Advani A, Goldstein M, Shahar Y, Musen MA.
Developing quality indicators and auditing protocols from formal guideline models: knowledge representation and transformations.
AMIA Annu Symp Proc. 2003;:11-5.

Developing Quality Indicators and Auditing Protocols from Formal Guideline Models: Knowledge Representation and Transformations

Aneel Advani, MD,[a,b] Mary Goldstein, MD,[a,b] Yuval Shahar, MD, PhD,[c] Mark A. Musen, MD, PhD[a]
[a] Stanford Medical Informatics, Stanford University School of Medicine, Stanford, California
[b] Department of Medicine, VA Palo Alto Health Care System, Palo Alto, California
[c] Department of Information Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Automated quality assessment of clinician actions and patient outcomes is a central problem in guideline- or standards-based medical care. In this paper we describe a model representation and algorithm for deriving structured quality indicators and auditing protocols from formalized specifications of guidelines used in decision support systems. We apply the model and algorithm to the assessment of physician concordance with a guideline knowledge model for hypertension used in a decision-support system. The properties of our solution include the ability to derive automatically (1) context-specific and (2) case-mix-adjusted quality indicators that (3) can model global or local levels of detail about the guideline (4) parameterized by defining the reliability of each indicator or element of the guideline.

Introduction

Clinical guidelines are increasingly being used as tools to improve the quality of medical care.[1] An important task in quality improvement using computerized guidelines is that of developing quality assessments to measure concordance of physician actions and patient outcomes in relation the guideline. One proposed approach to guideline-based quality assessment, suggested by the Agency for Healthcare Research and Quality (AHRQ) is to (manually) derive quality indicators (both outcome and performance measures) from the specification of the guideline itself.[2] However, currently used methods for quality indicators, such as the National Quality Measures Clearinghouse (NQMC) of the AHRQ,[3] are simply long lists of ratio-based measures that are generated and applied outside of any context of a full clinical guideline. For instance, the predecessor to the NQMC, the CONQUEST System, had about 1,200 measures spread over 50 conditions, giving an average of 24 different quality measures for a condition without any of the 24 being reconciled to the appropriate clinical guideline or guideline elements.[4]

Clearly, giving condition-specific assessments of quality with 24 different answers begs the questions of creating (a) a coherent modeling and reporting structure among these quality indicators preferably in relation to a clinical guideline for the condition, (b) a method of producing a "global assessment" or overall summary measure of quality, and (c) a method for designing an auditing protocol to know which of the measures to sample from the potential set in relation to the full clinical guideline. To answer these three questions, we extend our previous work on the development of guideline-based quality indicators. Our previous work has addressed questions (a) and especially (b) by showing that we can structure discrete ratio-based quality indicators from full clinical guidelines by modeling higher-level *intentions* of the guideline. [5,6] The *intentions* of the guideline allow us to model and evaluate higher-level, more global constraints that encapsulate properties relating to many sub-steps of the guideline processes and the relations between processes and outcomes stipulated by the guideline authors. [7,8]

In this paper, we address the problem of how to derive a global yet structured set of quality indicators when the guideline is represented as a formal specification used to drive an automated decision support system. We contrast this problem to that of developing quality indicators from guideline texts, where an automated approach to deriving the quality indicators would currently be unfeasible. Our work is based on the QUIL (Quality Indicator Language) system[5] for modeling and executing queries for guideline-based quality indicators. Below, we describe how the QUIL system can be applied to the problem of representing and deriving quality indicators from formalized guidelines. We discuss the implementation of the method in the QUIL Modeler component of the QUIL system. We then show how the QUIL representation is the basis for an automated method to derive quality indicators and to design auditing protocols that are (1) context-specific and (2) case-mix-adjusted and that (3) can represent global or local levels of detail about the guideline (4) parameterized by defining the reliability of each indicator or element of the guideline.

Methods

The QUIL system for automated quality assessment scores adherence to hierarchical sets of quality-indicators derived from guideline *plan elements* or higher-level *intentions*. Our method is designed to take guidelines expressed in guideline specification languages such as EON,[9] ASBRU,[10] or GLIF3,[11] and produce a set of related quality indicators as individual nodes in a hierarchical guideline-based *QUIL Structure*. In Figure 1 we present an example based on our current implementation of the QUIL system that starts with a

model of an EON guideline for hypertension used in the ATHENA clinical decision support system.[12]

Quality Indicator Language. The QUIL Modeler component takes guideline elements expressed as frames in the Protégé knowledge-modeling tool for the guideline and produces another Protégé knowledge model expressing the QUIL structure of quality indicators (shown with diagram widget in the Figure). To outline our method, we start first with the target language, QUIL, expanding on our previous exposition.[5] In this paper, we focus our discussion on the semantic features of the QUIL graph structure.

The *QUIL structure* is a directed acyclic graph (DAG) of nodes representing quality indicators, and (directed) arcs or edges representing *elaborations* of the quality indicators into more *context-restricted* or *consequential* nodes. In the QUIL structure, the higher-level indicators in the hierarchy can be considered higher-level *intentions* of the guideline while lower-level indicators may be more specific processes or adjusted outcomes. Individual performance criteria or outcome measures can be embedded as nodes in the guideline-based QUIL structure. QUIL nodes define *dyadic* (or ratio-based) queries consisting of *goal* (numerator) and *enabling* (denominator) constraints, preserving the form of ratio-based population quality indicators. Satisfaction of the goal constraint defines a clinical execution of the medical guideline that satisfies the quality indicator, given the appropriate ***evaluation context*** defined by the enabling constraint. The evaluation context can be used to model the case-mix associated with a quality indicator. Furthermore, the expansion of the quality indicators into lower-level nodes that are associated with more specific evaluation contexts is a way to model case-mix adjustment of the higher-level quality indicators.

For instance, the "Drug Treatment" process node can be considered a high-level intention of the guideline. Although we can still conceptually assign it a utility based on appropriate drug treatment from the disease-specific customizations of care in the child nodes, we cannot measure this complex process *directly*. We can, however, *elaborate* the node into more specific nodes that are directly queryable. These elaborated nodes may (1) measure indicators associated with a narrower patient evaluation context than their parent(s) created by elaborating the parents' enabling constraint. The nodes may also (2) represent consequents of the parent under some elaboration of the parent's goal constraint. Or both (1) and (2) may hold. Thus when we elaborate the "Drug Treatment" node by decomposing the

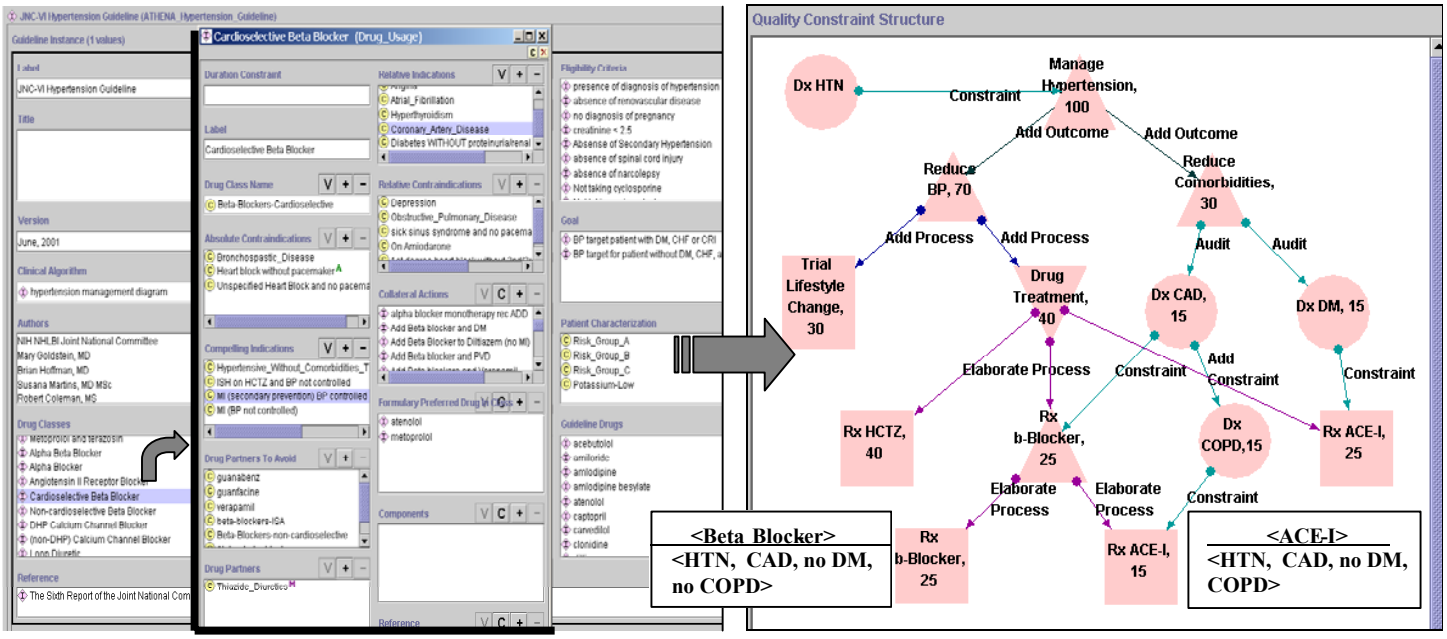

Figure 1. QUIL Structure for the ATHENA Hypertension Guideline. The screenshots show the EON guideline specification and the QUIL structure for the hypertension guideline used in the ATHENA decision support system. The image on the left shows the Protégé knowledge frames used in the EON model of the main guideline instance and the "Beta-Blocker" drug usage class. Our method takes the knowledge roles from the EON guideline and maps them onto quality indicators and elaboration relations in the QUIL structure, shown in the image on the right. Each node in the QUIL structure represents the quality indicator for a specific part of or collection of parts of the EON guideline. The nodes have logical relationships with their children, with downward-pointing triangles as OR nodes, and upward-pointing triangles as AND nodes. The numbers in the nodes refer to the utilities to the patient of satisfying the quality measure. The circular nodes represent *entry constraints* that narrow the evaluation contexts of their child nodes by adding to their enabling constraints. We illustrate the enabling and goal constraints defining the "Rx ACE-I" and "Rx β-Blocker" nodes. Note that all the co-morbidities of the enabling constraints of these two nodes have been inherited from circular nodes above.

guideline plan into condition-specific treatments with drugs in appropriate classes, we are elaborating both the enabling and goal constraints to produce directly measurable process nodes such as "Rx β-Blocker". For example, the lower-level "Rx β-Blocker" quality indicator is satisfied if β-Blockers were prescribed in those patients with diagnoses of hypertension and coronary artery disease, but no diagnosis of diabetes or obstructive pulmonary disease. The enabling constraint of "Drug Treatment" is elaborated by restricting the patient evaluation context of cases measured by the "Rx β-Blocker" node to exclude diabetes, and then further to exclude obstructive pulmonary disease. Similarly, the goal constraint of "Drug Treatment" defined as giving any drug intervention that is anti-hypertensive (thus satisfying that node's parent constraint to "Reduce BP") is elaborated into a disjunction of treatment options using more specific antihypertensive drug classes.

QUIL Structures from Guideline Specifications. Now that we have described our representation for QUIL structures, we outline our method for deriving QUIL structures from a guideline specification. Our algorithm depends on the existence of mapping relations between the elements or *roles* in knowledge base of the guideline specification and the QUIL elements that can be derived from these (see Table 1). For example, the evaluation context for a quality indicator can be derived from the knowledge roles representing the conditions or criteria for executing guideline elements. In the case of our "Drug Treatment" nodes, for instance, we can use the "Compelling or Relative Indications" and "Relative or Absolute Contraindications" to define the evaluation context for each drug class listed in the main guideline instance. The higher "Rx β-Blocker" node has been restricted to cases with coronary artery disease since that is an compelling indication, but the lower, more specific node, is defined on cases that are also without obstructive pulmonary disease since that is a relative contraindication.

The main algorithm then involves using the mapping relations to traverse the knowledge base containing the guideline specification in a context-consistent way so that the QUIL structure derived from the knowledge base is well defined. Our algorithm consists of the following steps:

1. Choose an anchor knowledge frame and parent

QUIL Structure element to map to	Frame-based EON guideline model element to map from	ASBRU or GLIF3 element with similar knowledge roles (for comparison)	Instantiated examples from ATHENA hypertension guideline	Minimum # of potential elements to map from ATHENA KB to QUIL Structure
Enabling constraint to define evaluation context	Drug usage relative and absolute (contra)indications	ASBRU: Filter, Setup conditions. GLIF3: Patient_State_Step, Decision_Step	Hypertension and coronary artery disease as compelling indications for beta-Blocker treatment	750
Goal constraint for process indicators	Drug usage drug class references	ASBRU: Plan body, process intention. GLIF3: Guideline	Using an anti-hypertensive for Drug Treatment in hypertension guideline.	60
Goal constraint for outcome indicators	Guideline goal criteria in guideline instance frames	ASBRU: State intention, Plan effect. GLIF 3: Patient_State_Step	Guideline goals for blood pressure targets	3
Elaborations based on decomposing plans	Slots for plan elements present in frames for plan and action steps	ASBRU: Plan body sub-plans. GLIF3: Sub-guidelines	Drug Treatment decomposed into specific "Drug_Usage" anti-hypertensive sub-classes	77
Elaborations based on materializing outcomes or plans	Slots for plan elements associated with slots for goals or outcomes.	ASBRU: Plan bodies associated with effects or intentions. GLIF3: Action step sequences	"Reduce BP" outcome materialized to "Drug_Treatment" and "Lifestyle_Changes"	5

Table 1. Mapping Relations Between Knowledge Roles in Guideline Specification and QUIL Structure. The table above outlines the mappings between knowledge roles in the EON guideline model and the QUIL Structure. The various knowledge elements required to generate the QUIL structure are listed in the first column. The second and third columns list the corresponding knowledge elements from the EON guideline model and (potentially) the ASBRU and GLIF3 guideline models that could be used to derive a QUIL structure. The next two columns show examples and enumerations of the knowledge elements corresponding to each type of mapping from the ATHENA hypertension guideline knowledge base.

QUIL node to elaborate into child nodes in the QUIL structure. For instance, the main "Guideline Instance" frame in Figure 1 can map to the "Manage HTN" node.

2. Enumerate the potential set of QUIL child nodes that can arise from the mapped frame. If a one-to-many relationship exists between the anchor frame and these QUIL nodes (such as many drug classes to use for "Drug Treatment"), create a disjunctive set of children elaborated from the parent QUIL node. The elaborations are derived by mapping slots in the anchor frame to consequence relations between QUIL nodes. For example, the slot "Drug Classes" can be used to elaborate the "Drug Treatment" node from "Manage HTN". The next level of indirection on "Drug Classes", pulling up the instances of the "Drug Usage" class can then be used to elaborate the "Drug Treatment" further.

3. Enumerate the set of evaluation contexts that apply to a particular node that has been elaborated from the anchor frame in the knowledge base. We can impute these evaluation contexts from the criteria associated with the frame instance. Here we use a mapping from slots in the frame instances to components of the node enabling constraints. So for "Manage HTN" we can put in the "Guideline Instance" slot "Eligibility Criteria" as the enabling constraints. For "Drug Treatment" we can use "Compelling Indications" as enabling constraints as described above.

4. Since the QUIL structure allows multiple parents, we need to make sure the elaborated nodes are inserted correctly in the graph structure. We do this by making sure our insertion is done in such a way that no matter how we have traversed the knowledge base, the set of parents of the QUIL inserted node will always be the same. This requires that we insert the node so that we use the least number of parents that are required to for the node to inherit its largest number of its enabling and goal constraints. That is, we place the node in its *minimally sufficient evaluation context*. The problem is very similar to that of classifying an instance in a description logic, but ours is a simpler form involving only a two scalar lists of properties. We defer the discussion of this algorithm to a later journal paper.

5. After inserting the node, we recursively repeat steps (1)-(4) by simply defining a new anchor frame and QUIL node for each of our newly inserted nodes that have been elaborated. We continue until we either exhaust the knowledge base or satisfy a stopping criterion. The stopping criterion is based on the lower bound on the number of patient cases satisfying the node as required by our auditing protocol. We expand on the lower bound and it's relation to designing auditing protocols below.

Results

Auditing Protocol Design with QUIL. As we traverse the knowledge base and continually add more context-specific node to the QUIL structure, the number patients falling to the bins associated with each of the nodes will continue to decrease. For example, in our hypertension knowledge base used to model an implemented, real-world guideline, there are over 700 criteria that may potentially define the evaluation context for a particular patient case. Clearly, a fully elaborated QUIL model of the hypertension guideline would be too densely connected and deep to be used for an understandable assessment of quality. It has been shown that as the number of patient cases satisfying the evaluation context for a quality indicator decreases, the reliability of the quality indicator also decreases.[6] For this reason, we introduce a stopping criterion for elaborating the guideline that depends on the number of patient cases falling into a leaf node in the generated QUIL structure. Thus, our algorithm with the bin-number stopping criterion may produce a set of guideline-based quality indicators that is optimally reliable for a given patient data set.

We present the results of applying our auditing protocol design method to a dataset comprising records of hypertension care for approximately 1000 patients in the seven clinic divisions of the Palo Alto VA Health Care

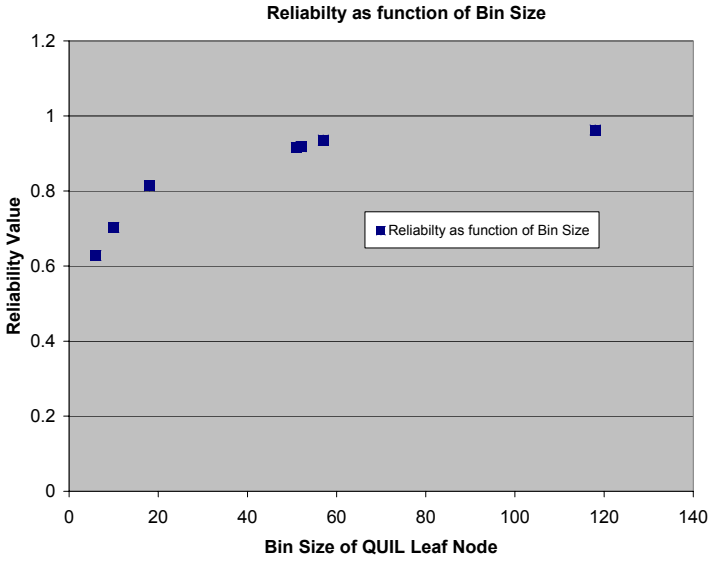

Figure 2. Reliability as a function of bin size. The reliability of the QUIL structure leaf node "Rx β-Blocker" is plotted for the patient populations of the seven primary care clinics at the VA Palo Alto. We can see that the three clinics where less than 20 patients satisfied the node's enabling constraints had reliability scores less than 80%. The QUIL structure would not include that indicator in the hypertension auditing protocol for those three clinics, but would include it for the other four sites.

System, California. We show how the bin-number stopping criterion behaves on the "Rx β-Blocker" leaf node in the QUIL structure generated in the example in Figure 1 for each of the seven clinic divisions. We have used equations based on split-half reliability and the Spearman-Brown prophecy formula, to model the effects of bin size on quality indicator reliability.[13] In Figure 2 we plot the reliability of the QUIL structure leaf node "Rx b-Blocker" is plotted for the patient populations of the seven primary care clinics at the VA Palo Alto. From Figure 2, we see that our lower bound for at least 80% reliable quality indicators places a lower bound of about 20 patient cases on the bin size of a node in the generated QUIL structure. We can see that the three clinics where less than 20 patients satisfied the node's enabling constraints had reliability scores less than 80%. The QUIL structure would not include that indicator in the hypertension auditing protocol for those three clinics, but would include it for the other four sites.

Discussion

In this paper we have described a model representation and algorithm for deriving structured quality indicators and auditing protocols from formalized specifications of guidelines used in decision support systems. We have applied the model and algorithm to the evaluation of physician concordance with recommendations from a guideline-based decision support system for hypertension. Our solution uses the concept of *evaluation contexts* to derive context-specific and case-mix-adjusted quality indicators quality indicators from guideline specifications. Through the use of the hierarchical graph structure of quality indicator nodes and their elaborations, we can model global or local levels of detail about the guideline in our quality assessment. Lastly, we have shown that by parameterizing our algorithm's stopping criterion, we can design auditing protocols that incorporate information about the reliability of using each element of the guideline as a quality indicator.

In addition to quality assessment, our method can be applied to problem of evaluating clinical trials of complex decision support systems, where many elements of the guideline recommendations must be simultaneously evaluated.[14] A major limitation of (and potential for future work on) the method described above is that it does not incorporate the ability to query temporal patterns between nodes and constraints. This limits us from automatically deriving quality indicators that measure clinician adherence to guideline recommendations involving the dynamic execution of the guideline plans. For example, we cannot derive an indicator that tries to measure whether a physician prescribed an ACE-I at a later time in response to a guideline recommendation to add an ACE-I to the patient's regimen.

Acknowledgements. This work was supported by the NIH grants LM07033, LM06806, LM05708 and VA grant HSR&D CPI 99-273. We would also like to thank Susanna Martins, Robert Coleman, Martin O'Connor, and Ravi Shankar for valuable contributions to the ATHENA knowledge base and decision-support system. Views expressed in this paper are those of the authors and do not necessarily reflect those of the VA.

References

[1] Fields MJ, Lohr NK, eds. Institute of Medicine (US). *Guidelines for Clinical Practice: From Development to Use*. Washington: National Academy Press; 1992.

[2] *Using Clinical Practice Guidelines to Evaluate Quality of Care*. Vol 2: Methods. Washington: US Dept. of HHS; 1995. (AHCPR/AHRQ Pub. No. 95-0046.)

[3] National Quality Measures Clearinghouse (NQMC) [Web site]. Rockville (MD): [cited 2003 Mar 10]. Available: http://www.qualitymeasures.ahrq.gov

[4] CONQUEST 2.0 User's Guide. Washington: US Dept of HHS; 1999 (AHCPR/AHRQ Pub. No. 99-0011)

[5] A. Advani, Y. Shahar, & M.A. Musen. Medical quality assessment by scoring adherence to guideline intentions. In: *J Am Med Inform Assoc* 2002; 9(90061):S92-S97.

[6] A. Advani, M. K. Goldstein, & M.A. Musen. A framework for evidence-based quality assessment that unifies guideline-based and performance-indicator approaches. In: *Proc of the 2002 AMIA Fall Symposium*, San Antonio, TX.

[7] Advani A, Lo K, Shahar Y. Intention-Based Critiquing of Guideline-Oriented Medical Care. *Proc. 1998 AMIA Fall Symposium*, Orlando, FL; 1998.

[8] Shahar Y, Miksch S, Johnson P. The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. *Artificial Intelligence in Medicine* 1998; 14:29-51.

[9] Tu, S.W. & M.A. Musen. A Flexible Approach to Guideline Modeling. *Proc 1999 AMIA Fall Symposium*, Washington, D.C.;1999.

[10] S. Miksch, Y. Shahar, and P. Johnson. Asbru: A task-specific, intention-based, and time-oriented language for representing skeletal plans. *Proc. of the 7th Workshop on Knowledge Engineering Methods and Languages* (KEML-97) Milton Keynes, UK; 1997; 9-1 – 9-20.

[11] M. Peleg, A. A. Boxwala, O. Ogunyemi, Q. Zeng, S. W. Tu, E. Bernstam, L. Ohno-Machado, E. H. Shortliffe, & R. A. Greenes. GLIF3: The Evolution of a Guideline Representation Format. *Proc. 2000 AMIA Fall Symposium*, Los Angeles, CA, (20 Suppl):645-649. 2000

[12] Goldstein MK, Hoffman BB, et al. Operationalizing clinical practice guidelines amidst changing evidence: ATHENA, an easily modifiable decision-support system for management of hypertension in primary care. *Proc. 2000 AMIA Fall Symposium*, Los Angeles, CA; 2000.

[13] Hofer TP, Hayward RA, et al. The unreliability of individual physician "report cards" for assessing the costs and quality of a chronic disease. *JAMA* 1999; 281:2098-105.

[14] A.S. Chan, R.W. Coleman, S.B. Martins, et al., Development of a Method to Evaluate a Trial of a Guideline-Based Decision Support System for Hypertension. *Proc. 2003 AMIA Fall Symposium,* Washington, DC; 2003, submitted.