Support for Guideline Development through Error Classification and Constraint Checking

Mor Peleg, Ph.D.[1], Vimla L. Patel, Ph.D., D.Sc.[2], Vincenza Snow, M.D.[3], Samson Tu, M.S.[1], Christel Mottur-Pilson, Ph. D.[3], Edward H. Shortliffe, M.D., Ph.D.[2], and Robert A. Greenes, M.D., Ph.D.[4]

[1]Stanford Medical Informatics, Stanford University School of Medicine, Stanford, CA
[2]Department of Medical Informatics, Columbia University, New York, NY
[3]American College of Physicians-American Society of Internal Medicine, Philadelphia, PA
[4]Decision Systems Group, Harvard Medical School, Brigham &Women's Hospital, Boston, MA

Clinical guidelines aim to eliminate clinician errors, reduce practice variation, and promote best medical practices. Computer-interpretable guidelines (CIGs) can deliver patient-specific advice during clinical encounters, which makes them more likely to affect clinician behavior than narrative guidelines. To reduce the number of errors that are introduced while developing narrative guidelines and CIGs, we studied the process used by the ACP-ASIM to develop clinical algorithms from narrative guidelines. We analyzed how changes progressed between subsequent versions of an algorithm and between a narrative guideline and its derived clinical algorithm. We recommend procedures that could limit the number of errors produced when generating clinical algorithms. In addition, we developed a tool for authoring CIGs in GLIF3 format and validating their syntax, data type matches, cardinality constraints, and structural integrity constraints. We used this tool to author guidelines and to check them for errors.

1 Introduction

The aim of evidence-based clinical guidelines is to eliminate errors, reduce practice variation, and encourage best practices in clinical medicine. Guideline implementations best affect clinician behavior if they deliver patient-specific advice during patient encounters[1,2]. Therefore, several groups have been developing methodologies for representing CIGs that can be linked to patient data[3]. Our group, InterMed (a consortium of researchers at Harvard, Stanford, and Columbia universities), has developed a language called GLIF3[4] and supporting tools for representing CIGs and sharing them.

A report of the Institute of Medicine (IOM), *To Err is Human*[5], sets forth an agenda for reducing medical errors and improving patient safety through the design of a safer health system. The report recommends that health professional licensing bodies recognize patient safety in practice guidelines. Because clinical guidelines aim to reduce practice errors, it is extremely important that narrative guidelines and CIGs not contain errors. One way to reduce errors in guidelines and CIGs is to subject them to rigorous trials before releasing them.

Another way to reduce errors is to assess guideline quality with specially designed assessment tools and to assure that specified attributes have been incorporated. The 1992 IOM report on the development of clinical guidelines[6] suggests eight attributes for assessing guideline quality. Four attributes relate to guideline content: validity, reliability and reproducibility, clinical applicability, and clinical flexibility. The other attributes relate to the process of guideline development or representation: clarity, multidisciplinary process, scheduled review, and documentation. A variety of guideline assessment tools have been published[6,7]. These tools evaluate guidelines according to desirable attributes that can be mapped to the IOM attributes. One tool, GEM-Q, is an application derived from the Guideline Elements Model (GEM)[7]. GEM-Q is intended to facilitate guideline quality evaluation based on published quality rating instruments that can be mapped to GEM elements. First, a guideline is marked-up in GEM. Then, GEM-Q selectively retrieves marked text components that are relevant for quality evaluation.

We are taking a complementary perspective to these approaches. Our approach is to identify the types of errors introduced during the CIG development process, identify their sources, and to devise methods and tools for limiting them. Traditionally, CIGs are created based on published narrative guidelines. Unfortunately, many considerations that are important for automating guidelines are not explicitly considered when the narrative guidelines are being developed. This situation creates great difficulties when the guidelines are to be encoded in a CIG formalism, such as GLIF3. In recognition of this problem, we are collaborating with the guideline development team from the American College of Physicians – American Society for Internal Medicine (ACP-ASIM) on studies that examine the process of narrative guideline creation. Specifically, we are studying the last step of this process, in which the ACP-ASIM team creates a

clinical algorithm in diagrammatic form based on the guideline text. We are introducing computer-science modeling considerations into these studies. This process will assist us in refining our authoring tools and it will assure robustness in GLIF3. Like the developers of the PRO*forma* CIG modeling methodology,[8] we also believe that decision-support systems should have design features aimed at ensuring safety.

With these issues in mind, we developed an authoring tool for GLIF3 that allows a CIG developer to encode a narrative guideline as a flowchart of guideline steps containing formal definitions of medical actions, decision and eligibility criteria, patient states, and control flow[9]. We also developed a module for the authoring tool for validating GLIF3-encoded guidelines. The tool can find errors in guideline specification by checking syntax, data type matches, cardinality constraints, and structural integrity constraints.

2 Methods

InterMed investigators observed ACP-ASIM experts as they created flowchart versions of clinical algorithms based on narrative guidelines that they had created previously. The two guidelines studied were *Pharmacological Management of Acute Attacks of Migraine Headache*[10] and *Pharmacological Management for Prevention of Migraine Headache*[11]. We recorded the experts as they "thought aloud" about what they were doing, and we also captured any conversations with the investigators and other guideline creators. We kept all the drafts of the algorithms (seven drafts of the first algorithm and three of the second), which showed progressive changes in the algorithms. The Institutional Review Board reviewed the protocol and approved it.

We used a classification scheme proposed by Knuth[12] to classify changes between narrative guideline text and the clinical algorithm produced from it.

We used Protégé-2000[13] to develop an authoring and validation tool for GLIF3. Protégé enables defining allowed data types and checking them, establishing cardinality constraints, and setting lower and upper limits on numerical values. We used Protégé's axiom language[14] to define constraints in a subset of first-order predicate logic written in the Knowledge Interchange Format syntax. We also checked two GLIF3-encoded guidelines[15,16] for unmet constraints.

3 Results

We describe the process of algorithm creation and our analysis of changes that were made between versions of the algorithms. We then describe our GLIF3 authoring and validation tool.

3.1 The ACP-ASIM process of creating algorithms

After the ACP-ASIM team has created a narrative guideline, it constructs a clinical algorithm and publishes it at www.acponline.org. The algorithm is first created by a medical expert, who reads the narrative guideline and creates versions of the algorithm in an iterative manner until she is satisfied with the results. No computer-based tools are used during this process. When finished, she delivers a clean copy to the director of scientific policy (DSP) at the ACP-ASIM. This person examines the algorithm and compares it to the narrative guideline. Then, in a face-to-face meeting with the expert, the DSP suggests clarifications and changes. The expert modifies the clinical algorithm and hands it over to a third member of the team, who uses software to generate flowcharts from the expert's hand-drawn algorithms. The expert then checks the flowcharts.

3.2 Changes made by the medical expert

In our analysis, the medical expert made the following types of changes between algorithm versions:

1. Logic changes, such as switching the order of steps or making a decision criterion more general or more specific. For example, "ischemic heart disease" was generalized to "contraindications to triptans".

2. Adding details, such as listing the preferred formulary drugs of a drug group.

3. Complexity management that reduced the number of steps in each algorithm by breaking an algorithm into several parts (nesting) or by identifying sequences of steps that occur in different paths of the algorithm and specifying these steps once. Links are used to connect each path to the sequence of steps.

4. Adding/omitting information – The expert frequently added guideline steps and sidebars. Occasionally, some information was not carried over to the next version. In most cases, the expert noticed this problem in subsequent versions.

3.3 Changes made by the DSP

The DSP made mostly clarification changes. For example, she suggested that drugs be arranged alphabetically when no ranking was available, so as to avoid the false impression that the drugs were ranked by preference. She also ensured that all terms used by the guideline were clearly defined and that details of actions, such as patient education, were provided. When a term could not be defined by formal criteria (e.g., "good response") she asked that this fact be explicitly stated. In addition, she suggested changing the organization of the two related algorithms so that

the one for acute migraine headache would be encountered first, and that it would link to the preventive treatment algorithm. In addition, the second algorithm would link back into the first algorithm.

3.4 Changes made by the flowchart designer

The medical expert created clinical algorithms containing boxes and arrows. She used a single type of box to represent medical actions, patient states, and decisions. Decisions were also represented as text written over arrows. The flowchart designer used commercial software to generate flowcharts. She classified the algorithm's boxes into different types of guideline steps. The step types were defined by instructions for creating algorithms as written by the Agency for Healthcare Research and Quality[17]. The step types were similar to those used by GLIF3, and included decisions, actions, patient states, start and end steps (represented by GLIF3's patient state step), and subguidelines. GLIF3 supports other kinds of steps that are not supported by the clinical algorithms. They include branch and synchronization steps that enable parallel paths and choice steps that define decision rules for and against decision alternatives.

Although the medical expert's clinical algorithm included only one type of step (box), the flowchart designer had little difficulty interpreting the kinds of steps that each box represented, because she relied on the textual information written in the boxes and clarifications that were made by the medical expert. Occasionally, the flowchart designer needed to split a box into two steps, or add a decision that was implicit in the model created by the medical expert.

3.5 Comparing algorithms to narrative guidelines

The medical expert developed a clinical algorithm for guiding physicians during patient encounters. Where evidence was not available, she included guideline steps and sidebars that were based on expert opinion. As a result, discrepancies between the original guideline text and the clinical algorithms most often involved additions that better reflected the flow of an actual patient encounter, rather than omissions.

We classified the changes between the original narrative guidelines and the final version of the clinical algorithms according to Knuth's classification scheme[12]. Knuth classified discrepancies between the requirements document for TeX and the resulting software. Twelve of the 15 change types that he suggested are applicable to the narrative guideline domain. The narrative guideline is analogous to the requirements document, whereas the clinical algorithm created from it is analogous to the software. We added *specialization* as a possible change type. Table 1 summarizes the changes found in the guideline for treating acute migraine headache.

Change type	#	Example	
		Narrative statement	Algorithm specification
Better user interaction	5	Follow up visit was not recommended	Follow up visit was added to the algorithm
Clarity (resolving ambiguity)	4	Decision: in patients whose migraine attacks have not responded to NSAIDs	Decision: in patients whose migraine attacks have not responded to NSAIDs, in past or current attack
Quality improvement	2	Previous history of triptans use not considered	Adding decision if the patient used a triptan with good response then recommending the same drug
Omission	2	IV metoclopramide is recommended as monotherapy for acute migraine treatment	IV metoclopramide is recommended as an antiemetic
Generalization	2	Eligibility: acute migraine attacks	Eligibility: acute headache
Specialization (narrowing a definition of a term)	1	Is the patient a candidate for prevention therapy? (Conjunction of several criteria)	Only one criterion: does patient have more than two attacks per week? If yes, refer to preventive guideline, which checks for all of the criteria
Blunder (e.g., confusing 'before' with 'after')	1	Once preventive treatment is underway, avoid overuse of acute medications	If the patient is on preventive treatment he exits the acute treatment algorithm. However, the intention was to allow the use of acute treatment (without overuse) while preventive treatment is under way

Table 1. Changes between the narrative guideline for treating acute migraine headache and the clinical algorithm derived from it. '#' designates the number of changes of each change type. Change types that were not found include: (1) algorithm awry, (2) misusing the modeling language, (3) mismatch between algorithms, (4) promotion of document organization, (5) considering a surprising scenario, and (6) typographical error.

3.6 Guideline authoring and validation tool

The ACP-ASIM team's flowcharting software was not specifically designed for authoring guidelines, and therefore does not support validating a clinical algorithm's content and logic. Clinical terms are defined in sidebars, using natural language. There is no support for checking that all the terms mentioned in the names of the guideline steps have been defined. Importantly, to create a GLIF3 guideline specification, all of these terms need to be defined by codes taken from controlled medical terminologies. Tools that ensure that every term in a clinical algorithm has a code would greatly ease the process of converting a narrative guideline into a CIG.

We created a GLIF3 authoring and validation tool using Protégé-2000. We configured our tool in two ways. One configuration is used for creating abstract flowcharts, while the other enables a detailed computable specification containing formal definitions of decision criteria and action specifications. The first configuration allows a guideline author to specify a clinical algorithm, codes of clinical terms, rules for ranking alternative treatment options written in natural language, and documentation attributes. Documentation attributes do not specify guideline logic but contain information that is important for guideline validity. Examples include links to support material, the target audience of the guideline, and strength of evidence associated with each guideline step. Specifying strength of evidence is one of the measures of guideline validity that the IOM defined in its report[6].

To support validation, we used Protégé-2000's axiom language (PAL) to define logical constraints. Figure 1 shows an example of a PAL constraint. Table 2 shows the errors found when we applied the constraints to four guidelines that the first author of this paper encoded in GLIF3. No errors were found in the headache guidelines. This may be due to experience gained by the guideline encoder.

Range

```
(defrange ?decision_step :FRAME Decision_Step)
(defrange ?option1 :FRAME Decision_Option)
(defrange ?option2 :FRAME Decision_Option)
```

Statement

```
(forall ?decision_step
  (and (own-slot-not-null options ?decision_step)
    (exists ?option1  (exists ?option2
    (and (options ?decision_step ?option1)
      (own-slot-not-null destination ?option1)
      (options ?decision_step ?option2)
      (own-slot-not-null destination ?option2)
      (/= (destination ?option1) (destination ?option2)))))))
```

Figure 1. A PAL constraint specifying that a decision step have at least two decision alternatives (decision options with destination guideline steps).

Table 2. Integrity constraints that were not met by two guidelines encoded in GLIF3. e(C), e(S), and e(H) mark the number of unmet constraints in the cough, stable angina, and two headache guidelines.

Integrity constraint	e(C)	e(S)	e(H)
A decision step should link to ≥2 options	7	10	0
A branch step should link to ≥2 branches	0	0	0
A synchronization step should not immediately follow a branch step	10	0	0
A guideline step must not connect to a step from a different algorithm	0	0	0
A step must be part of an Algorithm	0	2	0

4 Discussion

Currently, CIGs are developed based on narrative guidelines. Introducing computer-science related design issues into the process of generating narrative guidelines and clinical algorithms could make a narrative guideline less ambiguous and ease the process of generating CIGs. The ACP-ASIM team that develops clinical algorithms is already considering issues that are important for automating guidelines. These issues include: (1) validity of the algorithm, (2) reproducibility (i.e., the algorithm would always give the same behavior for the same patient situations), (3) clinical applicability (i.e., eligibility criteria), (4) clinical flexibility – considering different patient scenarios that occur during a patient encounter, (5) a clear definition of clinical terms, decision points, and medical actions, (6) a clear definition of control flow, (7) logical and easy-to-follow modes of presentation, and (8) distinction among clinical decisions, actions, patient states, and entry and exit points of the algorithm. The ACP-ASIM team uses a process of algorithm development that involves several people and several stages. This process includes face-to-face meetings and discussions of the algorithms that help spot errors or lack of clarity.

Despite the rigorous process of algorithm development, the informatician in our team, who is the first author of this paper, still found places in the algorithm requiring changes. The ACP-ASIM team agreed with these changes, which included adding definitions of terms and altering the control flow as a result of considering patient situations not addressed by the medical expert. It will next be important to study whether using the GLIF3 authoring and validation tool for creating clinical algorithms helps the ACP-ASIM team create algorithms that are valid and clear. It will also be interesting to see whether using such tools during the process of creating the narrative guideline will result in improved guidelines. Nevertheless, our approach cannot replace such complementary approaches as testing with sets of actual cases and testing in clinical practice settings.

Other tools might limit errors that result from forgetting to represent part of the narrative guideline or to copy part of the clinical algorithm or sidebars to the next version. A tool like GEM-Cutter[7] could be used to mark-up narrative guidelines using GEM elements. The tool could be used to view unmarked parts of the guideline, thus aiding in limiting omission errors.

We used the Knuth's classification scheme to categorize changes between a narrative guideline and the final version of its clinical algorithm. Although the classification scheme was developed for changes be-

tween requirements documents and software products, we found it appropriate for categorizing changes between narrative guidelines and clinical algorithms, as well as errors in CIG specifications. A different classification scheme was developed by Tierney and coauthors, who described the problems encountered while they encoded a heart failure guideline[18]. They found that the guideline often lacked definitions of terms and branch points, did not focus on errors of commission, and did not account for comorbid conditions, concurrent drug therapy, or the timing of most interventions and follow-up. Because our study did not examine implementation issues or the development of the narrative guideline prior to clinical algorithm generation, we only considered the first of these problem types, although the implementation of two other guidelines developed by the ACP-ASIM team has been reported by Patel et al[19].

Tierney and coauthors proposed recommendations to improve guidelines based on problems they identified. By looking specifically at the process of algorithm creation and following it closely, we can add more recommendations: (1) make sure that all relevant information is carried from the narrative guideline to all versions of the clinical algorithm, (2) provide all the information necessary to rank treatment options, and (3) consider different patient scenarios.

The life cycle of developing, implementing, and using CIGs includes several phases and involves many individuals. This paper describes our work in reducing guideline errors by looking at specific examples and examining the phases of algorithm creation and CIG encoding. Additional work can be done to examine other life-cycle phases.

5 Acknowledgements

Supported in part by Grant LM06594 from the National Library of Medicine and by the Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Material Command.

References

1. Overhage JM, Tierney WM, Zhou XH, McDonald CJ. A randomized trial of "corollary orders to prevent errors of omission." *JAMIA* 1997;4(5):364-375.
2. Shea S, DuMouchel W, Bahamonde L. A meta-analysis of 16 randomized controlled trials to evaluate computer-based clinical reminder systems for preventative care in the ambulatory setting. *JAMIA* 1996;3(6):399-409.
3. Elkin P, Peleg M, Lacson R, et al. Toward standardization of electronic guideline representation. *MD Computing* 2000;17(6):39-44.
4. Peleg M, Boxwala A, Ogunyemi O, et al. GLIF3: The evolution of a guideline representation format. *Proc AMIA Symp* 2000:645-649.
5. Kohn LT, Corrigan JM, (eds). To err is human: building a safer health system. Washington DC: Committee on Quality of Health Care in America, Institute of Medicine, National Academy Press; 1999.
6. Field MJ, Lohr KN. Guidelines for clinical practice: from development to use. Washington DC: Institute of Medicine, National Academy Press; 1992.
7. Argawal A, Shiffman RN. Evaluation of guideline quality using GEM-Q. *Medinfo* 2001:1097-1101.
8. Fox J, Das S. Safe and sound. AAAI Press; 2000.
9. Patel VL, Branch T, Wang D, Peleg M, Boxwala AA. Analysis of the process of encoding guidelines: an evaluation of GLIF3. *Methods of Information in Medicine,* in press 2002.
10. Snow V, Weiss K, Wall EM, Mottur-Pilson C, for the AAFP and the ACP-ASIM. Pharmacological management of acute attacks of migraine headache: clinical practice guideline. *Annals of Internal Medicine*, in press 2002.
11. Snow V, Weiss K, Wall EM, Mottur-Pilson C, for the AAFP and the ACP-ASIM. Pharmacological management for prevention of migraine headache: clinical practice guideline. *Accepted for publication by Annals of Internal Medicine* 2002.
12. Knuth DA. The errors of TeX. *Software - Practice and Experience* 1989;19(7):607-685.
13. Grosso WE, Eriksson H, Fergerson R, Gennari JH, Tu SW, Musen MA. Knowledge modeling at the millennium (The design and evolution of Protege-2000). The 12th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop 1999:7-4:1-36.
14. Tu SW, Musen MA. Modeling data and knowledge in the EON Guideline Architecture. *Medinfo* 2001:280-284.
15. Irwin RS, Boulet LS, Cloutier MM, et al. Managing cough as a defense mechanism and as a symptom, a consensus panel report of the ACCP. *Chest* 1998;114(2):133S-181S.
16. ACC/AHA/ACP-ASIM. Guidelines for the management of patients with chronic stable angina. *J Am Col Cardiol* 1999;33:2092-2197.
17. Banks NJ. Using Clinical Practice Guidelines To Evaluate Quality of Care: AHCPR 1995:No. 95-0046.
18. Tierney WM, Overhage JM, Takesue BY, et al. Computerizing guidelines to improve care and patient outcomes: the example of heart failure. *JAMIA* 1995;2(5):316-22.
19. Patel VL, Arocha JF, Diermeier M, How J, Mottur-Pilson C. Cognitive psychological studies of representation and use of clinical practice guidelines. *International Journal of Medical Informatics* 2001;63(3):147-168.