Clinical guidelines are increasingly being used to improve the quality of medical care. An important task in guideline-based quality improvement, the assessment of medical care quality, can be accomplished by retrospectively comparing clinician actions to the guideline recommendations. However, most automated quality assessment measures that are currently used, such as the Health Plan Employer Data and Information Set (HEDIS) quality indicator standards and benchmarks, are still limited to evaluating simple one-step elements of medical care. The use of automated systems for quality improvement is inhibited by the lack of a good quality-assessment methodology that can deal with the complex plans with multiple steps and multiple alternatives that clinicians execute when following medical guidelines.

Most computerized guideline-based quality improvement systems are based on medical guidelines that have been developed for physician education and point-of-care decision support. However, directly applying educational medical guidelines to create quality measures can present problems of external validity. These threats to validity include the lack of information about (1) an explicit specification of the data sources that quality criteria depend on, (2) a method for determining whether criteria are fully or partially complied with, (3) specification of the temporal parameters for observation and evaluation of criterion compliance, (4) a representation of criteria that reflect the intent of the guideline, (5) the specification of acceptable alternatives and exclusions to recommendations of the guideline, (6) the interpretation of the general guideline-based criteria in the context of patient-specific care.

We have created a system, called MedCritic, and a formal specification and query language, Quality Indicator Language (QUIL), that provide a solution to the preceding challenges to creating a quality assessment method from point-of-care medical guidelines. The MedCritic system has been designed to work within the EON or Asgaard architectures, and as a part of the ATHENA clinical decision support system for hypertension. In this paper, we discuss the methodology quality assessment of medical plans used in the MedCritic system. We begin by describing the process for modeling and representing the intentions of guideline authors from guideline documents in QUIL. Then we describe the methodology for assigning importance weights to guideline intentions and scoring the adherence to intentions of guidelines executed by clinicians.

Affiliations of the authors: Stanford Medical Informatics, Stanford University School of Medicine, Stanford, California, USA (AA, YS, MAM), Department of Information Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel (YS).
Modeling Guidelines with QUIL

The QUIL language allows clinicians to specify formally quality indicators that address the six threats to external validity of point-of-care guidelines outlined above. Consider the following guideline fragments from the Sixth Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC VI) representing high-level intentions of the guideline authors:

[A] “The goal of prevention and management of hypertension is to reduce morbidity and mortality. This may be accomplished by achieving and maintaining the SBP below 140 mmHg and DBP below 90 mmHg.”

[B] “The goal may be achieved by lifestyle modification, alone or with pharmacologic treatment. If there are no indications for another type of drug, a diuretic or a beta-blocker should be chosen.”

The intentions of the guideline authors can be target outcome goals (statement [A]) or required actions (statement [B]). More generally, the intentions of the guideline authors represent behavioral constraints on clinician actions and state constraints on patient outcomes. However, we need to annotate these general constraints with information that allows us to successfully create valid quality measures from the intentions. For example, from the first statement, we don’t know when the blood pressure cutoffs should apply. What is the time window of the measurements, six months, one year, or longer? Which classes of patients, such as those with essential hypertension, hypertension with diabetes, undiagnosed essential hypertension, untreated hypertension? These annotations must therefore include knowledge about enabling criteria that identify when it is appropriate to apply the intention in a particular patient case, specification of data sources, time frames for evaluation of compliance, and acceptable alternatives to guideline recommendations.

Single Intention Constraints

Since we view intentions as constraints on clinician behavior, we model intentions as pairs of enabling and goal temporal constraints corresponding to the denominator and numerator in standard rate-based quality indicators called performance measures. Performance measures are defined as ratios that determine the extent to which a clinician’s actions conform to the clinical practice guideline. The enabling constraints allow us to specify for which patients it is appropriate to apply the intention and evaluate the clinician. In QUIL, intentions are defined by the following axioms in Backus-Naur form (BNF) notation to make it easier to write a parser for the language and to create an equivalent XML definition form for document markup:

[1] <intention> ::= intention{IF <enabling_temporal_constraint>, GOAL <goal_temporal_constraint>, <importance_score> }

For example, the intention <beta-blocker-rx> to give a beta-blocker when a patient has uncomplicated essential hypertension and is already on a diuretic, may be specified using the following QUIL definitions.

[2] DEF <beta-blocker-rx> = intention {IF <pt-has-htn>, GOAL <pt-on-beta-blockers>, 70}

where the enabling constraint <pt-has-htn> corresponds to the “denominator” of the performance measure and the goal constraint <pt-on-beta-blockers> refers to the “numerator” of the performance measure. The number 70 is the importance weight of the intention, which defines how adherence to this intention is prioritized over other intentions in the guideline.

The temporal constraints are given by the following BNF definition:

[3] <temporal_constraint> ::= constraint{ TRUE | FALSE | (GENERAL, (<pattern> ((AND | OR) <pattern>)*) ) | (OUTCOME, (GOAL <effect> | <param_prop> | <goal_constraint> ) ) | (ENTRY, (IF <param_prop> | <enabling_constraint> ) ) | (PLAN | ACTION, (GOAL <param_prop> | <goal_constraint> ) ) | (AUDIT, (IF <audit-exp> | <enabling_constraint> ) ) }

Temporal constraints can be general combinations of temporal pattern queries or can be specified as: entry or outcome, for enabling or goal patient state constraints; plan or action, for constraints on clinician behavior (plan constraints are not patient-specific); or audit, for checking for the presence of required data sources and data elements. The use of the audit constraint allows MedCritic to judge whether is appropriate use the intention to evaluate the physician if data sources are not present.

Each of the temporal constraints is built up from complex patterns of temporal events and non-temporal parameter propositions defined by similar QUIL axioms. Finally, the parameter propositions and temporal events are parsed and interpreted based on queries to the temporal patient record database or
EON database mediator. For example, the temporal constraint <pt-on-beta-blockers> may be specified as a temporal pattern of regular daily therapeutic doses of any one drug in the beta-blocker class that is given continuously for the last six months:

[4] DEF <pt-on-beta-blockers> = constraint [PLAN, GOAL pattern \{<drug-is-beta-blocker> AND <drug-for-last-6-months>, MAINTAIN]]


[6] DEF <drug-for-last-6-months> = event [durexp [<drugs_table>, GR_EQ, duration [6, MONTHS]]]

We have shown that the QUIL representation allows us to specify the data sources for each parameter, the temporal aspects of the constraints, as well as the enabling conditions for the intention. Next we describe how QUIL can be used to model the multi-step nature of medical guidelines and the presence of valid alternative plans that can be used to satisfy a guideline’s intentions.

Intention Structures and Revisions

The graphical structure in Figure 1 shows the collection of intentions that we have modeled from the recommendations for hypertension care as described in the JNC VI guideline. The intentions in the figure include the intentions present in the guideline fragments extracted in statements [A] and [B] above. Note that intentions vary from the most general intention of managing hypertension, to the more specific intention to use drug treatment when appropriate, to the most specific intention of prescribing a particular drug such as hydrochlorothiazide (HCTZ) under the requisite enabling constraints. The purpose of creating such an intention structure is to integrate the many behavioral steps and expected patient states of a medical guideline into a set of pre-defined quality indicators that are logically and formally consistent with each other and to the guideline. We do this by organizing intentions into parent-child relations and defining how parents logically relate to their children.

The intention structure of a guideline is thus modeled using a multiparent tree, a directed acyclic graph (DAG), of individual intention nodes. Each of these nodes is given an intention importance weight, and these weights are used in the scoring adherence to the guideline (described below). For a given patient case, each intention is evaluated as either satisfied or unsatisfied. The intention nodes can be related to each other based on the way that satisfying a node’s children will affect satisfaction of the parent node’s intention. This relation is defined by the logical matching operators assigned to the parent node. For example, the node “Reduce BP” in the center of the intention structure is shaped as inverted triangle. The palette to the right shows that this shape denotes an OR intention node. Thus the node will be satisfied if either of its two child intention nodes, drug treatment or lifestyle changes, is satisfied. Similarly, other classes of nodes are AND, ordered (sequential) AND, k-of-n subset, and ordered k-of-n subset.

The nodes are connected by edges that represent intention node revisions of their parent intentions. An intention node revision is an operator on an intention that produces a new child intention. One class of node revision is the static specialization of the parent node. The specification of additional outcome, entry, audit, or plan constraints can be applied to the numerator (goal) or the denominator (enabling) parts of the intention. For example, the node “Reduce BP” inherits the entry constraints from the “Diagnose HTN” node and adds additional outcome goal constraints regarding levels of blood pressure through an outcome specification revision of its parents. Another class of node revisions is the dynamic plan revision. For example, the action specification from the “Drug Treatment” node to the ordered set of drug recommendation actions “HCTZ” and “Beta-Blocker” nodes results from the patient-specific application of the plan interpretation revision to extend the plan. Another revision is the plan substitution revision that replaces an action (b-Blocker therapy) with another action with the same effect (ACE-I therapy) if not contraindicated. We have developed a small ontology of static and dynamic revision classes, expanding upon previous work at Stanford. The new set of revision includes axioms for relaxing component clauses in constraints, allowing time shifts and interval scaling, and changing the sequence of plan components. This ontology can be used to generate additional allowed valid alternatives that the user of the MedCritic system may want to include in the intention structure. The use of these revision operators facilitates the definition of patient-specific quality measures that include multiple alternatives that have explicit and consistent relationships to the intentions of the guideline authors.

Scoring Guideline Adherence

Once the intentions of the guideline authors are modeled and annotated, the MedCritic system uses two
interacting run-time algorithms to produce its quality assessment: the guideline adherence algorithm and the intention recognition algorithm. We limit the main discussion here to the adherence-scoring algorithm.

### Intention Score Normalization

The first phase of the guideline adherence scoring algorithm is to *normalize* the intention weights so that any deviations can be scored consistently. We can see why this adjustment is necessary using an example from the JNC VI guideline. Figure 2 shows a diagrammatic subset of the JNC VI hypertension guideline with the nodes labeled with names U0–U8 and with the importance weights given to each individual intention. When the intention structure is ready to be evaluated against the patient records, the algorithm evaluates the leaf nodes of the intention structure first to see if the current patient record complies with these intentions. However, the higher-level intentions have to be given a truth-value as well. This can be done only by “propagating up” the matches and non-matches from the leaf nodes. The process must take into account the different logical matching operators associated with each parent node. For example, U2, an OR node, should receive credit for its entire score if only one of its child nodes is true, whereas U6, an AND node, should receive less than a full score in this case.

The solution to this problem is to create a new intention structure with each node assigned with a new normalized importance weight (see Figure 2b). The normalization is based on the conceptual abstraction of looking at each intention node with its child (null for leaf nodes) as an element of a *normed linear vector space*, with a metric norm defined on the space that is specific to the type of *logical matching operator* assigned to the parent node. Intuitively, the normalized score assigned to an intention node is the contribution to the total adherence score of the guideline of satisfying the intentions that are descendents of that node.
The main reason for using this conceptualization is that it can be shown (a) that the composition of different norms on a linear vector space is also a well-behaved norm on that vector space, and (b) all bounded linear norms are equivalent in value up to a constant factor related to the number of nodes.11 These results allow us to consistently propagate matches upwords from sub-trees with AND nodes (such as node N6) and from subtrees with OR nodes (such as node N2) to higher-level intentions. We use the special class of p-norms, defined as:

\[ \| x \|_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad 1 \leq p \leq \infty \]  

(7)

where \( x \) is the vector of scores from the children of a parent node and the value \( \| x \|_p \) is the score added to the weight of the parent node. Each logical matching operator is assigned to a different p-norm as follows: (1) for AND and ORDERED_AND nodes, we have \( p=1 \), which is the just sum of the child scores; (2) for OR nodes, we use \( p = \infty \), which reduces to the max of the child scores; (3) for the k-of-n SUBSET nodes, we set \( p = \exp(k/(n-k)) \). Alternatively we can set \( k=n \) for AND nodes and \( k=0 \) for OR nodes where \( p = \exp(k/(n-k)) \) as well. With this method, we can combine the different p-norms for each logical matching operator as we propagate truth-values up in our intention structure and still be confident that we are getting consistent results.

**Adherence-Scoring Algorithm**

In the second phase of the adherence-scoring algorithm, we query the patient data for compliance with the leaf nodes. In the example, we find that the physician: (1) did put the patient on a trial to improve his lifestyle by cutting down on drinking (node N5 matched); and (2) did prescribe HCTZ (node N7 matched) with an ACE-I instead of the suggested Beta-Blocker (node N6 unmatched). We therefore have to take these results into account and propagate the effects of the matches and unmatches on the scores for higher-level intentions (Figure 2c). The procedure for the propagation of scores is just to re-normalize the intention structure using the new truth-values for the leaf nodes. For node N2, since it is an OR node and since one of its children matched, it matches to true as well (i.e. we use the maximum value of its children scores since at least one of the children matched). For node N6, we remember that the 1-norm is just the sum of the child scores, so in this case we subtract the value of the unmatched child N8. Since node N6 didn’t match itself, we also subtract the value of its own original importance weight of 40. So we have N8 (70 - 70 = 0), N6 (180 - 70 - 40 = 70). The other nodes don’t change their truth-values, so we need not recalculate their scores.

Note that if we had accepted ACE-Inhibitor as a valid revision of Beta-Blocker (as some more recent evidence suggests), then we need not have taken the entire value of node N6 out from the score. A new lower penalty could simply have been applied to node N8 for allowing a plan substitution revision to

---

**Figure 2** Example of the guideline-adherence scoring algorithm. The diagram shows the intentions structure for the JNC VI guideline shown in Figure 1. First, the intention structure (A) is modified to create a normalized version (B). Then the leaf node matches are propagated up to top of the structure using the logical matching operators (C).
occur. If we had assigned a penalty factor to plan substitution revisions (such as 50%) then we could have gracefully scored this valid but not strictly adherence situation somewhere in between perfect compliance and total non-compliance.

The use of this scoring algorithm enables us to combine quality measures based guideline outcomes, processes, preconditions, and data constraints into a uniform method of evaluation. Moreover, the importance weight of intentions can be chosen to reflect evidence of their clinical effectiveness in contributing to more general intentions such as reductions in mortality or co-morbidity or in relation to the effectiveness of other alternatives. The judicious choice of intention weights and revision penalties permits the creation of measures consistent with multiple plan choices and congruences between processes and their outcomes.

Discussion

We have outlined a method for quality assessment that seeks to solve some of the problem with using medical guidelines as the basis for quality indicators. We have shown how the QUIL language can incorporate annotated knowledge to validly use point-of-care guidelines for quality assessment. By allowing patient-specific guideline revisions, our method reconciles the inconsistency in using guidelines built for individualized patient care and using criteria based on these guidelines as population-based quality measures. We have built a system, called MedCritic, which takes QUIL models of quality indicators and scores adherence to these indicators. The MedCritic system retains a correspondence between the definition of QUIL intentions and standard ratio-based performance measures. This allows a “graceful degradation” between our method for quality assessment and current standard quality measures defined as simple ratio-based benchmarks.

The second part of the work on our system deals with the problem of recognizing physician intentions. This effort builds on previous experience in quality improvement using medical expert critiquing systems. The HyperCritic and VQ-ATTENDING systems showed the importance of being able to represent and recognize the intentions of physicians executing plans when using medical guidelines as quality assessment tools. The QUIL language allows the user to model revisions to allow valid alternative plans to be generated from the original guideline. The use of repeated revisions to search for matches with physician actions is the basis for the construction of an algorithm that allows for the recognition of physician intentions in addition to modeling the guideline authors’ intentions.

This work was supported by the National Library of Medicine Grants LM07033, LM05708.

Reprinted from the Proceedings of the 2001 AMIA Annual Symposium, with permission.

References