MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press


# Modeling Data and Knowledge in the EON Guideline Architecture

**Samson W. Tu and Mark A. Musen**

*Stanford Medical Informatics, Stanford University School of Medicine, Stanford, California, U.S.A.*

**Abstract**

*Compared to guideline representation formalisms, data and knowledge modeling for clinical guidelines is a relatively neglected area. Yet it has enormous impact on the format and expressiveness of decision criteria that can be written, on the inferences that can be made from patient data, on the ease with which guidelines can be formalized, and on the method of integrating guideline-based decision-support services into implementation sites' information systems. We clarify the respective roles that data and knowledge modeling play in providing patient-specific decision support based on clinical guidelines. We show, in the context of the EON guideline architecture, how we use the Protégé-2000 knowledge-engineering environment to build (1) a patient-data information model, (2) a medical-specialty model, and (3) a guideline model that formalizes the knowledge needed to generate recommendations regarding clinical decisions and actions. We show how the use of such models allows development of alternative decision-criteria languages and allows systematic mapping of the data required for guideline execution from patient data contained in electronic medical record systems.*

***Keywords:***

Clinical Practice Guidelines, Knowledge Representation, Expert Systems, Computer-Assisted Decision Making

## Introduction

In recent years, guidelines and protocols have gained support as vehicles for promulgating best practices in clinical medicine. We see considerable interest in creating sharable computer-interpretable representation of guidelines [1, 2]. For providing guideline-based decision support, it is not sufficient to have a computer-interpretable guideline representation and algorithms for making use of it. As the experience with sharing of Medical Logic Modules (MLM) across multiple institutions indicates [3], a great deal of the incompatibilities among encodings of guidelines lie in the way data and medical concepts are represented and linked to an electronic medical record (EMR) system. A guideline-based decision-support system must reason with patient data, medical concepts, and the knowledge embodied in a clinical guideline. Having a clear model of the knowledge and data and how they can be acquired and maintained are critical for the success of implementing guidelines at points of care.

This paper seeks to clarify the respective roles that patient data, medical-specialty concepts and relations, and computer-interpretable guideline representation play in providing patient-specific decision support. We show, in the context of the EON guideline architecture, how we use the Protégé-2000 knowledge-engineering environment to build (1) a patient-data model that describes the structure of data that can be obtained from external sources and used to describe patient situations, (2) a medical-specialty model consisting of taxonomic hierarchies of medical concepts and their relationships, and (3) a guideline model that formalizes the knowledge needed to generate recommendations regarding clinical decisions and actions. We show that such data and knowledge models crucially affect the type of decision criteria that can be written to describe patient situations and the way patient data can be mapped from clinical information systems.

## Materials and methods

We first review the role that knowledge and data modeling play in the EON architecture. EON is an evolving suite of models and software components designed to create guideline-based applications (Figure 1). It contains a set of middleware servers that perform the computation necessary to support specific tasks in guideline-based patient care. One such server, the Padda guideline execution server [4], takes as inputs formalized clinical guidelines and relevant patient data to generate situation-specific recommendations. A second server, the Tzolkin temporal data mediator [5], extends the traditional relational database server to include capabilities to resolve queries involving complex temporal relationships and to create temporal intervals representing abstractions derived from primitive time-stamped data. A third component, known as WOZ [6], provides explanation services for other components.

All EON application, data mediation, and explanation servers access data models and knowledge bases created and maintained in the Protégé-2000 knowledge-engineering environment [7, 8]. Protégé-2000 has a frame-based knowledge model: all entities in a Protégé knowledge base—instances, classes, slots, facets, and constraints—are frames. Instances represent objects in the domain of interest

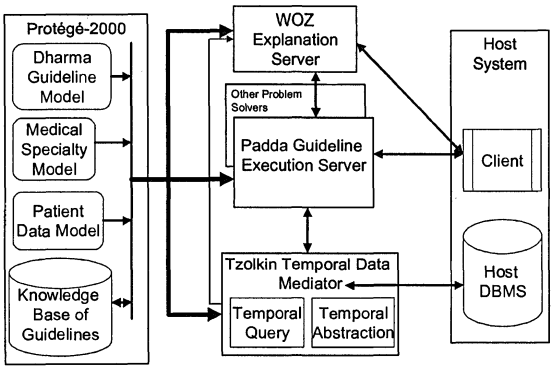


*Figure 1. Architecture of EON guideline applications. All problem-solving, explanation, and data mediating components access formal data models of clinical guidelines, patient data, and medical concepts created in the Protégé-2000 knowledge engineering environment.*

(e.g. a patient). Classes are either named collections of instances or abstract conceptual entities in the domain (e.g. the concept of a drug ingredient). Slots are binary relations describing properties of classes (e.g. the indications of a drug). Facets describe properties of slots (e.g. the data type of a slot's value). Constraints specify additional relationships that must hold among instances in a Protégé knowledge base.

We distinguish the concepts of *information model*, *ontology*, and *vocabulary*. An *information model,* in the sense of the HL7 Reference Information Model (RIM) [9], is a set of metadata specifications that define the structure of data to be stored or manipulated. Like the schema of a relational database, an information model specifies the categories of data and their attributes. An *ontology* is a logical theory that gives an explicit account of the conceptualization of a domain. It contains the formal description of concepts and relationships in a domain of interest. Often these concepts can be organized into taxonomic hierarchies. The ideas of information model and ontologies are closely related. A taxonomic hierarchy of concepts can be constructed according to a particular information model. In that case the categories of concepts and possible relationships in the taxonomic hierarchy are constrained by the information model. An information model itself is a kind of ontology. The domain of interest is the categorization of data that need to be represented. A *vocabulary*, on the other hand, is a collection of lexicons like the International Classification of Diseases (ICD), which may or may not be constructed according to a set of formal principles.

To represent data and knowledge needed for guideline-based decision support, we created (1) an information model of patient data as a set of classes and attributes in Protégé-2000, (2) a medical-specialty model consisting of an information model of the structure of domain concepts and relationship, an ontology of domain concepts organized into taxonomic hierarchies, and (3) a guideline ontology/information model called Dharma that defines the structure of the guideline knowledge required for performing guideline-based application tasks. The Dharma guideline model includes criteria and query languages that use the patient-data model and medical-specialty model to specify relevant patient situations. In the following subsections, we will describe each of the information models and ontologies, their relationship to each other, and how we use these models to link patient data in a medical record system to an EON guideline application.

**Patient Data Information Model**

The patient-data information model defines the classes and attributes of patient data that the rest of the system assumes. We model patient data as either static, time-stamped, or associated with a time interval during which the information represented by the patient data is valid. The EON patient-data model consists a *Patient* class whose instances hold demographic information about specific patients, a *Qualitative_Entry* class that describes qualitative observations about patients, a *Numeric_Entry* class that represent results of quantitative measurements, an *Adverse_Event* class that models adverse reactions to specific substances, a *Condition* class that represent medical conditions that persist over time, and two intervention classes, *Medication* and *Procedure,* that model drugs and other medical procedures that have been recommended, authorized, or used.

The defining characteristic of entities in the patient-data model is that they are assertions about demographic and clinical conditions of specific patients. We do not try to create a data model that replicates everything that an EMR holds, but only those distinctions that are relevant for the purpose of modeling guidelines and protocols.

**Medical-Specialty Model**

Unlike patient data, the concepts we want to model in the medical-specialty model are abstract entities that can be organized into taxonomic hierarchies. The medical-specialty model has an information model specifying the categories of concepts in the medical-specialty model and the attributes of these concepts. In Protégé-2000, the information model for an ontology can be specified as a set of *meta-classes*. Meta-classes are classes whose instances are themselves classes. Attributes of a metaclass are inherited by classes that are instances of that metaclass. These classes can have values assigned to these inherited attributes. For example, the information model for the ontology contains a *Defined_Condition_Metaclass* that has an associated *definition* slot. In the context of hypertension management, *hyonatremia* as a contraindication for the use of thiazide may be defined as a serum sodium measurement that is less than 135 mg/dl. Thus the concept of *hyonatremia*is is defined in terms of the range of values that some patient data can take. If there is an instance of *Numeric_Entry* whose *domain_term* slot is the term *Serum_Sodium* from the medical-specialty model and

whose *value* slot is less than 135, then the system should infer that *hyonatremia* is present for the patient.

**Use of Patient Data and Domain Concepts in the Guideline Ontology**

The design of the Dharma guideline ontology has been reported elsewhere [10]. Here we describe the use of patient data and medical-specialty domain concepts in the guideline model. The primary mechanism through which the guideline ontology interacts with the patient information model and medical-specialty model involves decision criteria written in EON's criterion languages.

Guideline authors and developers can write decision criteria in the Dharma guideline model in one of three criterion languages. We will use examples to illustrate these three possibilities:

1. Presence of *diabetes mellitus* and most recent *serum creatinine* is less than upper limit of normal
2. Presence of an *authorized medication* that is *contraindicated* by some *medical condition*
3. Presence of an episode of *uncontrolled blood pressure* that overlaps with *lisinopril* medication and that started within two weeks after the initiation of lisinopril

The first criterion language consists of a set of object templates that allows a guideline author to encode common but relatively simple criteria by filling in forms generated by Protégé-2000. The first examples can be encoded by filling in three forms: (1) one form (Figure 2) that allows the specification of the presence or absence of patient-data model instances that correspond to some domain terms (e.g. diabetes mellitus) from the medical-specialty model, (2) a similar form that allows description of the comparison of a numeric entry (e.g. serum creatinine value) with a cut-off value queried from the medical-specialty model (e.g. upper limit of normal of serum creatinine), and (3) a form that allows the specification of a Boolean combination of other criteria (e..g, a conjunction of the two statements in Example 1).

Criteria encoded in this object-based language evaluate to *true*, *false*, or *unknown*. In the absence of any serum creatinine result, the criterion that compares the serum creatinine value to its upper limit, for example, will evaluate to *unknown*. For medical problems specified in the medical-specialty model, we make the closed-world assumption where the criteria that checks the presence of these medical conditions evaluate to *false* if there is no record of such medical condition.

A guideline author can use these templates to write most common decision criteria with little training. The criteria that can be written in this form, however, are not sufficiently expressive to encode the second and third examples. Example 2 requires that, for each current medication, we determine its contraindications from the medical-specialty knowledge base, and check to see if there is any patient-data instance that suggests presence of one of

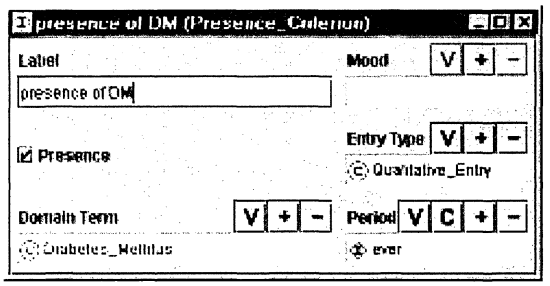


*Figure 2. Protégé-2000 graphical form for specifying the presence of some qualitative entry corresponding to the observation that the patient is diabetic.*

these contraindication. Instead of trying to resolve such complex criteria procedurally, we use Protégé-2000's PAL constraint language to write them (Figure 3). The PAL constraint language implements a subset of first-order predicate logic written in the Knowledge Interchange Format (KIF) syntax [11]. It makes full use of Protégé-2000's frame-based knowledge model. For example, variables can range over instances of Protégé-2000 classes (e.g. in Figure 3, the variable *?current_med* ranges over instances of *Medication* class). Attributes of classes (e.g. *Absolute_Contraindications* in Figure 3) can be used as a binary predicate to check that a constant or the value of a variable (e.g. *?contraindication*) is a value of the slot. An attribute can also be used as a single-arity function (e.g. *domain_term* in Figure 3) that returns the value of the slot for an instance. While we have a structure editor that aids guideline developers in writing these complex logical criteria, we don't expect domain experts not trained in logic to formulate them.

The third example criterion requires the system to make an abstraction (an episode of uncontrolled blood pressure) and to compare the temporal sequence of occurrences (the episode of uncontrolled blood pressure overlaps the use of lisinopril and it started within two weeks after initiation of lisinopril). For this type of criteria, our group has developed the Tzolkin temporal data mediator which is

```
(defrange ?current_med :FRAME Medication)
 . . .
(exists ?current_med
   (exists ?med_class
        (and (subclass-of
               (drug_name ?current_med) ?med_class)
             (exists ?contraindication
               (and (Absolute_Contraindications
                      ?med_class ?contraindication)
                    (exists ?problem
                      (subclass-of
                         (domain_term ?problem)
                          ?contraindication)))))))
```

*Figure 3. Simplified PAL criteria checking existence of contraindicated medication. The KIF syntax uses a prefix notation. The variable ?current_med ranges over instances of the Medication class, ?med_class ranges over drug classes, ?contraindication ranges over medical conditions, and ?problem ranges over instances of patient problems.*

capable of simultaneously making interval-based abstraction from primitive data (e.g. creating episodes of uncontrolled blood pressure) and resolving complex temporal comparisons [5]. Figure 4 shows a Tzolkin query that formalizes the patient condition described in Example 3.

The three criteria languages in the Dharma guideline model are informed by the same patient-data information model and serve complementary requirements. The object-oriented language provides form-based templates that domain experts can use to write the common decision criteria easily. The logic-based PAL language and the Tzolkin temporal-query language add additional expressiveness for writing specific types of complex criteria. The object-oriented templates and the PAL language make use of the medical-specialty taxonomic hierarchies to infer generalization/specialization relationships and to make abstractions based on definitions (e.g. the definition of *hyonatremia*) embedded in the hierarchies. The Tzolkin data mediator creates temporal abstractions using the Resume knowledge-based abstraction system [12].

A common evaluation-result interface unifies the usage of the three criteria languages. For criterion written in each language, the guideline execution engine invokes the appropriate criteria-evaluation engine. Each criteria-evaluation engine returns the result consisting of the criterion being evaluated, the truth value of the criterion as applied to available patient data, and annotations that can be used for explanation purpose.

**Linking to Medical Record System**

For criteria in an encoded guideline to evaluate specific patient situations, available patient data must be mapped to the terms and relations that are used in the guideline. Having a medical-specialty model where concepts and relations used in an encoded guideline model are described means that the mapping can be done systematically across the whole domain model. Unlike Arden Syntax, where the data slot of every MLM has to be modified, data mapping in EON can be done without changing the encoded guideline.

For the application of EON architecture to the Department of Veteran Affair's ATHENA hypertension advisory system [13], we used two data-mapping techniques. The first technique involves the creation of a mapping knowledge base and a set of software modules that transform the schema and terminologies of the data extracted from VA's VISTA system so that the patient data viewed by the Padda guideline execution server are already consistent with the patient-data information model assumed in the EON system [14]. The second technique simply augments the medical-specialty model so that appropriate ICD9 codes are represented as specializations of disease concepts in the model.

```
TEMPORAL SELECT domain_name
VALID INTERSECT(Condition, Medication)
FROM Condition, Medication
WHERE domain_name = "UNCONTROLLED_BP" AND
      drug_name = "lisinopril"
WHEN start(Condition) AFTER start(Medication) AND
     start(Condition) BEFORE
          (start(Medication) + weeks(2))
```

*Figure 4. An example of temporal query that checks for existence of an episode of uncontrolled blood pressure that overlaps with administration of lisinopril but occurring within two weeks of initiating lisinopril.*

## Discussion

We developed a framework for structuring patient information, medical concepts, and guideline knowledge. In this framework, clinical guidelines can be encoded using criteria languages that are informed by the structure of patient data, by a formal medical concept model, and by taxonomic hierarchies of medical terms. The patient-data information model and the taxonomic hierarchies present targets for schema and vocabulary mappings from an implementation site's clinical information system.

In contrast to our approach, the well-known Arden Syntax [1] for modeling medical decisions has a data and knowledge model involve variables that take series of time-stamped data as their values and a procedure-oriented language for expressing criteria for medical decisions. Definitions of variables are left in curly braces for local sites to implement. Our approach models patient data as declarative assertions whose meanings are constrained by concepts and relations in the patient-data information model and the medical-specialty model. Primitive terms in the medical-specialty model still require mapping to local EMR, but much of the semantics previously buried in the curly braces is made explicit in the data and medical-special models.

A recent proposal for integrating data and medical knowledge is the Unified Service Action Model (USAM) [15]. USAM defines an action-oriented information model for patient data, concept definitions, action plans, conditionals, and goals. Through the idea of *action moods* that distinguishes various ways an action can be conceived as having states (such as definitional, factual, possible, intended, or ordered), it uses the same conceptual structure to represent the patient data, domain ontology, and guideline encoding for which we created multiple models in EON. For the ease of integration with EMR systems, it is a good idea to make a patient-data model as close to that of the HL 7 RIM as possible. However, by combining all possible attributes of all possible action states, USAM does not allow a user to easily identify the knowledge roles relevant for alternative uses of the USAM conceptual structures. In the EON framework, by clearly defining the roles of guideline model, medical-specialty model, and patient data model, we allow the application of problem-solving and inference methods appropriate for each type of model.

Much remains to be done in developing the methodology for integrating data and knowledge modeling into guideline-based applications. We need to clarify the relationship between medical-specialty models with vocabulary systems. The maintenance of large multi-axial taxonomic hierarchies and their adaptation to local implementation environment are especially problematic. The information model for patient data need to be standardized so that what's available in electronic medical records and what's required in guideline models can be brought into conformance. Deployment of guideline applications also need to take into account the organizational model and workflow processes at implementation sites. By creating explicit models of patient data, medical specialties, and clinical guidelines as we have done in the EON architecture, we can draw on the research results of ontological engineering that has commanded much attention in research years.

## Acknowledgments

This work has been supported by National Library of Medicine, a grant from FastTrack Systems, Inc., DARPA contract #N66001-94-D6052, and grant LM06594 with support from the Department of the Army, Agency for Healthcare Research and Quality, and the National Library of Medicine.

## References

[1] Hripcsak G, Clayton PD, Pryor TA, Haug P, Wigertz OB, Van der lei J. The Arden Syntax for Medical Logic Modules. In: Proc Annu Symp Comput Appl Med Care. Washington, D.C., 1990:200-204.

[2] Peleg M, Boxwala A, Ogunyemi O, Zeng Q, Tu S, Lacson R, Bernstam E, Ash N, Mork P, Ohno-Machado L, Shortliffe EH, Greenes RA. GLIF3: The Evolution of a Guideline Representation Format. In: Proc AMIA Symp. Los Angeles, 2000:645-649.

[3] Pryor T, Hripcsak G. Sharing MLM's: An Experiment Between Columbia-Presbyterian and LDS Hospital. In: Proc Annu Symp Comput Appl Med Care, 1993:399-403.

[4] Tu SW, Musen MA. From Guideline Modeling to Guideline Execution: Defining Guideline-Based Decision-Support Services. In: Proc AMIA Symp. Los Angeles, USA, 2000:863-867.

[5] Nguyen JH, Shahar Y, Tu SW, Das AK, Musen MA. Integration of Temporal Reasoning and Temporal-Data Maintenance into a Reusable Database Mediator to Answer Abstract, Time-Oriented Queries: The Tzolkin System. Journal of Intelligent Information Systems 1999;13(1/2):121-145.

[6] Shankar RD, Musen MA. Justification of Automated Decision-Making: Medical Explanation or Medical Argument? In: Proc AMIA Symp. Washinton D.C., 1999:pp. 395-399.

[7] Musen MA, Fergerson RW, Grosso WE, Noy NF, Crubezy M, Gennari JH. Component-Based Support for Building Knowledge-Acquisition Systems. In: Conference on Intelligent Information Processing (IIP 2000) of the International Federation for Information Processing World Computer Congress (WCC 2000). Beijing, 2000:18-22.

[8] Noy NF, Fergerson RW, Musen MA. The Knowledge Model of Protege-2000: Combining Interoperability and Flexibility. In: 2th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2000). Juan-les-Pins, France, 2000.

[9] Health Level 7. HL 7 Reference Information Model. http://www.hl7.org/library/data-model/RIM/modelpage_non.htm: 2000.

[10] Tu SW, Musen MA. A Flexible Approach to Guideline Modeling. In: Proc AMIA Symp. Washinton D.C., 1999:420-424.

[11] Knowledge Interchange Format: Draft Proposed American National Standard (dpANS). 1998.

[12] Shahar Y. A Framework for Knowledge-Based Temporal Abstraction. Artificial Intelligence 1997;90(1-2):79-133.

[13] Goldstein MK, Hoffman BB, Coleman RW, Musen MA, Tu SW, Advani A, Shankar R, O'Connor M. Operationalizing Clinical Practice Guidelines While Taking Account of Changing Evidence: ATHENA, an Easily Modifiable Decision-Support System for Management of Hypertension in Primary Care. In: Proc AMIA Symp. Los Angeles, USA, 2000:280-284.

[14] Advani A, Tu S, O'Connor M, Coleman R, Goldstein MK, Musen M. Integrating a Modern Knowledge-Based System Architecture with a Legacy VA Database: The ATHENA and EON Projects at Stanford. In: Proc AMIA Symp. Washington, D.C., 1999:653-657.

[15] Schadow G, Russler DC, Mead CN, McDonald CJ. Integrating Medical Information and Knowledge in the HL7 RIM. In: Proc AMIA Symp, 2000:764-768.

**Address for correspondence**

Samson Tu, M.S.
Stanford Medical Informatics
Medical School Office Building x-259
251 Campus Drive
Stanford, CA 94305-5479
Phone: (650) 725-3391 Fax: (650) 725-7944
Email: tu@smi.stanford.edu
http://www.smi.stanford.edu/people/tu/