Imperative elements. A Conditional recommendation is appropriate for only a fraction of the target population and can be recognized by use of words like "if", "when", and "whenever". Imperative recommendations, on the other hand, apply to the entirety of the target population.

A Conditional recommendation can be structured as an "if–then" statement. Decision Variable elements store tests and observations that determine the appropriateness of related Action elements. For example, a recommendation states:

> *If the patient appears toxic, a blood culture should be obtained to rule out septicemia.*

In this case, "appears toxic" is stored in a Decision Variable element, "a blood culture should be obtained" is stored in an Action element, and "to rule out septicemia" is stored in a Reason element.

Decision Variables are often tests, which have Sensitivity, Specificity and Predictive Value information that can be stored in Decision Variable sub-elements. Likewise, Evidence Quality and Recommendation Strength elements store information specified by the guideline authors that attest to the validity and importance of individual recommendations. Cost elements may exist within the Decision Variable branch to reflect the cost of a specific test, within the Action branch to indicate the cost of carrying out an action, and as a branch of the higher level Conditional to specify the cost of the entire activity. The relationships among multiple decision variables and action clauses joined by logical ANDs and ORs are stored in the Logic element. The Flexibility element can store information about options available to the guideline user in exercising the appropriate actions. The Link element stores information that relates one Knowledge Component to another.

The second type of recommendation, as mentioned above, is an Imperative; it applies to the target population as a whole. An Imperative recommendation has branches similar to a Conditional's except there are no Decision Variables in an Imperative, and recommended actions in an Imperative are called Directives.

METHODS

Subjects represented a convenience sample of faculty, fellows, and residents, who represent informatics, pediatrics, and internal medicine at 4 institutions (Yale University, University of North Carolina, University of Alabama at Birmingham, and Johns Hopkins University). Inclusion criteria included (a) an interest in guideline translation, (b) access to a personal computer with Internet Explorer 5.0 and Word 95 or later, and (c) facility with Windows tree-views for collapsing and expanding computer-viewed data. Users were compensated $50 for participation. The project was approved by the Yale University Human Investigations Committee.

Subjects analyzed the American Academy of Pediatrics Practice Parameter on Neurodiagnostic Evaluation of the Child With a First Simple Febrile Seizure [15]. This guideline was chosen because in spite of its brevity (3 pages) it includes a wide variety of elements and recommendations with both conditional and imperative statements and is representative of a wide variety of national guidelines recently created using evidence-based methods.

One of the authors (BTK) oriented each subject to the Guideline Elements Model and the required task using a 3-page prepared text. Orientation required 15-20 minutes per individual. For those participants who were off-site, orientation was completed via telephone. Subjects were provided with an electronic version of the guideline (minimally abridged to remove redundant elements such as the names of several committee members) and a GEM template supplied as Microsoft Word outline file. Participants were asked to copy and paste text from the guideline into pertinent elements in the template and to add additional metadata as appropriate. Subjects were specifically instructed to analyze composite items into individual elements and to replicate branches of the template sufficiently to atomize guideline content.

Subjects also completed a demographics and skills survey before the task and a satisfaction survey afterwards. The satisfaction survey included subjective and objective assessments regarding how long the mark-up process took, which parts of the mark-up process were perceived as frustrating or straightforward, suggestions for modifications of GEM, and an overall assessment of the expressive adequacy of the model. Responses were collected confidentially and de-identified by a research associate, who is not otherwise involved with the study.

Survey responses were tabulated. For each of the 8 major branches of the GEM hierarchy, the number of *element types* used and the total number of elements were counted. Within the Knowledge

Components section, specific counts were made of the number of recommendations and within each recommendation the number of conditionals and imperatives was ascertained. SPSS (Version 8) was used for statistical analyses.

RESULTS

Eight subjects marked-up the guideline, 5 of whom had no involvement in the development of the model. Three were informatics faculty at 3 different institutions. Three of the subjects had pediatrics training, 2 were trained in internal medicine; the others were a neurologist, an anesthesiologist, and a graduate student. Satisfaction surveys were collected from the 5 subjects who were not involved in model development.

Time to Complete

The time to complete the task ranged from 90 to 169 minutes (median of 115). Time to complete the task was associated with the number of elements used. (reflecting a "lumpers" vs. "splitters" phenomenon). Two subjects made extraordinary efforts to complete the task, even extending beyond the tasks' parameters, e.g., going on-line to identify guidelines not mentioned in the AAP guideline to fill the Comparable Guidelines element. As might be expected, their efforts tended to take more time. The total *number of elements* highly correlated with *time* to complete (Pearson's r of 0.823 with a significance of .012—seeTable 1).

	Subjects							
	A	B	C	D	E	F	G	H
Identity	2	5	1	7	1	6	3	4
Developer	10	5	6	15	0	6	5	4
Purpose	28	8	7	7	12	7	4	5
Audience	10	3	4	12	11	1	1	2
Method	8	6	5	4	3	6	4	4
Knowledge	91	17	65	60	40	26	56	33
Testing	0	2	0	2	0	0	3	0
Review	0	0	0	0	0	0	0	0
TOTAL	149	46	88	107	67	52	76	52
Time (min)	169	95	109	123	130	90	105	120

Table 1. Total number of elements used and time to complete task by subject (A-H) and GEM category.

XML Template Analysis

We tabulated for each participant, both the number of *unique* elements and the total number of elements they used to categorize guideline content for each major branch of the GEM hierarchy. Because the number of available elements in each category varies widely, we calculated the percentage of available elements that each subject applied (Table 2).

The percentage of unique elements used to represent content from this guideline varied widely from user to user. Although all agreed that there was no content relating to Review, all other categories showed marked variation in the use of GEM elements. Only in the Audience category were all elements used and this was only by 3 subjects.

	Subject							
	A	B	C	D	E	F	G	H
Identity (10)	20%	50%	10%	70%	10%	60%	30%	40%
Developer (9)	56%	56%	56%	78%	0%	44%	56%	22%
Purpose (9)	89%	89%	78%	78%	67%	78%	44%	56%
Audience (3)	100%	100%	67%	100%	67%	33%	33%	67%
Method (14)	50%	43%	29%	29%	21%	43%	29%	29%
Knowledge (46)	52%	37%	22%	20%	15%	20%	20%	13%
Testing (3)	0%	67%	0%	67%	0%	0%	67%	0%
Review (2)	0%	0%	0%	0%	0%	0%	0%	0%
OVERALL	51%	48%	30%	41%	20%	34%	29%	24%

Table 2. Percent of unique elements used by subject (A-H) and GEM category (total elements).

The same variability noted in the use of unique elements carried over into our tabulation of the total absolute number of elements used to store content from the guideline. In this case, however, larger numbers represent more complete atomization of the content into repeated elements of the same type (Table 1).

There was disagreement as to the analysis of recommendations and their placement in the hierarchy. The guideline included a section titled *Recommendations* with four subsections—*Lumbar Puncture, EEG, Blood Studies, Neuroimaging*. Some subsections had multiple statements presenting separate conditional or imperatives.

	Subjects							
	A	B	C	D	E	F	G	H
Recommendation	4	1	9	6	4	4	4	4
Conditional	4	1	7	5	3	4	6	7
Imperative	3	1	4	3	3	0	4	0
Decision Variable	5	1	5	2	3	2	6	6

Table 3. Number of Recommendation elements and sub-elements used by subject (A-H)

Although most subjects conceptualized 4 major recommendations (as the guideline text did) some made each conditional or imperative statement a separate recommendation (Table 3). There was also disagreement among the participants as to whether a statement was a conditional or imperative. This is surprising in that all subjects found the distinction between the two constructs to be "straightforward".

A closer look at content showed dramatic style variation in how the same information was marked-up from one subject to another. Word count analysis noted a dichotomy among the participants. Some placed considerably more text into each element than did others who sought to abridge and abbreviate. For example one subject identified the reason for a particular recommendation as:

> ***clinical signs and symptoms of meningitis may be subtle***

while another subject included considerably more guideline text for the same element:

> *because* ***clinical signs and symptoms of meningitis may be subtle … In*** *approximately 13% to 16% of children with meningitis, seizures are the presenting sign of disease, and in approximately 30% to 35% of these children (primarily children younger than 18 months), meningeal signs and symptoms may be lacking…An increased risk of failure to diagnose meningitis occurs in children: (1) younger than 18 months who may show no signs and symptoms of meningitis; (2) who are evaluated by a less-experienced health care provider; or (3) who may be unavailable for follow-up.*

Survey Analysis

The subjects agreed that the GEM hierarchy was comprehensive enough to represent all (n=2) or most of (n=3) the information content of the practice guideline. Four of 5 subjects agreed with a statement that placement of text into Identity, Developer, Purpose, Intended Audience, Testing, and Revision Plan elements was "straightforward". Four of 5 found that analysis of Knowledge Component elements was confusing. Although most had no difficulty identifying actions and distinguishing conditionals from imperatives, 3 of 5 participants reported difficulty with identification of decision variables. There was considerable variation in the overall assessment of the straightforwardness of the task: 3 of 5 found it to be straightforward while 2 disagreed.

Participants expressed a need for clearer definition of the individual elements in the hierarchy and expressed concern about proper placement of text within elements

> *Although I was able to find a place for everything in the guideline, in one or two places I found myself assigning very different bits of information to the same term (element). I'm not sure where the balance between flexible enough to include most and so flexible that the terms lack precision is.*

Another commented

> *The definition of each element needs to be much clearer, perhaps by providing more examples…it seemed that some items fit in more than one place, but in neither place well.*

As evidenced by the wide variation in the time to complete the task, considerable variability was noted in the effort expended by individual participants. Some users did research beyond the described requirements of the task in an effort to fill empty elements. For example one subject went to the NGC web site and tried to find the URL of the task guideline in order to fill an Identity element.

Some users were inconsistent in the designation of tagged material as explicit, inferred or controlled vocabulary. Such distinctions could be determined automatically by a tool that assigns these attributes to the element and eliminate human error.

Subjects recognized that current limitations would be improved by a customized XML editing tool.

> *The major modifications will need to be in the tools for mark-up. The hierarchy worked well, once I got the hang of it.*

DISCUSSION

GEM is intended to be used throughout the entire guideline lifecycle to model information pertaining to guideline development, dissemination, implementation, and maintenance. Information at both high and low levels of abstraction can be accommodated. Use of XML facilitates computer processing of the guideline information.

In this pilot study we found that subjects from a number of institutions and backgrounds felt the GEM model was sufficiently comprehensive to model the information content of the practice guideline that we tested. However, there was substantial variation in their use of elements and the atomization of concepts. They had particular difficulties in analyzing and categorizing recommendations, the content necessary for electronic guideline implementation.

The subjects pointed out a need for clearer definition of model components, indicating a role for improved training and mark-up tools. Their difficulties may also reflect factors extrinsic to the model, such as their understanding of the task, their background knowledge, their motivation and/or the underlying complexity of the guideline that was chosen. Further testing and analysis will be necessary to sort out these confounders.

The generalizability of this study is limited by its small sample size, and the fact that only a single guideline was analyzed. In addition, information

derived from the surveys depended on self-reporting, which may be a source of error. Also, time to complete task is, at best, an indirect measure of usability and may be more reflective of extrinsic factors.

Future Directions

This preliminary evaluation represents an early stage in the iterative refinement of the GEM model and the specification of GEM-related tools. As a result of feedback from the participants, the GEM hierarchy has been modified to simplify placement of content and definitions are being tightened. The study has helped to define functional requirements for an XML guideline editor and to clarify factors that will be used to evaluate its success. GEM Cutter, currently in development, is a computer application to facilitate mark-up of a prose guideline document into proper GEM, automating the generation of XML output. To diminish some of the marked variability noted in this study, it will be necessary for GEM Cutter to offer guidance to the inexperienced user that facilitates consistent mark-up. Improved training materials, contextual help, and examples will be key to the success of GEM.

Ultimately, our goal is for GEM to be used to facilitate translation of guidelines into formats that can be processed by computer without requiring programming knowledge. GEM-encoded XML documents could be used to verify the completeness and consistency of proposed guidelines, to facilitate Web dissemination of guideline knowledge, and to interact with clinical databases to provide guideline-based decision support.

For more information about this model:

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

* Since this study was performed, GEM has been revised to contain 9 major branches with the addition of elements for Target Population.

Acknowledgements

The authors appreciate the contributions made by Steve Downs, Tom Heil, Kevin Johnson, Andy Spooner, Nick Tosches, G. Weissman, and Cindy Brandt, and recognize significant contribution made in the development of GEM by Roland Chen, Abha Agrawal and Luis N. Marenco. This work was supported by NLM grants 1 R29 LM 05552 and T-15 LM-07056 and NIST award NANB 7H3035. Dr. Shiffman is a Robert Wood Johnson Generalist Physician Faculty Scholar

References

1. Hripcsak G, Ludemann P, Pryor, et al. Rationale for the Arden Syntax. Computers and Biomedical Research 1994;27:291-324.

2. Lobach DF, Gadd CS, Hales JW. Structuring clinical practice guidelines in a relational database model for decision support on the internet. In: Masys D, ed. Proc AMIA Annu Fall Symp. Nashville, TN: Hanley and Belfus, 1997: 158-62.

3. Gordon C, Herbert I, Johnson P. Knowledge representation and clinical practice guidelines: the DILEMMA and PRESTIGE projects. In: Medical Informatics Europe '96: human facets in information technologies. Amsterdam: IOS Press, 1996: 511-5.

4. Dolin RH, Alschuler L, Biron PV, et al. Clinical practice guidelines on the internet: a structured, scalable approach. MD Computing 1999;16:60-4.

5. Tu SW, Musen MA. A flexible approach to guideline modeling. In: Lorenzi N, ed. Proc AMIA Symp. Washington, DC: Hanley & Belfus, 1999: 420-4.

6. Institute of Medicine. Guidelines for clinical practice: from development to use. Washington, DC: National Academy Press, 1992.

7. Hayward RSA, Wilson MC, Tunis SR, Bass EB, Rubin HR, Haynes RB. More informative abstracts of articles describing clinical practice guidelines. Ann Intern Med 1993;118:731-7.

8. 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;281:1900-5.

9. Shiffman RN, Karras BT, Agrawal A, Chen R, Marenco L, Nath S. GEM: a proposal for a more comprehensive guideline document model using XML. J Am Med Informatics Assoc 2000 (in press).

10. Tierney WM, Overhage JM, Takesue BY, et al. Computerizing guidelines to improve care and patient outcomes: the example of heart failure. J Am Med Inform Assoc 1995;2:316-22.

11. Ohno-Machado L, Gennari JH, Murphy SN, et al. The guideline interchange format: a model for representing guidelines. J Am Med Informatics Assoc 1998;5:357-72.

12. Patel VL, Allen VG, Arocha JF, Shortliffe EH. Representing clinical guidelines in GLIF: individual and collaborative expertise. J Am Med Informatics Assoc 1998;5:467-83.

13. W3C. Extensible Markup Language (XML). www.w3c.org (2/28/00).

14. AHRQ. National Guidelines Clearinghouse. www.guideline.gov (2/15/00).

15. American Academy of Pediatrics. The neurodiagnostic evaluation of the child with a first simple febrile seizure. Pediatrics 1996; 97:769-72.