

# **Analysis of the Process of Encoding Guidelines:**

## **An Evaluation of GLIF3**

Vimla L. Patel, Ph.D., D.Sc.<sup>1, 2</sup>

Timothy Branch, B.Sc.<sup>1</sup>

Dongwen Wang, M.Phil.<sup>2</sup>

Mor Peleg, Ph.D.<sup>3</sup>

Aziz Boxwala, M.B.B.S., Ph.D.<sup>4</sup>

<sup>1</sup>Centre for Medical Education, McGill University, Montreal, Canada

<sup>2</sup>Department of Medical Informatics, Columbia University, New York, NY

<sup>3</sup>Stanford Medical Informatics, Stanford University School of Medicine, Stanford, CA

<sup>4</sup>Decision Systems Group, Harvard Medical School, Boston, MA

All correspondence may be sent to:

Vimla L. Patel (patel@dmi.columbia.edu)

Department of Medical Informatics

College of Physicians and Surgeons, Columbia Presbyterian Medical Centre

Columbia University

622 West 168<sup>th</sup> Street

New York, NY

10032-372

Tel/Fax: (212)305-5643/(212) 305-3302

**ABSTRACT:** We evaluated the use of an updated version of the Guideline Interchange Format (GLIF), GLIF3, in the translation of clinical guidelines into an electronically encoded form. The goal of GLIF3 is to enable the effective sharing of such guidelines among various clinical institutions and settings. The newer version of GLIF was developed in response to limitations in the expressiveness of the earlier version, GLIF2. We videotaped and analyzed the encoding of two clinical guidelines both into GLIF3 and into GLIF2 by two medical informaticians. GLIF3 was found to be more robust than GLIF2 for representing the content and logical structure of the clinical guidelines studied. Our formative evaluation demonstrated that GLIF3's intended improvements in expressiveness were achieved.

**KEYWORDS:** Knowledge representation, clinical guidelines, evaluation, cognitive analysis, knowledge sharing

## INTRODUCTION

Computer-based approaches to the representation of clinical practice guidelines have received increased attention in response to the recent trends in health care delivery. With the current demand for the standardization of health care, such structured, computer-interpretable guidelines offer the opportunity to improve both the quality and efficiency of treatment provided by physicians through enabling decision support at the point of care[1]. Furthermore, the advent of emerging information technologies in health care, such as computerized patient record systems (CPR) and the Internet, has led to greater possibilities for the dissemination of clinical guidelines to physicians. This widespread dissemination may provide a common standard of care both within a single health care institution and across many different institutions, as well as providing a broad range of applications such as workflow management support, quality assurance evaluation, and simulation for educational purposes[2]. Thus, if these computer-based guidelines are to be used effectively in the standardization of care across institutions, it is critical that they be encoded in a common representation language that may be shared electronically. Such a language must be unambiguous in its representation of medical concepts and procedures if “shareability” is to be achieved. If the goals, intentions, and representations of all users do not match those of the system being used a range of problems and frustrations can result[3].

### GUIDELINE ENCODING AND THE GUIDELINE INTERCHANGE FORMAT (GLIF)

Several models have been developed for the encoding of computer-based guidelines that enable clinical decision support. For example, the *PROforma* model assists patient care through active decision support and workflow management by representing guidelines as constraint satisfaction graphs, where the nodes of the graph represent tasks that can include actions, decisions, enquiries or complex plans[4]. Alternatively, the *PRODIGY* model focuses on patient scenarios that drive decision-making and structures guidelines as a set of choices for the physician which have to be made between alternative actions[5]. The current study focuses on the Guideline Interchange Format (GLIF), a

representational language developed by the InterMed Collaboratory[6]. The InterMed Collaboratory consists of informatics researchers from Columbia University, Harvard University, McGill University, and Stanford University, where the principle mandate for InterMed's participants has been to join in the development of a sharable computer-interpretable guideline representation format (GLIF), and shared infrastructural software, tools, and system components that will facilitate and support the development of diverse, institution-specific applications[7, 8]. Specifically, the GLIF format allows for a formal specification of medical concepts and data, decision and eligibility criteria, and medical actions. This will enable structuring the mapping of terms used in guideline encoding into the codes used by different institutional CPRs, thus facilitating the sharability of encoded guidelines. Furthermore, this set of tools will include an execution engine for GLIF and a guideline server that will store GLIF representations of authoritative clinical guidelines that could be downloaded by healthcare organizations, adapted to their local setting, and implemented in their institutional CPR systems.

The original GLIF specification published in 1998, GLIF version 2.0 (GLIF2), consists of two parts: the GLIF model and the GLIF syntax[6]. The GLIF model consists of a set of classes for guideline entities, attributes of those classes and data types for the attribute values. A particular guideline encoded in GLIF is represented as an object consisting of a name, a list of authors, a characterization of the guideline's intention, a specification of the patient-eligibility criteria, a list of supporting material, and a clinical algorithm represented as a list of all of the steps in the guideline and an indication of the starting step. Specifically, there are four types of guideline steps that can be combined to form a clinical algorithm in GLIF2: action steps, which are used to represent clinical actions, conditional steps, which are used to represent if-then-else statements, and branch and synchronization steps, which are used to represent parallelism. Furthermore, action steps can be nested into sub-guidelines. The GLIF syntax is used to express an encoded guideline as instances of classes of the GLIF model.

While GLIF2 has been found to be adequately expressive in the encoding of guidelines, a comparison among multiple encoders revealed a high level of variability in the encoding processes that were used by each[6]. Specifically, sources of variation included differences in the level of detail represented, differences in the specification of

data elements, and differences in the order in which data elements were specified. Furthermore, variability in encoding has also been found to be dependent on the degree of prior experience and domain-knowledge possessed by the encoder, where representations developed by physicians differed both in content and structure when compared to representations developed by computer scientists[9]. Thus, variation in the encoding of guidelines into GLIF led to qualitative differences between the encoded guidelines and the original guidelines, as well as qualitative differences among the encoders with varying levels of expertise. In order to more fully explain the implications of such variation, and how it may be reduced, it is important to understand the cognitive requirements by which an original guideline may be accurately and efficiently translated into an encoded representation of the guideline's logic.

#### **GUIDELINE EQUIVALENCE AND FLEXIBILITY**

The primary focus of the encoding process is the extent to which GLIF can be used for developing site-specific guidelines that can be accurately and efficiently used by practitioners. Assuming that these criteria are satisfied by the original text-based guideline, the adequacy of the GLIF encoded guideline depends on two factors: the equivalence of the original guideline and the GLIF-encoded guideline, and the flexibility of the GLIF representations for use in institutional settings with varying goals, priorities, cultures, and practical constraints[9].

Given that standardized care can be compromised by differences in the way clinical guidelines are understood, the ability to construct GLIF representations that are equivalent to the original text guideline is crucial. Specifically, there are two perspectives from which the original guideline and the GLIF-encoded guideline may be judged as equivalent: informationally and computationally[10]. If all of the information that can be inferred from the original guideline can also be inferred from the GLIF-encoded guideline, and vice-versa, then the two guidelines are said to be *informationally equivalent*. Thus, both must contain the same concepts and relations to be informationally equivalent. On the other hand, if all inferences that can be drawn from one guideline can also be drawn from the other, using the same cognitive operations, then the two guidelines are said to be

*computationally equivalent*. For example, if a relation is explicitly given in the original guideline but is implicit in the GLIF-encoded guideline then computational equivalence is not satisfied. Therefore, the variation found in the encoding of clinical guidelines in GLIF, such as variations in the level of detail and the order in which data is specified in the GLIF representations, suggests that equivalence between the original guideline and the GLIF-encoded guidelines was not achieved in past efforts. In fact, earlier evaluations of GLIF-encoded guidelines revealed that different recommendations would be given on the basis of the same clinical case, depending on whose GLIF-encoding of the guideline was followed[6]. Therefore, a lack of informational and computational equivalence between original guidelines and their GLIF-encoded counterparts has serious implications for their use in clinical settings, where the standardization of practice may be severely compromised.

In order for GLIF-encoded guidelines to be implemented in different institutions it is crucial that they be encoded in a manner that ensures representational flexibility[11]. Specifically, a flexible representation is generic enough to be applicable at a variety of institutions while allowing for easy adaptation at each specific site. This allows a generic guideline to be used among institutions for different reasons, depending on the specific goals of each institution. However, for GLIF-encoded guidelines to be implemented successfully, the information contained in the guidelines must be interpreted in accordance with its intended meaning. If the shared information encoded in guidelines is represented in a manner such that it is too general or too specific, then the guidelines may not be useful in any particular context. Furthermore, the users of guidelines may interpret them at different levels of abstraction, leading to different representations. For example, research in diagnostic reasoning has shown that the problem solving and decision making approaches applied by experts in a medical specialty contrast considerably with the strategies used by non-experts[12]. Given these differences in reasoning, GLIF-encoded guidelines may be interpreted differently depending on the level of expertise of the user, where the guidelines must be flexible enough to accommodate such differences. Thus, the encoding of guidelines in GLIF involves a fine balance between the flexibility of the guideline so that it may be used for a wide variety of purposes in a wide variety of settings and the inclusion of details necessary for informational and computational equivalence.

Reflecting the attempt to reach a balance between equivalence and flexibility, previous studies have led to the extension and clarification of GLIF, which remains an evolving language. Version 3.0 of GLIF (GLIF3) was introduced in order to overcome several of GLIF2's limitations, including substantive changes to GLIF2's model and syntax[13]. GLIF3 aims to build upon the framework set by GLIF2 and augment it by introducing several new constructs, or program functions, and extending existing GLIF2 constructs to allow for a more formal definition of decision criteria, action specifications and patient data. In addition, GLIF3 is intended to be sufficiently expressive to support the encoding of guidelines that differ in their medical purpose, intended uses, intended users and utilization sites. Thus, the purpose of the current study is to characterize the processes involved in the encoding of guidelines into GLIF2 and GLIF3, where this will allow for an evaluation of the current state of the GLIF representational language. This will involve the observation and recording of encoder-computer interactions in the GLIF-encoding of two guidelines developed by the American College of Physicians – American Society of Internal Medicine (ACP-ASIM): one designed for the pharmacological treatment of depression[14] and another to be used in the screening for thyroid disease[15]. As in our previous research, we will employ a variety of methods from cognitive science which will allow for the identification of characteristic processes in the encoding of guidelines unique to GLIF2 and GLIF3, potential problems related to GLIF2 and GLIF3, and potential problems related to the original guidelines[16]. The current study employs such methods to evaluate the ability of GLIF3 to overcome the previous shortcomings identified with GLIF2 and to achieve its aims of increased formality.

## **METHOD**

In general, an open framework that allows for the convergence of methodological approaches from various research disciplines may provide a useful foundation for evaluation in medical informatics. Through past research, a number of methods have been developed for collecting and analyzing video-based data involving human-computer interactions that support the identification of both cognitive aspects of the system, as well as problems in the interaction between the subject and the system[17]. A practical intent of

using such methods here is to provide a detailed evaluation of the encoding processes used in GLIF2 and GLIF3.

The study focused on two encoders from the InterMed collaboratory who were asked to generate GLIF2 and GLIF3 representations of the depression and thyroid guidelines in their natural, everyday work environment, where both subjects had extensive backgrounds in computer science: a medical informatician at the Harvard site(denoted HMI for Harvard medical informatician) who encoded the guidelines in GLIF2 and a medical informatician at the Stanford site(denoted SMI for Stanford medical informatician) who encoded the guidelines in GLIF3. The instructions given to the subjects were to use the computer system to encode the guidelines into the relevant version of the GLIF format, working from a graphic representation of the guideline and having access to the on-line text of the guideline. In addition, the subjects were instructed to verbalize their thoughts while they were encoding the guidelines.

## **DATA COLLECTION**

The interactions of the subjects during the encoding process were recorded using a methodology similar to that used in our previous research[18]. This approach to data collection involved full video recording of user interactions with the guideline authoring software during the process of encoding guidelines into GLIF. The recording was conducted by remotely logging on to the subjects' application (i.e., the authoring software) at our evaluation site. By doing so, we were able to record the computer screens of the subjects' interaction with the computer system during the task of encoding guidelines into the GLIF language by inputting the resulting screens into a PC-video recorder which then outputs to a VCR. Furthermore, by using a speaker-phone the subject's verbalizations were audio recorded at the evaluation site and merged with the corresponding video of the computer screens. In this way we were able to obtain a complete video recording of the interaction at the remote site in real time. It should be noted that the HMI encoded the guidelines into GLIF2 using different authoring software than the SMI encoding the guidelines into GLIF3, however issues related to the authoring software used to encode guidelines are beyond the scope of the current study. Instead, the underlying processes

inherent in GLIF2 and GLIF3, as related to the original guidelines, will be focused on in order to provide a detailed evaluation of the current state of the GLIF representational language itself.

## **DATA ANALYSIS**

The videotapes of the subjects' interactions as they encoded the guidelines were transcribed and coded based on their encoding-related actions and verbalizations. To facilitate the coding a commercially available video annotation package known as CVideo was used, which allows researchers to document sections of a videotape with textual annotations and codes on a computer[19]. This approach allows for the indexing of the tape to identify exact video sequences where events, such as problems, occur in encoding the guidelines, where the codes that are entered into the text file are time-stamped in relation to the corresponding sequences on the tape.

An important aspect of analyzing video data is the development of a principled coding scheme containing identifiable categories of actions and verbalizations[18]. Based on our knowledge of the models and syntaxes used in GLIF2 and GLIF3, the coding scheme that was used here consisted of identifying and classifying the subjects' encoding-related actions, such as the constructs used to represent the guideline in GLIF, as well as identifying and classifying any problems that were experienced during the encoding of the guidelines. For example, this included encoding-related actions such as the categorization of the guidelines, the definition of data items, the creation of references, and the addition of guideline steps, decision nodes, and links between representations. Furthermore, the identification of problems in encoding the guidelines included problems with the guideline, such as the lack of information, and lack of clarity, where required information is specified in an ambiguous way, as well as GLIF-related problems, such as difficulty in modeling statements, representing temporal sequences, and modeling steps. Thus, the coding scheme that was used focused on providing a detailed description of how each GLIF representation was implemented and developed over time. Such a description may be used to shed light on the encoding processes that were used by the subjects through an examination of the manner in which they approached and completed the encoding task,

and the methods they used to resolve any ambiguities. Issues related to the authoring software, such as problems with the user interface, as well as generic issues in the translation of natural language into a computer-based format were not included in the analysis, given the focus of the current study.

## **RESULTS AND DISCUSSION**

In this section we define and discuss the process of encoding a guideline into GLIF, and the problems encountered during that process.

### **CHARACTERIZATION OF THE GUIDELINE ENCODING PROCESS**

An analysis of the coded transcripts revealed several variations in the processes used in the encoding of the guidelines by the subjects. These variations can be related to differences between the GLIF2 and GLIF3 languages. Specifically, the encoding process demonstrated by the subjects may be interpreted through the consideration of the GLIF constructs used in the representation of the guidelines, the level of detail in the subjects' resulting representations based on the GLIF2 and GLIF3 models, and the representation of temporal sequences and decisions among guideline steps.

#### ***REPRESENTATIONAL CONSTRUCTS USED IN ENCODING***

Table 1 presents the number of steps and actions used by the subjects in encoding the guidelines into GLIF2 and GLIF3, as well as the proportion of the total constructs used in the encoding process that is represented by each individual construct. It should be noted that while the conditional step exists only in GLIF2 model, the case step, choice step, and patient state step, as well as the exception condition, exist only in the GLIF3 model. The conditional step of GLIF2 can be modeled by GLIF3's case step, however the new GLIF3 constructs cannot be modeled by GLIF2. As indicated by Table 1, the encoding of the depression guideline in GLIF2 by the HMI consisted of adding action (11), branch (3), and conditional(6) steps, used to represent clinical actions and decisions in GLIF2, where

the use of these constructs represented 100% of the total encoding process. On the other hand, the encoding of the depression guideline in GLIF3 by the SMI consisted of adding action(8), choice(5), and patient state(1) steps, used to represent clinical actions, decisions and patient states in GLIF3, where the use of these constructs represented only 42% of the total encoding process. In addition, adding data items(15), such as defining guideline concepts, creating references based on information found in the guideline(3), and the use of an exception condition(1) accounted for the remaining 58%. The encoding of the thyroid guideline in GLIF2 by the HMI mainly consisted of adding action(4), branch(5), and conditional(9) steps, where the use of these constructs represented 78% of the total encoding process and adding data items(4) and creating references(1) accounted for the remaining 22%. The encoding of the thyroid guideline in GLIF3 consisted of adding action(9), case(3), choice(1) and patient state(7) steps, where the use of these constructs represented 57% of the total encoding process and adding data items(12) and creating references(3) accounted for the remaining 43%. Thus, the characteristic constructs used in the encoding of the depression guideline varied significantly between the subjects. Indeed, these differences are a reflection of the characteristic encoding processes of GLIF2 and GLIF3. Specifically, guidelines are modeled as a flowchart in GLIF2, where four basic classes of steps, or primitives, are used as nodes for representing recommendations, decisions, and simultaneous paths through the guidelines[20]. Thus, these low-level primitives can be used to encode a variety of guidelines, such as the depression and thyroid guidelines, through the flexibility that they provide. The nature of representing steps in GLIF2 is reflected in the following excerpt taken from the encoding of the thyroid guideline by the HMI while thinking aloud:

There are four steps in GLIF. I'm going to use the conditional step because that's... I can put the logic forth for the eligibility there...Which is really the same as what we were doing earlier...female, greater than fifty...add that here... equals female greater than or equal to fifty years...That's my first step in the guideline. Screening only eligible patients...If that is true then I want to screen for TSH test.. and...that is a, let's see, screening for TSH's....it's an action step, so I'm going to put in an action step...because it will only work if it's true. If a female greater than fifty turns out to be true.

Specifically, in GLIF2 the patient eligibility for the thyroid guideline (i.e., patients who are women over the age of fifty years) is represented first as a conditional step. This step is then followed by a screen sensitive TSH test which is represented as an action step that applies to eligible patients. Nonetheless, the flexibility associated with representing guidelines in these primitives has also been found to result in an encoding process that is subject to variability in the representation of steps [6].

-----  
Insert Table 1 here  
-----

In contrast to GLIF2, GLIF3 enables the encoding of guidelines at three separate levels: a conceptual level, a computable level and an implementable level [13]. At the conceptual level guidelines are represented as flowcharts that can be used for browsing. At the computable level guidelines are specified formally, which allows them to be verified for logical consistency and completeness. At the implementable level guidelines are appropriate for incorporation into a particular institution's information system, after mappings of guideline terms to the CPR codes and methods are done. Furthermore, GLIF3 includes a wider range of constructs for representing guidelines not available in GLIF2, such as the patient state step, in which the patient's condition may be summarized, automatic decisions versus user choices based on rules for and against each decision option, and exception conditions. For example, in the following excerpt from the think-aloud protocol the SMI uses an exception condition to model the occurrence of adverse effects:

So what I would do is I would have an exception to this entire guideline and then I would, uh, so during the entire guideline no matter where you are you would watch for that exception and if it happens then I would direct, if you want to link back to the guideline... Then I would say this was linking back with this step...because I think that as soon as you go off the guideline I would like to begin at the beginning of it. Okay, you gave the medication and you have real serious effects and you still want to see how it affected the depression. Maybe it changes...before you had mild and now you have severe. That may be one of the changes so I would want to come back to the beginning of the guideline, then trace forwards and move along...I mean, the physician, I think, should examine if something happened to the degree of depression and then decide if he could order non-traditional medicine or the traditional medicine.

Thus, GLIF3 allows for a more fully defined encoding of guideline logic than GLIF2, as demonstrated by the SMI's encoding of the depression guideline. Rather than consisting mainly of the addition of steps in the flowchart, the encoding process used by the SMI in GLIF3 also focused on adding information and creating references that more fully defines and supports the concepts found in the guidelines.

Further differences were found with regard to the representation constructs used for the encoding of "if-then-else statements". Specifically, these statements are represented in GLIF2 through the use of conditional steps. For example, the HMI used conditional steps to represent such points in the depression guideline as severity of depression, evaluation for clinical improvement, therapeutic level of drug, and drug class. In contrast to this method of representing if-then-else statements, GLIF3 uses several types of decision steps organized as a decision hierarchy to encode conditional traversals of the guideline rather than the conditional step. This decision hierarchy distinguishes between *case steps* and *choice steps*, where case steps are used to encode decision points that are automated and contain any number of alternatives (similar to the conditional steps of GLIF2), and choice steps are used to encode decision points that require an external agent, such as a physician or other healthcare provider, where arguments are given in favor and against each decision option. For example, in the following excerpt the SMI uses a choice step in representing the evaluation of therapeutic level, which must be done by the physician, in the encoding of the depression guideline:

Okay. So I have this decision node, this decision link, so let me open this. If there's no clinical improvement...then assess dose of serum-level of TCA. And then you ask, you have to make a decision if you should, uh, if you adjust their therapeutic level or not.

Therefore, the subjects were found to use different constructs in the representation of if-then-else statements in the guideline, where these differences are related to the constructs used in the GLIF2 and GLIF3 models.

### *LEVEL OF REPRESENTATIONAL DETAIL IN THE ENCODED GUIDELINES*

In analyzing the encoding process the SMI was found to encode a greater level of detail in representing the guidelines than the HMI. Thus, the encoded guidelines that were produced by the subjects, as well as the associated processes of encoding, were found to differ with respect to the level of detail contained in the representations. These differences may be related to the different granularities used in the GLIF2 and GLIF3 models. Specifically, GLIF3 uses a formal syntax for expressing logical criteria and medical concepts and data, while GLIF2 uses natural language for this purpose. Indeed, GLIF2 does not specify how to structure important attributes of guideline steps, such as patient data and action specifications [13]. Also, values of most attributes are expressed using simple plain text strings, making them unusable for drawing automatic inferences. In contrast, GLIF3 often requires expressing representations formally rather than just using plain text. For example, when defining the eligibility criteria for the depression guideline, the SMI represents “adult” as a quantity (the HMI defines it using plain text):

What I want to say is that there is a data item I want to call date of birth, so what we do is that we are saying “now” minus date of birth. That is the name of the data item...the date of birth is greater or equal to eighteen years.

This requirement for formality results in a greater level of detail when defining concepts in GLIF3. Furthermore, certain statements that cannot be defined in GLIF2 may be modeled using the formal syntax in GLIF3. For example, in the following excerpt the HMI notes that GLIF3 is able to include the level of evidence for a recommendation:

It's really difficult to do that much more. I mean the other thing I could do is add some kind of annotation which is at the level of evidence. GLIF3 for me can do that, I could do it as a didactic. I could say level A evidence, recommendation as evidence.

In contrast, GLIF2 is not able to include this detail. Therefore, the ability to indicate the level of recommendation in GLIF3 is an improvement over the GLIF2 language, where the result is a representation that contains a greater level of detail.

#### ***AMBIGUITY IN THE REPRESENTATION OF TEMPORAL SEQUENCES AND DECISION POINTS***

The representation of temporal sequences and decision points by the subjects in encoding the guidelines was also analyzed. For example, the HMI used branch steps to represent both temporal sequences and decisions in encoding the guidelines, while the SMI did not use any branch steps. Instead, the SMI used case and choice steps to represent key choices in a sequence of events. Thus, differences in the representation of temporal sequences and decision points were also found in the encoding of the guidelines, where this may be related to the way that time (control-flow) is modeled in GLIF2 and GLIF3. In GLIF2 branch steps can be used both for representing the concurrent execution of multiple actions and for making a selection among a set of alternatives. However, this allows for an overlapping of functionality, where decision points among a set of alternatives can be represented by both conditional and branch steps. In contrast, GLIF3 avoids such overlaps by modifying the branch step, where it may no longer be used to represent a decision point in the guideline. Instead, decisions are modeled using case and choice steps, as described

earlier. For example, when modeling a change in drug class in the depression guideline the HMI used a branch step to select one of the remaining drug classes:

If there is no clinical improvement, then we need to go down another pathway. Then I am going to follow this up with two steps to review side effect as an action and then after that there's a branch step which is sort of selecting...I'm just going to select one of the following four drugs.

However, the SMI, modeling the same change in drug class, represented the sequence as a decision node involving a choice step:

So what it says is in case the answer here is "yes" to switch the drug or drug class, which means what you do is you go directly to this step. So I need I a decision destination...There's SSRI's or SNRI's or 5-HT2 antagonists...because it says, you know, after this you are already at the maximum dose you will not give TCA again. So you will go back to decide what drug to give now.

In order to avoid ambiguity, the change in drug class is modeled differently in GLIF3, where this difference is related to the way that temporal sequences and decisions are modeled in GLIF2 and GLIF3. Specifically, the use of case and choice steps in GLIF3 avoids the overlap in functionality of the branch step used in GLIF2 and, thus, reduces the ambiguity found in the GLIF2 model.

### **CHARACTERIZATION OF THE PROBLEMS DURING ENCODING**

The results include an analysis of the type and frequency of problems encountered in the encoding of guidelines. The problems identified from the coded transcripts indicate that they can be broken down into two basic categories: problems related to the original guideline and problems related to representing information in GLIF2 and GLIF3. In order to summarize this data, Table 2 presents the number of problems related to the original guidelines encountered in the encoding sessions with the HMI and SMI.

-----  
Insert Table 2 here  
-----

During the encoding of the thyroid guideline the HMI experienced problems due to a lack of information(6) and a lack of clarity(1), where information that is required is specified using vague or ambiguous statements in the original guideline. During the encoding of the depression guideline the HMI experienced one problem due to a lack of clarity. Thus, the problems related to the original guidelines that the HMI experienced were mostly due to a lack of information(6) in the guidelines which did not allow information to be specified in as much detail as wished. For example, this included the need for knowing exact values from the materials for encoding normal ranges for FT4 tests in the thyroid guideline. As seen in the following excerpt, the HMI must omit certain details in the encoding process as a result of this lack of information:

And I'm going to look up the definition of what the normal FT4 test is. And it is not given in the supporting documentation...I can't find it. I don't know what the normal FT4 test should be so I'm just going to leave the specification down here, where you can see my cursor, blank.

This leads to an incomplete representation of the guidelines, where this lack of information creates ambiguity in the encoded guideline. Furthermore, the SMI also experienced this problem while encoding the thyroid guideline. This lack of information is also demonstrated in the following excerpt where a need for further information about treatment options is stated:

Now we're down to if there are signs or symptoms of hyperthyroidism then we have to consider treatment. And I can't specify the action in more detail because the guideline, does not specify what the treatment should be so we'll leave it at that... So we'll go to the branch that says free thyroxine is elevated. That's on the free thyroxine test was abnormal. With this situation we're just going to have to type in the diagnosis in here, into a branch...and then...a consider treatment box.

Therefore, the problems related to the guidelines that the HMI encountered during the encoding process were mainly a result of information not present in the guideline that was sought to be included in the GLIF2 representation.

While the HMI only encountered one problem that was related to the depression guideline during the encoding process, the SMI referred to problems related to the depression guideline eighteen times, where these problems were due to a lack of information(11) and a lack of clarity(7). During the encoding of the thyroid guideline the SMI experienced six problems, where these problems were also due to a lack of information(4) and a lack of clarity(2). Specifically, the problems that the SMI encountered during the encoding of both guidelines may be related to the more granular nature of GLIF3, where this led to a higher number of problems related to the original guidelines than the HMI experienced. The greater level of detail that is required in the encoding of guidelines in GLIF3 led to the exposure of many additional instances of information missing from the original guidelines and vague statements that were not problematic for the HMI, given the less detailed syntax that is used in GLIF2. For example, in the following excerpt the SMI encountered a problem due to a lack of clarity in the depression guideline related to the formal scale used to define moderate depression:

They say mild is a score set between 7 and 17 on the Hamilton Depression Rating Scale and...now I've noticed another bad thing about it. So they say if it's mild to moderate then you go this way. If it's moderate to severe you go that way. So what does it mean if it's moderate? Where do you go? So they're really unclear about that.

Thus, the lack of clarity related to the severity scale is exposed due to the formal syntax of GLIF3, which requires that the severity of depression be represented formally. In contrast, in the encoding of severity in GLIF2 this ambiguity in the guideline would be hard to discover, given that severity would be represented simply as a plain text string. Thus, additional information missing from guidelines and ambiguities may be discovered as a result of the greater level of detail required in GLIF3.

The number of problems related to the representation of information in GLIF2 and GLIF3 that occurred in the encoding of the guidelines can be seen in Table 3. The problems experienced representing the depression guideline in GLIF2 included problems modeling steps as found in the original guideline(1), modeling informational statements(3) and modeling temporal sequences(2). The problems experienced representing the thyroid guideline in GLIF2 included problems modeling a statement(1) and modeling a temporal sequence(1). For example, in the following excerpt the HMI has a problem modeling a step in the original guideline and is unable to decide whether ordering a free thyroxine test based on TSH-level should be represented as an action step or a conditional step. This problem was further complicated by the awkward way in which action and decision steps were represented in the original algorithm, where this method of representation is incongruent with the way guidelines are encoded in GLIF2. Specifically, the original guideline did not differentiate between the action of ordering a test and the decision made based on the results of this test. Instead, both the action and the decision are represented in a single step. As a result of this ambiguity the HMI is forced to make modifications to the algorithm:

And when TSH is undetectable then what we're going to do is...free thyroxine test.

And,...I'm running into a little problem because I have to decide whether it's an action step or conditional step in ordering a free thyroxine test. Conditional is when making a decision based on the results of free thyroxine test. This particular algorithm has that confused according to GLIF's specifications. I'm going to make this a...conditional step so I'm going to modify the algorithm to a slightly different way of putting it.

Thus, the nature of representing information in GLIF2, as related to congruency with the way information is represented in original guidelines, may lead to certain modifications in the guideline, where the encoded guideline will differ qualitatively from the original guideline.

-----  
Insert Table 3 here  
-----

The HMI also encountered problems representing statements in GLIF2 while encoding both the thyroid and depression guidelines. Similar to the problem above, in the following excerpt the HMI is unable to represent a reevaluation of clinical improvement in the treatment of depression in one step, as it is in the guideline, and consequently he is forced to stray from the flowchart:

I have to do a slight deviation from what's in the flow chart right now. They reevaluate in clinical improvement as one step. I have to do it as two steps here. I have to do a reevaluation followed by something as whether there was an improvement...Because re-evaluation is an action whether there's clinical improvement as a deficient, I need to model different types of steps. You can see that now, two sequential step, first one, reevaluate patient or not, after that you will check whether the clinical improvement, based on your evaluation

Once again, the nature of the GLIF2 language and its lack of congruency with the representation of information in the original algorithm make it difficult for the HMI to encode the guideline as indicated. Specifically, GLIF2 requires the user to distinguish between actions and decisions, where a test is ordered prior to a decision being made

based on the test result. The original algorithm does not display such a decoupling between actions and decisions. Thus, the resulting output that is produced will be qualitatively different from the original guideline, where these differences may lead to alternative representations of the guideline. Nonetheless, this decoupling between actions and decisions avoids the omission of information and possible confusion found in the original algorithm.

The problems related to the representation of information experienced while encoding the depression guideline in GLIF3 included problems modeling statements(4) and a problem modeling a temporal sequence(1). The SMI did not experience any problems related to representing information in GLIF3 during the encoding of the thyroid guideline. While the SMI did encounter problems representing statements in the GLIF3 language, the nature of these problems appears to differ from those of the HMI. Specifically, the HMI's problems were related to variations from the guideline in the number and types of steps required to represent certain statements in GLIF2. In contrast, the SMI's problems were related to the formal syntax defined in GLIF3. For example, in the following excerpt a problem arises when there is a lack of an appropriate concept in the Unified Medical Language System(UMLS), for defining "depressed mood most of the day, nearly every day", where concepts from UMLS are used in the specification of patient data inGLIF3[13]:

So the first one is "depressed mood most of the day nearly every day"...so either there will be a code for "depressed mood most of the day", but probably there is not...I don't think anyone would fill in, uh, the direct period would not be filled in, nobody would fill in rest from five to seven p.m. So someone has to record that the comment followed these...But what we would like to see is this chart gets a code in our ULMS or we can define the code ourselves, then you can use this code...so really I should look for a code that means this very specific term. I don't think there will be one at this moment. The thing that I can do is to define this new code. They have a variable code like I showed you before, I can go and put a UMLS code if I have one I will have to define this "depressed mood almost every day" as a data item literal. I don't want to do all the specific's because it's very long. I will have to put in this code and to define some local I.D. for it.

Here GLIF3 requires that “depressed mood most of the day, nearly every day” be defined formally, where GLIF2 does not have this requirement given that plain text definitions are used instead. Consequently, the SMI must create a new concept to define it, where concepts that are not part of a standard vocabulary can be defined. However, it takes a very long time to encode it formally based on UMLS terms. Furthermore, the SMI goes on to suggest an improvement to the nature of the GLIF3 language with respect to defining new concepts, stating:

And the other thing to do not encode any of these just to put it as a comment so when I have this name acute major depression maybe the thing we need to add except for name is straight definition. So let me write this down it will be useful for GLIF...I am just writing myself a comment that I would like to be able to put a sentence in English that I would explain what acute major depression is. Right now I can only write a formal string.

Thus, the current GLIF3 language led to difficulties in the representation of the depression guideline, where the lack of a textual definition field was identified as an issue to be considered. Nonetheless, the formal syntax used in GLIF3 affords for a much more detailed and unambiguous encoding of the guidelines than GLIF2.

Finally, both subjects had problems representing temporal statements in the GLIF2(3) and GLIF3(1) languages. These problems occurred for statements involving time intervals, as well as statements involving a specific, recorded time. For example, when trying to represent a condition for eligibility that involved a time interval, the HMI was unable to specify the statement easily, saying:

In here to sort of to put it down some syntax to say patient has major depression for the last two weeks and decrease the appetite and insomnia /hypersomnia...some of those can be said more easily, something like recurrent thoughts of death...I really don't have the syntax for specifying that...So I can't model that any more...in the GLIF2, the logic statements.

Thus, it was difficult to represent the statement given that GLIF2 uses plain text to describe such criteria. On the other hand, the formal syntax of GLIF3 makes such statements possible to encode. In another instance, the SMI experienced difficulty representing the onset of adverse effects due to treatment, as seen in the following excerpt:

Adverse effect is serious. I would say something like this “adverse effect would be,” ...we’ll say it’s common. When it started... Yeah, I’m just thinking now how should I say it. So there are two things that I want to say. I want to say that this is an event that now took place and when I’m thinking about the time stamp it’s basically...the event is related to the time stamp not the time interval...I’m not thinking it’s a critical time, it’s the time that this was recorded...so I might say that this is related to recording time or to say that the critical time...both the starting point and the end point may actually be the same time point. I don’t want to relate it to the recording time, but I really want to relate it to the critical time because this is the time it occurred. And I would say something about it’s severity. I would say something like about the severity of the adverse effects.... Severity is greater than, let’s say on a scale from, uh, 1 to 10...I need a physician to say but it’s probably let’s say greater than seven. Now I have to say that this thing is current. Oh, actually you know I don’t need to say it’s current because what triggers this to happen is the event when an adverse effect is recorded. Actually all I need to do is look at the latest adverse effects and check for severity. So all I need to say is “latest adverse effect that occurred is greater than some number” and I take seven to be reasonable.

This difficulty can be related to the specificity required by the GLIF3 data model, where insufficient information in the guideline may make it difficult to represent certain temporal expressions. Therefore, both subjects had problems representing statements that contained temporal expressions that were related to the constraints of the syntax used in GLIF2 and GLIF3, including difficulties encoding time intervals and time stamps.

## CONCLUSIONS

By investigating the encoding processes that underlie the development of clinical practice guidelines in GLIF2 and GLIF3, we have sought to provide a detailed evaluation of the current version of the GLIF representational language. Based on the resulting encoded guidelines that were produced by the subjects, the use of GLIF3 appears to offer several improvements over GLIF2 in the ability to encode guidelines accurately and efficiently.

The principal differences that were identified in the encoding processes demonstrated by the subjects were related to the formality associated with the encoding of guidelines in GLIF2 and GLIF3. First, the subjects were found to encode the guidelines using different representational constructs based on the way guidelines are modeled in GLIF2 and GLIF3. Whereas guidelines are modeled as a flowchart using only four basic classes of steps in GLIF2, the encoding of guidelines in GLIF3 is done at multiple levels, thus focusing on the addition of information as well as the addition of basic steps. Furthermore, several new constructs were used in the encoding of guidelines in GLIF3, such as the case, choice and patient state steps, reflecting the augmentation of GLIF3 over the earlier version. Second, the representations of the guidelines encoded in GLIF3 were found to contain a greater level of detail than those found in the guidelines encoded in GLIF2. This is a result of the model used in GLIF3, which requires information to be encoded formally, thereby reducing ambiguity in the encoded guidelines. Indeed, the GLIF3 syntax was noted to have increased capabilities for representing details over GLIF2, such as the ability to include the level of evidence associated with recommendations. In contrast, GLIF2 models data using plain text strings which limits the amount of detail that is contained in the resulting representations, as well as the automatic inferences that can be drawn from them. Furthermore, GLIF2 does not specify how to structure many important attributes of guideline steps, such as patient data. Finally, the representation of temporal sequences and decision points in the guidelines also differed between the subjects, where the GLIF3 model avoids ambiguities found in the GLIF2 model. Specifically, the branch and conditional steps found in the GLIF2 model can both

be used to model decisions, while the branch step may also be used to represent temporal sequences. In order to avoid this overlap in functionality, GLIF3 uses several types of decision steps, which are organized into a decision hierarchy to represent decision points in a guideline. In addition, GLIF3 uses a modified branch step such that it may no longer be used to represent a decision point in a guideline and is limited to temporal representations. This clearly reduces the potential for ambiguities in representing such statements in GLIF2.

The present analysis of the encoding of the depression and thyroid guidelines also revealed difficulties in the encoding process that may be related to differences between the GLIF2 and GLIF3 languages. The problems that occurred during the encoding in GLIF2 appear to be a result of its lack of formality, where the GLIF2 model uses only four basic classes of steps and the syntax used to encode information in GLIF2 consists only of plain text strings. Thus, the accurate and efficient representation of statements is severely limited, where this lack of formality often led to ambiguity in the encoded guidelines. The possibilities for misinterpretation of the GLIF2-encoded guidelines in clinical settings would clearly be increased as a result of this ambiguity. Conversely, difficulties during the encoding in GLIF3 were mainly related to the formality of the language, which requires a greater level of detail in the encoding process. Specifically, missing or ambiguous information in the original guidelines led to difficulties in encoding when the GLIF3 model required such information. Thus, the formality of the GLIF3 syntax may actually expose deficiencies with the original guidelines that would not have been problematic in GLIF2, given the greater level of representational detail in GLIF3.

In conclusion, GLIF3's intended improvements in formality and expressiveness were achieved, including an improved ability for the accurate and efficient translation of clinical practice guidelines into a shareable electronic format. The increased formality of the GLIF3 model and syntax leads to a guideline-encoding process which contains both a greater level of detail and less ambiguity than that in previous versions. The result of such improvements is an encoded guideline which is better equipped to aid practitioners as they make decisions in a variety of clinical settings.

## ACKNOWLEDGEMENTS

This work was supported in part by National Library of Medicine Grant LM06594, awarded in cooperation with the Department of the Army and the Agency for Healthcare Research and Quality, and in part by grant MRC-MA 13439 from the Medical Research Council of Canada awarded to Vimla Patel. We thank Andre Kushniruk and Jose Arocha for their assistance in the collection and analysis of data. This paper has benefited from the critical comments of Edward H. Shortliffe.

## REFERENCES

1. Lobach D, Hammond W. Computerized decision support based on a clinical practice guideline improves compliance with care standards. *American Journal of Medicine* 1997;102:89-98.
2. Boxwala A, Greenes R, Deibel S. Architecture for a multipurpose guideline execution engine. In: *Proceedings of the AMIA Annual Fall Symposium*; 1999; p. 701-705.
3. Patel V, Kushniruk A. Understanding, navigating and communicating knowledge: Issues and challenges. *Methods of Information in Medicine* 1998;37:460-470.
4. Fox J, Johns N, Rahmzadeh A. Disseminating medical knowledge: The PROforma approach. *Artificial Intelligence in Medicine* 1998;14:157-181.
5. Purves I, Sugden B, Booth N, Sowerby M. The PRODIGY project - The interactive development of the release one model. In: *Proceedings of the AMIA Annual Fall Symposium*; 1999; p. 359-363.
6. Ohno-Machado L, Gennari J, Murphy S, Jain N, Tu S, Oliver D, et al. The guideline interchange format: A model for representing guidelines. *Journal of the American Medical Informatics Association* 1998;5(4):357-372.
7. Shortliffe EH, Patel VL, Cimino JJ, Barnett G, Greenes RA. A study of collaboration among medical informatics research laboratories. *Artificial Intelligence in Medicine* 1998;12:97-123.

8. Shortliffe E, Barnett G, Cimino J, Greenes R, Huff S, Patel V. Collaborative medical informatics research using the Internet and the World Wide Web. In: Proceedings of the AMIA Annual Fall Symposium; 1996; p. 125-129.
9. Patel V, Allen V, Arocha J, Shortliffe E. Representing clinical guidelines in GLIF: Individual and collaborative expertise. Journal of the American Medical Informatics Association 1998;5(5):467-483.
10. Larkin J, Simon H. Why a diagram is (sometimes) worth ten thousand words. Cognitive Science 1987;11(1):65-99.
11. Fridsma D, Gennari J, Musen M. Making generic guidelines site-specific. In: Proceedings of the AMIA Annual Fall Symposium; 1996; p. 597-601.
12. Patel V, Arocha J, Kaufman D. Diagnostic reasoning and medical expertise. In: Medin D, editor. The Psychology of Learning and Motivation. New York: Academic Press; 1994. p. 189-252.
13. Peleg M, Boxwala A, Ogunyemi O, Zeng Q, Tu S, Lacson R, et al. GLIF3: The evolution of a guideline representation format. In: Proceedings of the AMIA Annual Fall Symposium; 2000; p. 646-649.
14. Snow V, Lascher S, Mottur-Pilson C. Pharmacological treatment of acute major depression and dysthymia. Annals of Internal Medicine 2000;132(9):738-742.
15. Helfand M, Redfern C. Screening for thyroid disease: An update. Annals of Internal Medicine 1998;129:144-158.

16. Patel V, Kushniruk A, Yang S, Yale J. Impact of a computer-based patient record system on data collection, knowledge organization, and reasoning. *Journal of the American Medical Informatics Association* 2000;7(6):569-585.
17. Kushniruk A, Patel V. Cognitive evaluation of decision making processes and assessment of information technology in medicine. *International Journal of Medical Informatics* 1998;51:83-90.
18. Kushniruk A, Kaufman D, Patel V, Levesque Y, Lottin P. Assessment of a computerized patient record system: A cognitive approach to evaluating an emerging medical technology. *MD Computing* 1996;5:406-415.
19. Roschelle J, Sibley J. *CVideo user guide*. San Francisco: Envisionology; 1992.
20. Boxwala A, Mehta P, Peleg M, Lacson R, Ash N, Bury J, et al. Modeling guidelines using domain-level knowledge representation components. In: *Proceedings of the AMIA Annual Fall Symposium*; 2000; p. 645.

**TABLE 1:** Number and type of constructs in GLIF2 and GLIF3 encoded by the HMI and SMI

TYPE OF EVENT	GLIF2		GLIF3	
	(#)	(%)	(#)	(%)
<i>Depression Guideline:</i>				
Action step	11	55.00	8	23.53
Branch step	3	15.00	0	0.00
Case step	-	-	0	0.00
Choice step	-	-	5	14.71
Conditional step	6	30.00	-	-
Create references	0	0.00	3	8.82
Data item	0	0.00	16	47.06
Exception	-	-	1	2.94
Patient state step	-	-	1	2.94
<i>Thyroid Guideline:</i>				
Action step	4	17.39	9	25.71
Branch step	5	21.74	0	0.0
Case step	-	-	3	8.57
Choice step	-	-	1	2.86
Conditional step	9	39.13	-	-
Create references	1	4.35	3	8.57
Data item	4	17.39	12	34.29
Exception	-	-	0	0.00
Patient state step	-	-	7	20.00

**TABLE 2:** Number and type of problems related to the original guidelines during the encoding into GLIF2 and GLIF3

<b>TYPE OF PROBLEM</b>	<b>GLIF2</b>	<b>GLIF3</b>
<i>Depression Guideline:</i>		
Lack of information	0	11
Lack of clarity	1	7
<i>Thyroid Guideline:</i>		
Lack of information	6	4
Lack of clarity	1	2
<b>TOTAL</b>	<b>8</b>	<b>24</b>

**TABLE 3:** Number and type of problems related to GLIF2 and GLIF3 during the encoding of the guidelines

<b>TYPE OF PROBLEM</b>	<b>GLIF2</b>	<b>GLIF3</b>
<i>Depression Guideline:</i>		
Modeling a step	1	0
Modeling a statement	3	4
Temporal sequence	2	1
<i>Thyroid Guideline:</i>		
Modeling a step	0	0
Modeling a statement	1	0
Temporal sequence	1	0
<b>TOTAL</b>	<b>8</b>	<b>5</b>