Journal of Biomedical Informatics 41 (2008) 889–903

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A quantitative assessment of a methodology for collaborative specification and evaluation of clinical guidelines

Erez Shalom [a,*], Yuval Shahar [a], Meirav Taieb-Maimon [a], Guy Bar [b], Avi Yarkoni [b], Ohad Young [a], Susana B. Martins [c], Laszlo Vaszar [c], Mary K. Goldstein [c], Yair Liel [b], Akiva Leibowitz [b], Tal Marom [d], Eitan Lunenfeld [b]

[a] *Medical Informatics Research Center, Department of Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel*
[b] *Soroka Medical Center, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel*
[c] *Geriatric Research Educational and Clinical Center (GRECC), VA Palo Alto Health Care System, Palo Alto, CA, USA*
[d] *Edith Wolfson Medical Center, Holon 58100, Israel*

ARTICLE INFO

Article history:
Received 1 August 2007
Available online 6 May 2008

Keywords:
Clinical decision support systems
Clinical guidelines
Ontologies
Knowledge acquisition
Knowledge bases
Evaluation
Mark-up
Structuring
Completeness
Correctness

ABSTRACT

We introduce a three-phase, nine-step methodology for specification of clinical guidelines (GLs) by expert physicians, clinical editors, and knowledge engineers and for quantitative evaluation of the specification's quality. We applied this methodology to a particular framework for incremental GL structuring (mark-up) and to GLs in three clinical domains. A gold-standard mark-up was created, including 196 plans and subplans, and 326 instances of ontological knowledge roles (KRs). A completeness measure of the acquired knowledge revealed that 97% of the plans and 91% of the KR instances of the GLs were recreated by the clinical editors. A correctness measure often revealed high variability within clinical editor pairs structuring each GL, but for all GLs and clinical editors the specification quality was significantly higher than random ($p < 0.01$). Procedural KRs were more difficult to mark-up than declarative KRs. We conclude that given an ontology-specific consensus, clinical editors with mark-up training can structure GL knowledge with high *completeness*, whereas the main demand for *correct* structuring is training in the ontology's semantics.

1. Introduction

1.1. Clinical guidelines and the importance of their formal representation

Medical practitioners, overloaded with information, do not always have the time, or the computational means, to use the valuable knowledge encoded in clinical guidelines (GLs) during actual patient treatment. Such GLs have the potential both to improve the quality of medical care [1,2] and to contribute to the containment of the costs of care. Although there are thousands of text-based GLs, there is usually no automated support for their specification and application, even though clinicians at the point of care would obviously benefit from such support. Thus, over the past decade, a number of attempts have been made to support complex GL-based care in an automated fashion. Such automated support could assist in a graphical, interactive specification of GLs, search and retrieval of the GLs, patient eligibility determination, runtime application of GLs, and retrospective quality assurance (adherence to GLs). Despite these efforts, most GLs are still text based. Thus, implementing GLs within a computer-based clinical decision support system, i.e., formal GL representation, is fast becoming a critical issue [3].

A recent review [4] has identified the four main areas involved in the development of GL-based decision support systems: (1) GL *modeling and representation*, i.e., the internal format by which a GL is represented in the digital library; (2) GL *specification*, i.e., the act in which an editor creates that representation, typically from a text-based input; (3) GL *verification* and *testing*, i.e., confirming that the GL is in the appropriate format and (potentially) achieves its objectives; and (4) GL *application*, i.e., executing the GL at the point of care. In the current study, we focus on GL specification.

Our recently developed framework for support of GL-based care, the Digital electronic Guideline Library (DeGeL) architecture [5], handles most of the desiderata for GL specification, such as facilitating a gradual, multiple-phase specification process, including mark-up of the GL. (Performing a *mark-up* here means structuring the GL text by labeling portions of the text, using semantic

* Corresponding author. Fax: +972 8 6477161.
E-mail addresses: erezsh@bgu.ac.il (E. Shalom), yshahar@bgu.ac.il (Y. Shahar), meiravta@bgu.ac.il (M. Taieb-Maimon), guybar@bgu.ac.il (G. Bar), yarkoni@bgu.ac.il (A. Yarkoni), ohadyn@bgu.ac.il (O. Young), Susana.Martins@va.gov (S.B. Martins), vaszar.laszlo@mayo.edu (L. Vaszar), Mary.Goldstein@va.gov (M.K. Goldstein), liel@bgu.ac.il (Y. Liel), akival@bgu.ac.il (A. Leibowitz), maromtal@013.net.il (T. Marom), lunenfld@bgu.ac.il (E. Lunenfeld).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.04.009

labels from a chosen target GL-specification language, sometimes even modifying the text.) The gradual-specification process supports different types of users, such as: *expert physicians*, namely, senior, domain-expert clinicians who assist in formation of a clinical consensus that disambiguates the GL; *clinical editors*, namely, medically trained editors who mark-up the GL, and *knowledge engineers*, typically informatics experts who can create a formal GL representation.

1.2. Problems in guideline specification

Despite the considerable work already done in the GL area, the following three challenges have not been considered in sufficient depth, and the current study thus focuses on their clarification:

(1) A comprehensive methodology for the GL-specification process remains to be developed.
(2) Likewise, an *evaluation* methodology to assess the results of the specification is still to be developed.
(3) There are very few quantitative evaluations of GL-specification methodologies in the literature. In particular, there is a lack of appropriate evaluations of GL specification using a gold standard, mainly because of the significant effort required to create such a gold standard and to use it rigorously to evaluate the quality of GL specification. This is especially true in the case of the GL-specification methodology we have evaluated here, since there could be considerable interobserver variability (among different GL-knowledge editors or knowledge engineers) during semantic mark-up [6,7].

With the last challenge in mind, the current study was designed, first, to develop comprehensive, detailed GL specification and evaluation methodologies, and then to answer three specific research questions, defined in the context of these methodological frameworks:

(i) Can clinical editors actually mark-up a GL, and if so, at what quality level?
(ii) Are there differences in the quality of the mark-ups between different clinical editors marking-up the same GL?
(iii) Are there differences in the quality of the mark-ups of different specific aspects of the GL (e.g., eligibility conditions versus objectives)?

To address these three challenges and to answer the three specific research questions raised by assessment of the specification methodology, the current study was performed in three main parts (see Section 3 for details):

(A) We introduced a general methodology for the use of GL-specification tools to specify GLs.
(B) We introduced a general methodology for evaluation of the GL-specification tools.
(C) We then assessed the actual use of that methodology in the case of a particular instance of a GL-specification framework and associated software tool, when used for specification of GLs within three different clinical domains.

2. Background

2.1. Formal representation and specification of clinical guidelines

Automated support for GL application requires formal GL-modeling methods. During the past decade, a number of research groups have devoted considerable efforts to developing computer-interpretable clinical guidelines (CIGs) to support decision-making during clinical encounters [3,8,9]. Most GL-modeling methods use knowledge acquisition tools for eliciting the medical knowledge needed for the knowledge role (KR) classes and subclasses of the GL-specification ontology (i.e., the key concepts and their properties and interrelations) assumed by each method so as to specify it in a formal, executable format. According to the terminology used in the Stepper tool [10,11], there are two main approaches to GL specification: *model-centric*, i.e., modeling the GL *de novo* using a predefined ontology and computational model and referring to the source text solely for documentation, including multiple projects and related tools, such as the EON and PROforma frameworks and the Protégé and Arrezo tools, respectively [12–23]; and, *document-centric*, i.e., starting from a free-text document and mapping it to a given GL ontology manifested in another set of projects and associated tools, such as the GEM Cutter or Delt/A tools [24–28].

2.2. The Asbru guideline-specification language

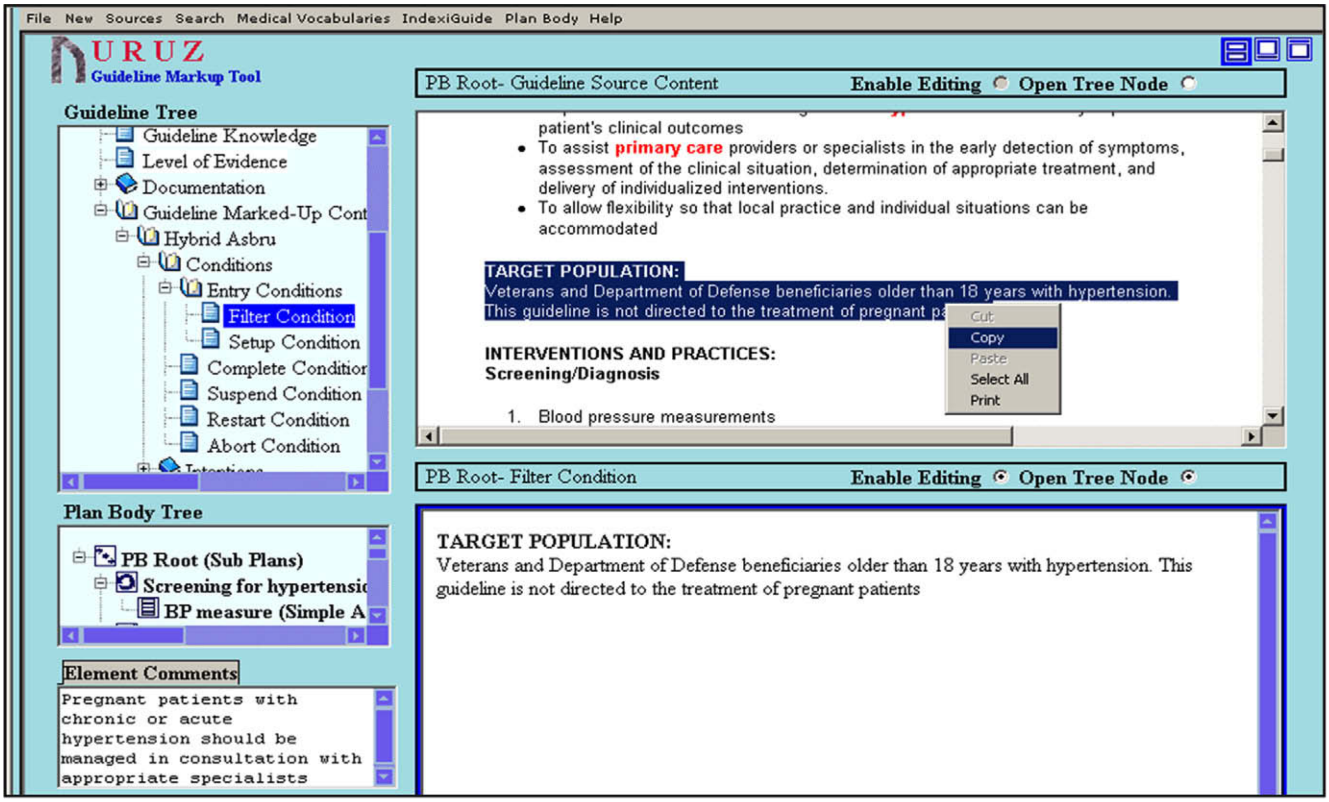
In this study, we used the Asbru language [21] as the underlying GL-representation language. The Asbru-specification language includes semantic KRs organized into classes including (1) *Conditions* (containing, for example, *the filter condition* subclass, which represents obligatory eligibility criteria, *the complete condition* subclass, which halts the GL execution when some predefined criterion is true, and *the abort condition* subclass, which aborts the GL execution when some predefined criterion is true); (2) control structures for the GL's *Plan-Body* (containing, for example, the *sequential, concurrent*, and *repeating* combinations of actions or subguidelines); (3) the GL's *Intentions* (containing, for example, the *process and outcome* intentions subclasses), and (4) the *Context* class of the activities in the GL (containing, for example, the *actors, and clinical context* subclasses). A detailed description of all Asbru KR classes and their constituent KR subclasses can be found in Appendix A.

The Asbru language enables specification of a GL in terms of a hierarchical procedural structure, consisting of *plans* and *subplans*. The plans and subplans are defined through the very act of an editor marking-up the GL so as to segment it into a hierarchical structure of plans, although, presumably, these plans already exist implicitly within the GL.

2.3. The Uruz guideline-specification tool

As discussed above, several challenges relevant to GL specification still require an integrated solution. Thus, to support GL classification, semantic mark-up, context-sensitive search, browsing, run-time application, and retrospective quality assessment, we previously developed the DeGeL architecture and set of tools for classification [5], search, and retrieval [29] and runtime application of the GLs [30]. One of these tools is the document-centric Uruz GL-specification tool (Fig. 1).

Uruz solves a common problem: clinical editors cannot (and need not) program in GL-specification languages, while programmers and knowledge engineers do not always understand the clinical semantics of the GL. One way of addressing this problem is to perform the specification process gradually through several intermediate, semi-structured phases. Uruz enables clinical editors and knowledge engineers to open a text-based GL within it, select a target GL ontology (e.g., Asbru) by which to structure the GL, and drag and drop portions of the text into various nodes and leaves (terminal modes) of the selected GL-ontology's tree, such as into the "entry conditions" and "outcome intentions" knowledge roles. The text is thus implicitly labeled ("*marked-up*") by these semantic tags (this is the "*semi-structured*" representation format). The text



patient's clinical outcomes

- To assist primary care providers or specialists in the early detection of symptoms, assessment of the clinical situation, determination of appropriate treatment, and delivery of individualized interventions.
- To allow flexibility so that local practice and individual situations can be accommodated

TARGET POPULATION:
Veterans and Department of Defense beneficiaries older than 18 years with hypertension.
This guideline is not directed to the treatment of pregnant pa

INTERVENTIONS AND PRACTICES:
Screening/Diagnosis

1. Blood pressure measurements

TARGET POPULATION:
Veterans and Department of Defense beneficiaries older than 18 years with hypertension. This guideline is not directed to the treatment of pregnant patients

```
Pregnant patients with
chronic or acute
hypertension should be
managed in consultation with
appropriate specialists
```

Fig. 1. The Uruz Web-based guideline (GL) mark-up tool in the DeGeL architecture. The tool's basic semi-structuring interface is uniform across all GL ontologies. The target ontology selected by the clinical editor, in this case, Asbru, is displayed in the upper left tree; the GL source is opened in the upper right frame. The clinical editor highlights a portion of the source text (including tables or figures) and drags it for further modification into the bottom frame's Editing Window tab labeled by a semantic role chosen from the target ontology (here, the Asbru filter condition). Contents can be aggregated from different source locations [5].

in the various nodes can then be further modified. The tool also enables the editor to create a hierarchical plan–subplan structure using various types of plans, such as sequential and parallel, and to control operators, such as "if-then-else" (this is the "*semi-formal*" specification level).

Use of the Uruz–DeGeL infrastructure enables clinical editors and knowledge engineers at different sites to collaborate in the process of GL specification and to mark-up the GLs in any of the following three representation levels—(1) a format that we refer to as hybrid representation, i.e., *semi-structured* (labeling, or marking-up, portions of the text with names of KR subclasses from the selected GL ontology, a process typically performed by the clinical editor); (2) *semi-formal* (adding control knowledge and structure, such as a hierarchy of plans and subplans and the order between them—such as in sequence or in parallel—a process typically performed by the clinical editor in collaboration with the knowledge engineer); and (3) *formal* (performing, typically by the knowledge engineer, an executable, ontology-specific representation).

In addition, as part of the current research project, we tested the URUZ tool's usability with the System Usability Scale (SUS) questionnaire [31]. A description of the results is beyond the scope of the current study.

2.4. Previous evaluations of guideline-specification processes and tools

According to Shadbolt et al. [32], "*the main problem in evaluating knowledge acquisition techniques and tools is that they are designed to elicit quality knowledge from human experts. Therefore, when a tool is evaluated, the knowledge that the expert has delivered must be tested too*...". Conducting a comprehensive evaluation is both complex and difficult. However, a formal evaluation of a GL-specification process and/or tools is important, since it will demonstrate to physicians the soundness of the technology and will pinpoint the most important elements of the methodology. Only very few such studies have appeared in the literature; among them are at least four studies that have focused on: (a) evaluations of four different methods—two model-centric [33,34] and two document-centric [27,35]—for procedural knowledge acquisition and representation, and (b) their respective knowledge acquisition tools.

All the evaluations used a small number (two to eight) of experts. Several methods used knowledge engineers as the only users of the knowledge acquisition tools, several used clinical editors as the users, and several, not necessarily in the medical domain, used "domain experts". No method used a gold standard, except the Protégé evaluation (that used a military manual as a gold standard), which is the only method to include a subjective usability measure [33]. None of the evaluations included detailed objective measures to quantify the completeness and correctness of the acquired knowledge; their qualitative results were based mainly on the personal observations of the evaluators.

It has previously been demonstrated that the specification process should involve both clinical editors and knowledge engineers, leading to improved results, as shown by Patel et al. [7], and that inclusion of knowledge engineers is crucial for detection of errors, as recently shown by Peleg et al. [36]. Indeed, the DeGeL architecture has always supported such a joint process [5].

In Section 3, we present an integrated specification and evaluation methodology, which greatly extends the insights of previous studies, regarding the importance of collaboration during the mark-up phase, into the pre-mark-up (consensus formation) and post-mark-up (detailed evaluation) phases.

3. Methods

3.1. The overall guideline-specification and evaluation methodology

The activities in the overall specification and evaluation process included three main phases, i.e., before, during and after mark-up (Fig. 2). The three main specification and evaluation phases and subphases are described below.

3.1.1. Pre-mark-up activities

3.1.1.1. Choosing the specification language. The first step towards specification of GLs was to select the target GL ontology for specification (e.g., GLIF, Asbru). This step was performed by a knowledge engineer who is familiar with the specification language. The choice of the language depended on the purpose of the modeling: for example, although the GEM ontology has been sporadically used within a partial processing engine [37], its main advantage lies in supporting *documentation* of the specification, for example, of the GL's recommendations (i.e., its underlying reasoning, quality of evidence and strength), whereas Asbru is more appropriate for GL planning, runtime application and quality assurance.

3.1.1.2. Learning the specification language. Before performing the mark-up, the expert physicians and clinical editors participated in a special 2-day course, given by the knowledge engineers, to learn the essential concepts and aspects required for the specification process. In particular, the course included: (1) the core semantics and KRs of the specification language (Asbru in our case); (2) the hybrid model and its multiple representation levels; (3) the overall GL representation framework (in this case DeGeL); and (4) the relevant specification tools. For this training program, as for all other phases of the evaluation, a user manual kit was created (known as a "Mark-up Kit"). This kit included explanations and examples of essential aspects such as the specification language (Asbru in our case) and how to use the tools.

3.1.1.3. Selecting a guideline for specification. Typically, a good candidate GL for specification is a GL designed for a common disease with agreement between the majority of expert physicians as to the methods of diagnosis and treatment and with a clear, well-defined clinical pathway. The GL sources, in addition to the expert physicians' own knowledge and interpretations, served as the basis for creating a consensus regarding the GL's semantics. Note that the chosen GL might be selected specifically because of its current or intended intensive use in a local clinical setting. (This was, in fact, the case for all of the three GLs used in our evaluation.)

3.1.1.4. Creating a clinical and an ontology-specific consensus. In general, the textual content of the GL is not always complete or self-evident within itself: it might lack of sufficient information, suffer from ambiguousness, or require customization to local settings. Therefore, local, site-specific customizations of the GL (say, motivated by the availability of resources in the local clinical setting, by local practices, or by personal experience) must be specified explicitly to increase the probability of site-specific successful application. In addition, the same free-text GL might sometimes be interpreted differently by several local expert clinicians; in such cases, much discussion can be prevented by an explicit agreement on a common local interpretation.

Thus, similarly to Miller et al. [38], we too found the creation of a local clinical consensus regarding the semantics of the GL to be an indispensable, mandatory step before mark-up.

We decomposed the crucial mark-up phase into two steps: in the first step, the local most senior expert physicians first created a *clinical consensus,* which was independent of any GL-specification ontology. The clinical consensus was created, for each GL, by the local senior expert physicians and a knowledge engineer. The clinical consensus was always a structured document that described in a schematic-only, but explicit, fashion the interpretation of the clinical directives of the GL, as agreed upon by the local expert physicians.

In the second step, we created what we refer to as an *ontology-specific consensus* (OSC), which specifies the consensus, in terms of the chosen target ontology KRs (e.g., *entry conditions*). The OSC was created by senior expert physicians, who had considerable practical knowledge and experience in the relevant clinical domain, in collaboration with a knowledge engineer, who was familiar with the target-specification language. An example of a free-text segment of a GL, a part of an OSC generated from that GL, and a marked-up plan within the Uruz tool referring to that Figs. B.1–B.3 in Appendix B. A detailed exposition regarding the process of creating an OSC and the evaluation of its effects can be found elsewhere [39].

3.1.1.5. Training the clinical editors in the mark-up tool. The clinical editors who were to perform the mark-ups received instruction from the knowledge engineers regarding the specification tool (e.g., the Uruz mark-up tool), the OSC, and the user manual guide. Within this activity, the clinical editors performed a sample mark-up based on a small portion of a "warm-up" GL. Note that up to this point, all the steps were identical, irrespective of whether an evaluation was being performed or a GL for operational use was being specified.

3.1.1.6. Creating a gold standard. The creation of a gold standard was relevant only for evaluation or for quality-control or quality-assessment purposes. For each of the GLs (or subGLs selected as a sample for quality control), a meticulous, detailed gold standard

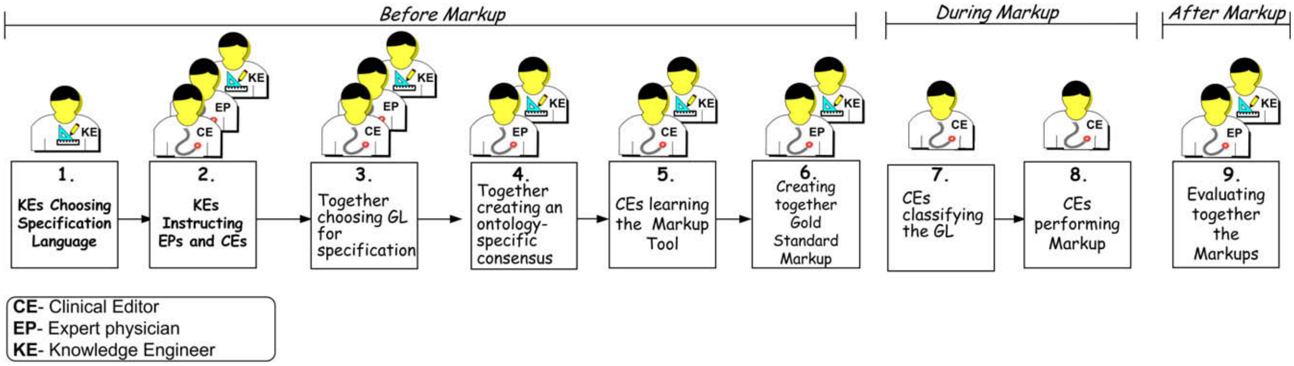

Before Markup
During Markup
After Markup
KE
1. KEs Choosing Specification Language
KE
EP
CE
2. KEs Instructing EPs and CEs
KE
EP
CE
3. Together choosing GL for specification
KE
EP
4. Together creating an ontology-specific consensus
KE
CE
5. CEs learning the Markup Tool
KE
EP
6. Creating together Gold Standard Markup
CE
7. CEs classifying the GL
CE
8. CEs performing Markup
KE
EP
9. Evaluating together the Markups
CE- Clinical Editor
EP- Expert physician
KE- Knowledge Engineer

Fig. 2. The three main phases of the overall specification and evaluation methodology, before, during and after the mark-up, and the activities in each phase. Note the descriptions under each activity. Activity six (creation of a gold standard) can be performed at the beginning or in parallel with activities seven and eight (editors' mark-up). Note also the participants in each activity.

mark-up was created, using the OSC and the GL source(s), by an experienced knowledge engineer, who was well versed in the syntax and semantics of the representation ontology, in cooperation with an expert in the relevant medical domain. Thus, we considered this method for creation of a gold standard to be at least reasonably appropriate for judging whether a mark-up that was created by several less experienced clinical editors was is sufficiently complete and correct. This step was performed before or in parallel to the activity of performing the mark-ups.

3.1.2. Activities during the mark-up

After the clinical editor felt sufficiently confident, he/she began to specify the GL with the mark-up tool (e.g., Uruz) according to the OSC and the GL sources. In addition, to facilitate GL retrieval, the clinical editor classified the GL in terms of multiple clinical indices, by associating the GL with one or more intermediate or leaf nodes of one or more hierarchical *semantic classification axes* (e.g., diagnostic findings/diagnostic ultrasound/chest, as well as treatment/surgery/musculoskeletal System/head), using for that purpose another tool within the DeGeL framework, *IndexiGuide* [5].

In most cases, a knowledge engineer was not a involved in the mark-up session, but he/she was available to help the clinical editor in the event of technical problems (in particular, as our results indicate, it was beneficial to have him/her assist the clinical editor during the specification of complex procedural semi-formal knowledge). Thus, the main activity undertaken during this phase was performing the specification process by the clinical editor, using the knowledge acquisition tool and the OSC.

3.1.3. Post-mark-up activities

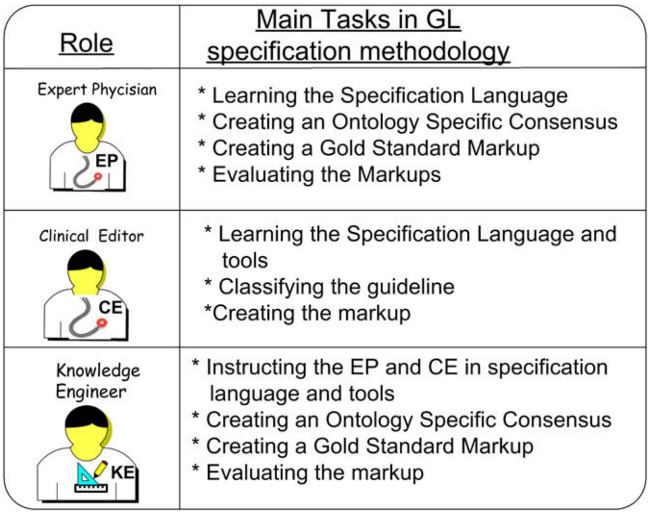
After the mark-ups had been completed, they were evaluated by comparison to the gold standard mark-up according to objective measures. The objective measures were defined in two main categories: a *completeness* measure of the acquired knowledge, i.e., how much content from the gold standard was present in (or absent from) each of the semi-formal mark-ups of each expert physician (for example, a predefined set of plans), and a *correctness* measure, i.e., the correctness of the acquired knowledge in terms of (1) clinical semantics and (2) Asbru semantics (see Section 3.3). Evaluation of the mark-up was important because it helped to evaluate the quality of the mark-up in qualitative and quantitative measures. The evaluation was undertaken by an expert physician in collaboration with a knowledge engineer, using the task-specific evaluation tool that we developed (see Section 3.4); this step enabled scoring of these objective measures for each plan, subplan, and KR subclass instances of an evaluated mark-up. Fig. 3 summarizes the different roles and their tasks within the methodology.

3.2. Implementing and assessing the proposed specification and evaluation methodology

We decided to apply our GL-specification and evaluation methodology to the DeGeL GL architecture, including, in particular, its Uruz Web-based GL-specification tool. As a target GL-specification ontology, we chose the hybrid-Asbru language, which has a structure of multiple representation levels that lent itself well to our collaborative process and which is compatible with the Uruz tool.

3.2.1. Choosing the experts and the clinical domains

Five expert physicians and four clinical editors from three different institutions and three different clinical domains—obstetrics and gynecology, pulmonology, and endocrinology—and two knowledge engineers were enlisted for the mark-up and gold standard tasks. The descriptive details of all expert physicians, clinical editors, and knowledge engineers are given in Table 1. (*Note:* For technical reasons, the role of one of the clinical editors in the obstetrics and gynecology domain was filled by one of the expert physicians, who was involved in the clinical consensus creation task but not in the creation of the gold standard mark-up in that domain.) Three established, published clinical GLs, one from each of these three clinical domains were selected (Table 2). These GLs were chosen because they represent ubiquitous diseases with agreement between the majority of expert physicians as to the methods of diagnosis and treatment and with clear, well-defined clinical pathways.



Role	Main Tasks in GL specification methodology
Expert Phycisian EP	* Learning the Specification Language * Creating an Ontology Specific Consensus * Creating a Gold Standard Markup * Evaluating the Markups
Clinical Editor CE	* Learning the Specification Language and tools * Classifying the guideline *Creating the markup
Knowledge Engineer KE	* Instructing the EP and CE in specification language and tools * Creating an Ontology Specific Consensus * Creating a Gold Standard Markup * Evaluating the markup

Fig. 3. The roles participated in the methodology and their main tasks.

3.2.2. Creating an ontology-specific consensus for each guideline

For each GL, an OSC was established in increasing levels of detail, through the collaboration of a senior expert physician and a knowledge engineer, with minor variations regarding the source of the GL, as follows:

(A) The pelvic inflammatory disease (PID) OSC was created from the CDC source of that GL [40].
(B) The chronic obstructive pulmonary disease (COPD) OSC was intimately related to the source [41] and included practically its entire contents, perhaps because the source was very well organized, structured and divided into sections, which facilitated its conversion into steps.
(C) The primary hypothyroidism (PHT) OSC was based on a standard GL [42]; however, many small, site-specific modifications were added by the expert physician, using mostly knowledge drawn from his personal experience rather than from additional textual sources.

3.2.3. Creating the gold-standard mark-up

For each of the GLs, a gold standard was created together by an expert physician and a knowledge engineer. The expert physicians and the knowledge engineers who specified the gold standard used a pre-defined OSC, which was the same OSC that the expert physicians had used for mark-up.

3.2.4. Performing the mark-ups

For each GL, two mark-ups were created by the clinical editors. Each clinical editor, after training in the use of Uruz, used the mark-up kit, the OSC, the GL sources, and his own knowledge. The PID clinical editors performed the mark-ups in their clinical setting, sometimes in-between treating patients. In addition, in each mark-up, the clinical editor classified the GL using the DeGeL IndexiGuide Tool.

Table 1
The expert physicians (EPs), clinical editors (CEs), and knowledge engineers (KEs) who participated in the study and the tasks they had performed

Participant	Level of training	Site	Tasks
EP1/CE1	Senior clinician (OB/GYN)	BGU-SMC	OSC (PID) Mark-up (PID) Evaluation(PID)
EP2	3rd Yr. Resident (OB/GYN)	BGU-SMC	OSC (PID) Gold standard (PID)
EP3	Senior clinician (Endocrinology)	BGU-SMC	Gold standard (PHT) Evaluation(PHT)
EP4	3rd Yr. Resident (ENT)	Wolfson MC	OSC (PHT)
EP5	Senior clinician (pulmonary diseases and critical care)	VA-PAHCS	OSC (COPD) Gold standard (COPD) Evaluation(COPD)
CE2	Senior clinician (OB/GYN)	BGU-SMC	Mark-up (PID)
CE3	Intern (internal medicine)	BGU-SMC	Mark-up (COPD) Mark-up(PHT)
CE4	Senior clinician (pediatrics)	Stanford, VA-PAHCS	Mark-up (COPD) Mark-up(PHT)

KE1	Information Systems Engineer (Medical informatics)	BGU-MIRC	OSC(PID,COPD)
			Gold standard (PID,COPD)
			Evaluation(PID,COPD, PHT)

KE2	Information Systems Engineer (Medical informatics)	BGU-MIRC	OSC(PHT) Gold standard(PHT)

Note: PID, pelvic inflammatory disease; COPD, chronic obstructive pulmonary disease; PHT, primary hypothyroidism; OSC, ontology-specific consensus; BGU-MC, Ben Gurion University-Soroka Medical Center, Israel; Wolfson MC, Wolfson Medical Center, Israel; BGU-MIRC, Ben-Gurion University-Medical Informatics research center; VA-PAHCS, Veterans Affairs-Palo Alto Health Care System, CA, USA.

Table 2
Selected domains and guidelines (GLs) for evaluation

Domain	GL	Length of textual source	GL description
Obstetrics and Gynecology	Pelvic inflammatory disease (PID) [38,39]	5–8 p.	The GL is intended for use by gynecologists to treat patients suffering from an inflammatory disease related to the pelvis
Pulmonology	Chronic obstructive pulmonary disease (COPD) [40]	7 p.	The GL is intended for use by emergency department physicians and/or general medical ward physicians to treat patients suffering from a low respiratory rate related to chronic obstruction of the pulmonary system
Endocrinology	Primary hypothyroidism (PHT) [41]	13 p.	The GL is intended for use by family practitioners to treat patients suffering from hypothyroidism that is directly related to the thyroid gland (i.e., primary)

3.2.5. Assessing the mark-ups by using objective measures

To obtain meaningful quantitative results, it was decided to measure the mark-up outputs on two scales, i.e., *completeness* of the mark-up and *correctness* of the mark-up. The correctness scale was further split into two aspects, clinical correctness and ontological-semantics correctness (the Asbru ontology, in our case). Here, we stress that this objective measurement was achieved by comparing the mark-ups with their respective gold standards. The evaluation process was performed through the collaboration of an expert physician and a knowledge engineer. Fig. 4 summarizes the study's design, listing the participants in each stage of the evaluation methodology.

3.3. Objective measures used in the study

3.3.1. Completeness measures

When comparing the mark-ups to the gold standard, in general, three types of KRs were defined:

(1) Those that appeared in the gold standard but not in the marked-up version ("Missing").
(2) Those that appeared in the marked-up version, but not in the gold standard ("Redundant").
(3) Those that appeared in both the gold standard and in the marked-up GL ("Existing").

The entire contents of the gold standard should ideally have been included in the mark-ups, i.e., full completeness should have been obtained ("Existing"). However, certain mistakes could have happened, such as the clinical editors skipping some source-GL content or one of the ontological KR subclasses instances. In such a case, content would have been missing from the mark-ups, and the level of completeness would therefore have been lower ("Missing"). Another possible situation was that clinical editors could have added to the mark-up some content that did not exist in the gold standard ("Redundant"); in such a case, the level of completeness would not have been lower, since it had been decided, for the sake of simplicity, to define 100% completeness as the case in which the entire content that existed in the gold standard would also have existed in the mark-up ("Existing" and "Missing").

Completeness denoted the proportion of KR-subclass instances of each marked-up GL relative to the gold standard (a KR subclass instance was defined here as an appearance of a gold standard KR subclass within a plan in the mark-up), but it did not reflect the quality (correctness) of each mark-up KR subclass instance (see Section 3.3.2 for a description of that assessment). We thus defined a *Mean Completeness Score* (MCS) measure for calculating the overall completeness of a mark-up. The MCS of a marked-up GL was defined as the weighted mean of the completeness of the constituent KR classes (e.g., Conditions KR class), each including several KR subclasses (e.g., filter condition). Each marked-up GL was composed of multiple KR subclass instances, and these were compared to the KR subclass instances that should have appeared according to the gold standard. Therefore, MCS_{jk}, the MCS of a mark-up by clinical editor j of GL k, composed of n KR classes c_i out of M possible KR classes ($i = 1,\ldots,M$, where M is the number of KR classes in the ontology; here, $M = 4$), was defined as follows:

$$\mathrm{MCS}_{jk}\% = \frac{\sum_{i=1}^{i=n} \text{Completness of KR class } c_i * \text{No. of gold standard KR subclass instances from class } c_i}{\sum_{i=1}^{i=M} \text{No. of gold standard KR subclass instances from class } c_i} \tag{1}$$

where the completeness of a KR class c was the proportion of KR subclass instances of all KR subclasses in the KR class that existed in the mark-up, compared to those subclass instances that were listed in the gold standard (note that missing KR subclasses were assigned a zero score, although this situation was rare).

Example: If, for the PID GL, marked-up by a clinical editor, 45 of the original 50 gold standard KR subclass instances of the Conditions KR class (of all KR subclasses) were marked-up, i.e., with a completeness of 90%; 67 of the original 70 gold standard KR subclass instances of the Context KR class were marked-up, i.e., with a completeness of 96%; 80 of the original 80 gold standard KR subclass instances of the Intentions KR class were marked-up, i.e., with a completeness of 100%; and 76 of the 80 gold standard KR subclass instances of the Plan-Body KR class were marked-up, i.e., with a completeness of 95%, then the MCS measure for the overall mark-up was:

$$\mathrm{MCS}_{\text{clinical editor PID}} = \frac{50 * 90\% + 70 * 96\% + 80 * 100\% + 80 * 95\%}{50 + 70 + 80 + 80} = 95.7\%$$

3.3.2. Correctness measures

We defined two measures of correctness to quantify the quality of the elicited content of the mark-ups, i.e., a *Clinical Correctness Measure* (CCM) score to measure the clinical correctness of the content and an *Ontological Semantics Correctness Measure* (OSCM) score to measure the ontological correctness of the content (Asbru ontology semantics, in our case). Note that a score was always assigned for both types of correctness measure by comparing the content of the mark-up of the KR subclass instances by an clinical editor to the content of the gold standard KR subclass instances (other more comprehensive measures may be found in Ref. [9]). A correct example of text, OSC representation and mark-up of a GL portion is given in Appendix B. An example of a complete, but incorrect, mark-up might be a cyclical plan which was correctly identified as the gold standard suggested, but that should have used a different dose or repetition frequency.

Table 3 presents the quality scale measures and the possible scores that were assigned to a clinical editor's mark-up content.

Therefore, we defined a *KR Subclass Instance Correctness Measure* (KRCM) for calculating the correctness of each KR subclass instance. A KRCM was assigned to each marked-up KR subclass instance in each plan. Therefore, KRCMl_{lijk}, the KRCM for a mark-up of GL k by clinical editor j for each KR subclass instance l in KR class C_i, can be defined as follows:

$$\mathrm{KRCM}_{lijk} = \frac{\mathrm{CCM}_{lijk} + \mathrm{OSCM}_{lijk}}{2} \tag{2}$$

Thus, each KR subclass instance that was originally marked-up as part of the gold standard was assigned a score in the range of $[-1, 1]$, depending on the quality of the mark-up performed for that KR subclass instance (if any) by the clinical editor who performed the mark-up. We therefore defined the *Mean Quality Score* (MQS) measure for calculating the overall correctness of a mark-up. The MQS of a mark-up was the weighted correctness of its constituent KR classes. Therefore, MQS_{jk}, the MQS of a mark-up by clinical editor j of GL k comprising M KR classes c_i ($i = 1, \ldots, M$), was defined as follows:

$$\mathrm{MQS}_{jk}\% = \frac{\sum_{i=1}^{i=M} \text{Mean correctness of KR class } c_i * \text{No. of gold standard KR subclass instances from class } c_i}{\sum_{i=1}^{i=M} \text{No. of gold standard}\mathit{KR} \text{ subclass instances from class } c_i} \tag{3}$$

where the mean correctness of KR class c_i (relative to the gold standard) was the mean correctness of the KRCM of all its KR subclass instances as they appeared in the gold standard, after having scored them on the basis of the actual mark-up.

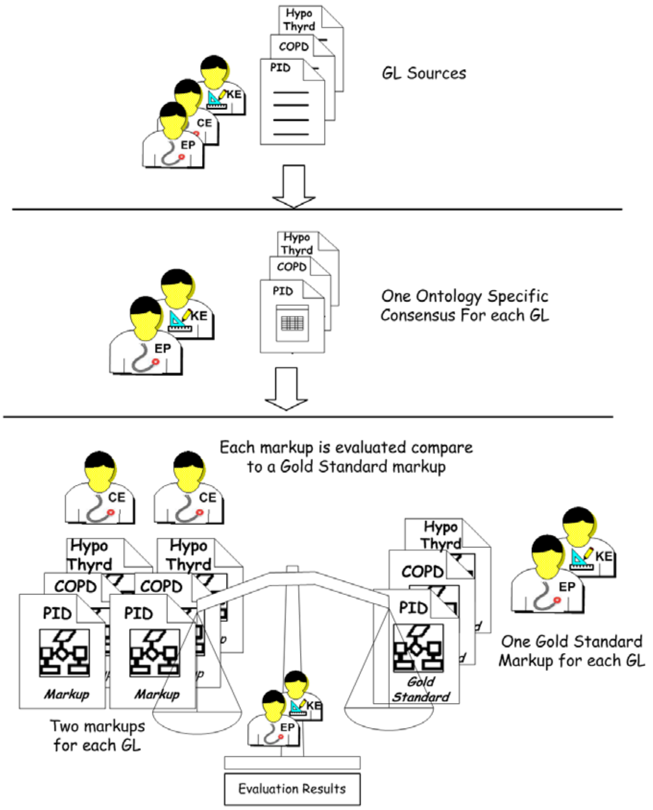

Hypo Thyrd
COPD
PID
KE
CE
EP
GL Sources
One Ontology Specific Consensus For each GL
Each markup is evaluated compare to a Gold Standard markup
Markup
Gold Standard
Two markups for each GL
One Gold Standard Markup for each GL
Evaluation Results

Fig. 4. The evaluation design. Three guideline (GL) sources were used. For each GL, an ontology-specific consensus and gold standard mark-up were created. For each GL, two different mark-ups were performed by separate clinical editors. Each mark-up was evaluated in comparison to the gold standard in clinical and semantic measures.

Table 3
Scoring scales used to grade the quality of the mark-ups, for both the clinical and ontological semantics measures

Scale of measure	Possible score	Description
Clinical correctness measure	1	Clinically correct
	0	Clinically incorrect, without potentially worsening the patient's prognosis
	−1	Clinically incorrect and potentially worsening the patient's prognosis
Ontological semantics correctness measure	1	Correct according to the GL-ontology semantics
	0	Incorrect according to the GL-ontology semantics, without potentially worsening the patient's prognosis
	−1	Incorrect according to the GL-ontology semantics, and potentially worsening the patient's prognosis

A score is always given for both types of measure. When a gold standard knowledge role subclass instance is missing in the mark-up, it is assigned a (−1) score in both measures. The clinical correctness measure is assigned by the expert physician, and the ontological semantic measure is assigned by the knowledge engineer.

Example: If, for the PID GL, as marked-up by a clinical editor, the 50 KR subclass instances of the Conditions KR class (of all its subclasses) that appeared in the gold standard were marked-up with a mean correctness of 0.7; the 70 gold standard KR subclass instances of the Context KR class were marked-up with a mean correctness of 0.7; the 80 gold standard KR subclass instances of the Intentions KR class were marked-up with a mean correctness of 0.8; and the 80 gold standard KR subclass instances of the Plan-Body KR class were marked-up with a mean correctness of 0.8, then the MQS measure for the overall mark-up of the PID GL by the clinical editor was:

$$\mathrm{MQS}_{\text{clinical editor PID}} = \frac{0.7 * 50 + 0.8 * 70 + 0.7 * 80 + 0.8 * 80}{50 + 70 + 80 + 80} = 0.75\%$$

3.4. Mark-up assessment tool

The evaluation methodology was implemented by developing a tool designed to produce completeness and correctness scores, the *Markup-Assessment Tool* (MAT). The MAT is a Web-based desktop application, which was developed using Dot.Net technology and which enabled sharing and collaboration between different sites and users (Fig. 5). The MAT enabled the expert physician, the knowledge engineer, and other guests to select and browse any particular evaluated mark-up from the DeGeL library. An evaluation session usually included the expert physician relevant to the clinical domain of the evaluated mark-up and a knowledge engineer familiar with ontological semantics.

There were two possible working modes in MAT in each evaluation session, *View* mode and *Evaluation* mode. When one of the evaluation managers (usually the knowledge engineer) started the evaluation session, he/she opened the MAT in evaluation mode and entered the participants—expert physicians, knowledge engineers and optional guests—in the appropriate fields in the session. To select and view the mark-up, all other participants at different sites and locations in the session could open the MAT in parallel in view mode. MAT's functionality enabled all the participants to see the changes created by the expert physician and the knowledge engineer who had entered in evaluation mode online during the session. The evaluation manager was required to attach a relevant OSC file when he/she entered the MAT. For example, for the PID mark-up, the OSC of the PID GL was attached. In addition, the content of each KR subclass instance of each plan was evaluated according to clinical and ontological correctness measures.

For each score [−1, 0, 1] of the two correctness measures, there was an appropriate checkbox, in which the evaluation manager could insert a tick after discussion with the expert physician (Fig. 5). Thus, during the evaluation session, the knowledge engineer and the expert physician collaborated by ticking the appropriate completeness and correctness checkboxes for each plan and KR subclasses instances in the gold standard and the evaluated mark-up subclasses instances. There were also checkboxes for each type of error, enabling the knowledge engineer and expert physician to report the error and its type, i.e., clinical or ontological semantics.

3.5. Specific research questions and measurement methods

To evaluate our specification and evaluation methodologies in the case of the URUZ tool and the particular GLs used, we defined three major dimensions for comparison of completeness and correctness: Comparison among the three GLs, among the four editors actually marking the GLs, and among the multiple ontological KR subclasses.

3.5.1. Can clinical editors perform a complete and correct mark-up?

Method of measurement

(1) The completeness level of the plans specified in each mark-up was assessed using the MCS measure (Eq. (1)).
(2) The correctness of the mark-up of each GL was assessed using the MQS measure (Eqs. (2) and (3)).
(3) Testing whether the proportion of scores of 1 (for both CCM and OSCM) was significantly higher than 1/3 for each clinical editor was performed using a *binomial proportions test*[1] [43], in which the scores of −1 and 0 were aggregated as one score versus the score of +1 for the whole mark-up of a clinical editor.

3.5.2. Is there a difference in the completeness and correctness of the mark-ups among clinical editors editing the same GL? Is there a correlation among their KR scores?

Method of measurement

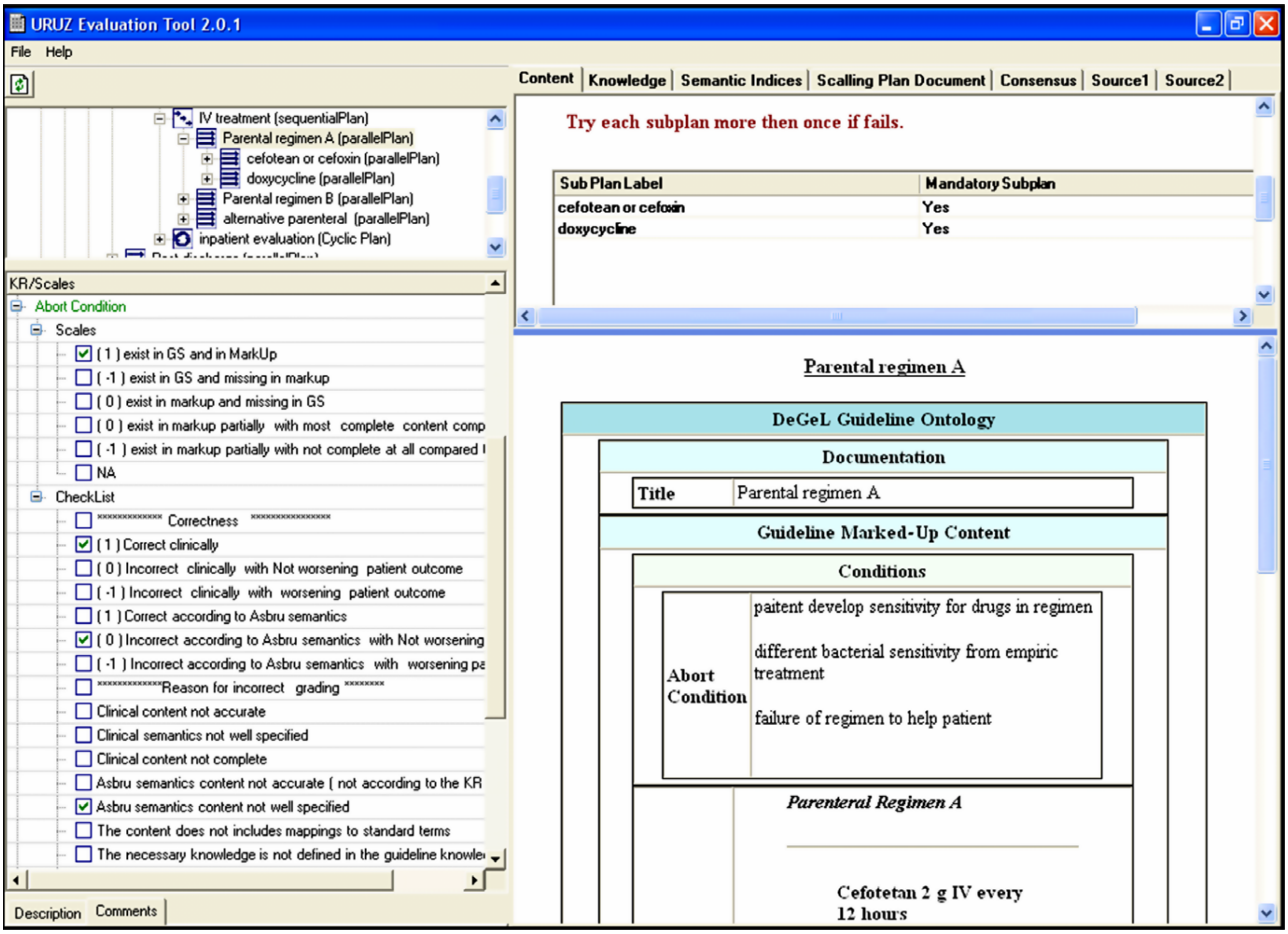
(1) Testing whether there was a significant difference in the completeness level between two clinical editors editing the same GL was performed using a *proportion test*[2] [44] on the number of plans specified by each pair of editors in each GL.
(2) Testing whether there was a significant difference in the KR subclass instances correctness scores between the clinical editor's mark-ups of each GL was performed using the *Wilcoxon signed-ranks test*[3] [43].
(3) Testing whether there was significant correlation between the KR subclass instances correctness scores of the two clinical editors in each GL was performed using *a Gamma correlation test*[4] [43] because the data distribution was *polichotomic* (i.e., in this case, many KR subclass instances with the same

[1] The binomial proportions test procedure compares the observed frequencies of the two categories of a dichotomous variable to the frequencies expected under a binomial distribution with a specified probability parameter (1/3 in our case, that is, a random selection out of three possible scores).

[2] The proportion test considers the problem of testing the hypothesis that the proportion of success in a binomial experiment of two populations is equal.

[3] The Wilcoxon signed-ranks considers the magnitude and direction of the differences between two samples. It gives more weight to a pair that shows a large difference between the two conditions than to a pair which shows a small difference.

[4] The Gamma statistic G measures the relationship between two ordinal scaled variables.



IV treatment (sequentialPlan)
Parental regimen A (parallelPlan)
cefotean or cefoxin (parallelPlan)
doxycycline (parallelPlan)
Parental regimen B (parallelPlan)
alternative parenteral (parallelPlan)
inpatient evaluation (Cyclic Plan)

KR/Scales

Abort Condition

Scales

- ☑ (1) exist in GS and in MarkUp
- □ (-1) exist in GS and missing in markup
- □ (0) exist in markup and missing in GS
- □ (0) exist in markup partially with most complete content comp
- □ (-1) exist in markup partially with not complete at all compared
- □ NA

CheckList

- □ xxxxxxxxxxxxx Correctness xxxxxxxxxxxxxxxx
- ☑ (1) Correct clinically
- □ (0) Incorrect clinically with Not worsening patient outcome
- □ (-1) Incorrect clinically with worsening patient outcome
- □ (1) Correct according to Asbru semantics
- ☑ (0) Incorrect according to Asbru semantics with Not worsening
- □ (-1) Incorrect according to Asbru semantics with worsening pa
- □ xxxxxxxxxxxxxReason for incorrect grading xxxxxxxx
- □ Clinical content not accurate
- □ Clinical semantics not well specified
- □ Clinical content not complete
- □ Asbru semantics content not accurate (not according to the KR
- ☑ Asbru semantics content not well specified
- □ The content does not includes mappings to standard terms
- □ The necessary knowledge is not defined in the guideline knowle

Try each subplan more then once if fails.

Sub Plan Label	Mandatory Subplan
cefotean or cefoxin	Yes
doxycycline	Yes

Parental regimen A

DeGeL Guideline Ontology

Documentation

Title	Parental regimen A

Guideline Marked-Up Content

Conditions

Abort Condition	paitent develop sensitivity for drugs in regimen different bacterial sensitivity from empiric treatment failure of regimen to help patient

Parenteral Regimen A

Cefotetan 2 g IV every 12 hours

Fig. 5. Assessment of the mark-up with the mark-up assessment tool (MAT). Note the following features: selected plan in the top left panel; the checkboxes for the completeness and correctness measures in the bottom left panel (in this case of the Abort Condition knowledge role (KR) subclass); the procedural content of the KR subclass instance in the top right above panel, and its textual content in the bottom right panel.

value of three distinct scores of [−1,0,1]). Thus, a correlation test that casts the data in the form of a contingency table was more appropriate than a standard correlation test (such as a *Spearman* test).

3.5.3. Is there a difference in the correctness scores of the specification of different knowledge-roles?

In particular, we sought to investigate: whether the proportion of scores of 1 (out of −1,0,1) was high for each KR class and for each KR subclass (i.e., significantly greater than 1/3, 1/2, etc.) over all the clinical editors' mark-ups; which KR classes and KR subclasses were easy (or not) to structure; and whether there was a significant difference between the KR classes and between the KR subclasses.

Method of measurement

(1) Testing whether the proportion of correctness scores of 1 (taking the clinical and ontological scores together) was high (i.e., significantly greater than 1/3, 1/2, etc.) for each KR class and for each KR subclass was performed using a *binomial proportions test*, in which the scores of −1 and 0 were aggregated as one score versus the score of +1 for the whole KR class and KR subclass.

(2) We then split the KR classes by their proportion of correctness scores of 1 into several groups. Testing whether these groups indeed represented several different clusters was performed by first demonstrating their homogeneity using a Kruskal–Wallis test [43]. To ascertain that the difference between the groups was indeed significant, a proportions test for the proportion of scores of 1 between each pair of homogenous groups of KR subclasses found by us was performed.

4. Results

4.1. Feasibility of mark-up by clinical editors

Although not always available, over the course of several months, all clinical editors eventually found sufficient time to edit the GLs assigned to them to the point at which they were personally satisfied that the structured GL faithfully represented the text-based GL. Tables 4–6 summarize the results for completeness and correctness of the mark-ups for all GLs, using the MCS and MQS measures defined above. Overall, the gold standard mark-ups included a total of 196 different plans and subplans that should have been marked-up for the three GLs by the clinical editors, i.e., 106 plans for the PID GLs, 59 for the COPD GLs, and 31 for the PHT GL (Table 4). In addition, a total of 326 KR subclass instances defined in the gold standard mark-up were assessed.

A CCM score and an OSCM score were assigned for each KR subclass instance within each plan and subplan, in each mark-up for

Table 4
Summary of the completeness of the mark-up of each clinical editor (CE) regarding number of gold standard (GS) plans created in the mark-up

		Missing group—plans that exist in the GS but not in the mark-up (%)	Existing group—plans that exist in the GS and in the mark-up (%)	Redundant group—plans that do not exist in GS but do exist in the mark-up (%)	No. of GS plans
PID	CE1	0 (0)	106 (100)	1 (1)	106
	CE2	3 (3)	103 (97)	0 (0)	106
COPD	CE3	0 (0)	59 (100)	1 (2)	59
	CE4	4 (7)	55 (93)	2 (3)	59
PHT	CE3	3 (10)	28 (90)	0 (0)	31
	CE4	2 (6)	29 (94)	0 (0)	31
Weighted mean		2 (3)	63 (97)	1 (1)	

Table 5
Summary of the completeness and correctness measures for all guideline mark-ups categorized by knowledge roles classes

KR class	Measure—all GLs		
	Number	Completeness (MCS), [−1,1]% ± SD	Correctness (MQS), [−1,1] ± SD
Context	47	68 ± 0.01	0.39 ± 0.92
Intentions	26	96 ± 0.04	0.89 ± 0.41
Conditions	60	88 ± 0.1	0.45 ± 0.69
Plan-Body	193	97 ± 0.02	0.66 ± 0.65
All KR classes	326	91 ± 0.11	0.6 ± 0.7

Completeness is presented as a percentage; correctness is presented on a scale of [−1,1].

Table 6
Proportion of each score of all knowledge role subclass instances in each mark-up of a clinical editor (CE) categorized by the correctness scores [−1,0,1][a]

Correctness score	PID (No. KRs = 368)		COPD (No. KRs = 194)		PHT (No. KRs = 96)	
	CE1****	CE2**	CE3*	CE4***	CE3***	CE4****
1	85.87	71.74	43.30	76.29	79.17	92.71
0	10.87	15.22	24.74	8.76	10.42	1.04
−1	3.26	13.04	31.96	14.95	10.42	6.25

[a] Values in the table are given in %.
* Proportion of scores of “1” significantly ($p < 0.05$) greater than 0.33.
** Proportion of scores of “1” significantly greater than 0.65.
*** Proportion of scores of “1” significantly greater than 0.7.
**** Proportion of scores of “1” significantly greater than 0.75.

all GLs, i.e., 180 KR subclass instances for the PID GLs, 97 KR subclass instances for the COPD GLs, and 49 for the PHT GL.

The completeness of the mark-ups was very high, with a mean of 97% for the plans re-created by the clinical editors marking-up the GLs and 91% ± 0.11 for the KR subclass instances marked-up for all the GLs. In contrast, there was some variability between the correctness levels of the KR subclasses instances of each KR class (mean of 0.6 ± 0.7) (Table 5).

Since the correctness scores [−1,0,1] were not continuous, we looked at a different aspect for analyzing these results, i.e., the proportion of “1” quality scores (both CCM and OSCM, Table 3) assigned to all KR subclass instances for each GL and each clinical editor. We tested whether the proportion of scores of “1” assigned to all KR subclass instances was high (significantly higher than 0.33, 0.65, etc.) for each clinical editor and each GL (Table 6). The proportion of scores of 1 in all the mark-ups of the clinical editors was significantly higher than the baseline random proportion of 0.33 ($p < 0.01$). Furthermore, except for one of the clinical editors in the case of the COPD GL, the proportion of scores of 1 was significantly higher than 0.65, and for one of the clinical editors in the case of the PID GL and one of the clinical editors in the case of the PHT GL, it was significantly higher than 0.75.

Our conclusion, given the high completeness of the mark-up performed by all clinical editors, with respect to both subplans and KR subclass instances, and the significantly high proportions of scores of 1 assigned to almost all of the clinical editor and GL combinations, is that with appropriate, albeit quite limited training, clinical editors can perform semi-structured and semi-formal mark-ups.

4.2. Differences between mark-ups performed by different clinical editors

The results of the analyses showed that in almost all KR classes there were no significant differences ($p > 0.05$) in the completeness measure between the two clinical editors marking-up the same GL: they both performed the mark-up with a high completeness for all GLs. However, the correctness measure varied depending on the particular clinical editor: there were often significant differences ($p < 0.05$) between the correctness scores of the clinical editor pairs structuring each GL, and the correlation coefficients between the KR subclass instances correctness scores of the two clinical editors in each GL were low and not significant ($p > 0.05$) (for detailed results and further description of all results see Ref. [9]). An example of disagreement and/or different interpretations of the same GL may be found in the definition of which plans are mandatory for successful execution in a set of subplans: one clinical editor defined all subplans as mandatory, while the other clinical editor defined, for the same set of subplans, only some of the plans as being mandatory.

Thus, our conclusion, given those results, is that clinical editors can perform mark-ups with high completeness. In addition, although it seems that the OSC helped the clinical editors to structure the GLs using the Asbru ontological semantics, there were still disagreements and different interpretations of the same GL, a fact that further emphasizes a need for a more detailed OSC and perhaps a need for additional training of the mark-up editors.

4.3. Differences in the quality of mark-ups among different KRs

The results of the analyses showed that for all KR subclasses, the proportions of correctness scores of 1 was significantly ($p < 0.05$) higher than 0.33. For most of the KR classes and subclasses, the proportion was even significantly higher than 0.6. There were no significant differences ($p > 0.05$) between the clinical and ontological parts in the proportions of scores of 1 for all KR subclasses. Sorting of the KR subclasses revealed that they fell into the following three groups, under the assumption that the easier it is to structure a KR subclass, the higher the resulting correctness scores:

(1) *Easy:* to structure (a group whose proportion of mark-up scores of 1 was significantly higher than 0.75)—this group included all the Intention and most of the Plan-Body KR subclasses.
(2) *Intermediate:* difficulty to structure (a group whose proportion of scores of 1 was significantly higher than 0.6 and 0.5, but not significantly higher than 0.75)—this group comprised mostly the Context KR subclasses, the Plan-Body's sequential order KR subclass and Abort Condition KR subclasses.
(3) *Difficult:* to structure (a group whose proportion of scores of 1 was significantly higher only than 0.33)—this group included mostly procedural KR subclasses (such as Plan-Body's parallel order and repeating plan KR subclasses), but also the Complete Condition and Filter Condition KR subclasses.

When we performed a Kruskal–Wallis test on the scores of each set of KR subclasses in each group, we found that there were no significant differences ($p > 0.05$) between the scores of the KR subclasses within each group, i.e., each group was homogenous. Finally, the proportions tests showed that for each pair of groups, the proportions of scores of 1 were significantly different from each other ($p < 0.05$).

Our conclusions, given those results, are that clinical editors can perform mark-up with a high proportion of scores of 1 for all KR classes and KR subclasses and with high correctness of both the clinical and ontological score measures. Declarative KR subclasses are easier for clinical editors to structure than procedural KR subclasses.

Although there seemed to be a problem in marking-up the Conditions KR class, in particular the Filter (compulsory eligibility) condition, we felt that this finding might be somewhat spurious and could have arisen due to the lack of explicit specification of the AND/OR operators in the OSC, thus making it more difficult for the expert physicians who marked-up the GL. We therefore concluded that logical operators must be carefully defined during the creation of the OSC.

If the Condition KR class is ignored in this clustering, it is clear that the KR subclasses that were difficult to mark-up were the procedural subclasses, namely, those that expressed different types of procedural control (e.g., Plan-Body's parallel order).

5. Discussion

In this study, we proposed a methodological remedy for the lack of comprehensive specification and evaluation methodologies and demonstrated its validity in the particular case of the Uruz tool, while answering several specific questions regarding the feasibility of GL specification.

5.1. Implications of the study

Although it seems that the creation of an OSC is time consuming and creates a bottle neck in the specification process, the opposite might be true, i.e., in practical terms, it might actually save time and money: instead of a programmer creating and maintaining a version of a GL, which had not necessarily been created in collaboration with an clinical editor, we suggest a process in which GLs are specified, as OSCs, as soon as possible in the process with agreement among all participants. It is our feeling, based on the results of the study, that the OSC, when properly created by the expert physicians and knowledge engineers, extracts most of the tacit knowledge from the source GL and thus enables clinicians, who are not necessarily experts in the domain but who are proficient in the use of the mark-up tool, to correctly mark-up the source GL.

Furthermore, as we further elaborate in the conclusions (see Section 6.2), the results of the study suggest that the specification (i.e., clinical editing) process itself can be performed by physicians who are less professionally advanced but who have sufficiently high mark-up skills.

It should be emphasized that demonstrating the capability of clinicians at any level to perform key portions of the mark-up (given an OSC) is a major contribution to the field, since many frameworks are based on GL-specification tools that are used mostly or solely by knowledge engineers, who obviously have significant computational skills [10,12–14,23,26,27,34,35].

5.2. Limitations of the study

At first glance, it might seem that an apparent limitation of this study lies in the small numbers of GLs, expert physicians and clinical editors, which is generally a common limitation in knowledge acquisition evaluations. However, in fact, an overall total of 196 subplans and 326 KR subclass instances, in three clinical domains, were structured by the clinical editors in all the mark-ups, enabling us to evaluate in detail all the specification phases from free-text representation, through the semi-structured format, into the semi-formal representation. We consider these data, collected over more than four years of the study from beginning to end, not only to be sufficient for proving the feasibility of the overall specification and evaluation methodology, and for multiple types of sophisticated statistical analysis performed in this study, but also to be unique in their very existence, particularly since the lack of a gold standard for clinical knowledge specification in many studies has prevented an objective assessment of the capabilities of different methodologies and tools for specification of procedural clinical knowledge. However, the small number of clinical editors was insufficient for performance of a cluster analysis among the clinical editors.

Another potential limitation of the current study was the lack of careful measurement of the required time: since the clinical editors worked mainly in their spare time (which was limited), it was difficult to evaluate the influence of the time variable on this research and to measure, say, the precise time it took a clinical editor to structure a GL or each KR subclass instance. On the other hand, the lack of time constraints enabled us to obtain realistic results, since the interaction with most of the clinical editors took place in their own "playgrounds", i.e., the editing mark-up task was often performed within clinical settings in the clinical editors' spare time between their immediate tasks of treating patients. This situation was thus as close as possible to our objective of examining the option of specification of GLs by clinical editors, who are themselves involved at the point of care, and not in the "artificial" environment of a laboratory.

5.3. Future work

A future study, employing our methodology, and performed using a sufficiently large number of clinical editors, might be able to answer such questions. A study with a large number of experts and editors might be able to relate the usability of the GL-specification tool to expert and editor features such as gender, computer expertise, and clinical training level.

6. Conclusions

In this study, a three-phase, nine-step methodology for the overall specification and evaluation process of GLs from a textual representation into a semi-formal representation was developed, and an actual, fully quantitative assessment of that methodology using a particular GL-specification set of tools (the DeGeL architecture) was performed in three clinical domains. The core specification and evaluation methodology was also validated in this study. Furthermore, the implementation study was designed to answer three categories of specific research questions, for which now have answers, at least in the context of our particular study:

(1) Clinical editors with appropriate training could actually mark-up a GL with high levels of completeness, and with variable, although mostly high or very high, levels of correctness.
(2) There were no significant differences between different clinical editors in the *completeness* of the mark-ups of the same GL; however, there were significant differences in *correctness*.
(3) Clinical editors seemed to have greater difficulty in correctly marking-up certain KR subclasses, particularly, procedural KR subclasses.

In the light of these conclusions, four recommendations for improving the GL-specification process can now be proposed (Sections 6.1–6.4).

6.1. The importance of creating an ontology-specific consensus

We found that collaboration between a clinical editor and a knowledge engineer was crucial to the success of the formal specification of a GL. In particular, we feel that creation of an OSC was an indispensable, crucial step. The experience we gained in this study suggests that the creation of the OSC should begin with creating a *clinical consensus* by a group of expert physicians (probably the most senior expert physicians in the chosen clinical domain) of the local medical setting in collaboration with a knowledge engineer (or knowledge engineers) familiar with the specification language and, probably, with the clinical editors who would actually mark-up the GL. The final consensus should be reached in the terms of the eventual target GL ontology in which the GL is to be represented. Indeed, we found a positive and significant correlation between the quality of the OSC and the overall clinical and ontological correctness scores [39]. In the creation of an OSC, consideration should be given to the possibility of involving psychologists to explore the issue of "group thinking" so as to reap optimal benefit from this step. It might also be useful to employ graphical tools to facilitate creation of the OSC. Finally, we suggest that re-using and sharing of OSCs among expert physicians and clinical settings should be supported by saving the OSCs in an appropriate digital library. The DeGeL framework would be suitable for this task.

6.2. Medical and computational qualifications needed for GL specification

Once an OSC has been prepared by expert physicians, assisted by a knowledge engineer, the results suggest that any clinical editor (senior, non-senior or a general physician) can structure the GL's knowledge in a semi-formal representation. In other words, clinical domain expertise is not essential for mark-up (for example, only two of the mark-up editors in this study were experts in the domain of the GL they marked-up).

However, due to the computational semantics of GL ontologies and the nature of GL mark-up tools, correct mark-up of a GL requires a clinical editor with a good understanding of the target ontology and at least a minimal set of computer skills. The clinical editor could be less clinically experienced, but should be able to perform complex tasks involving computer tools, such as mark-up. This assessment is strengthened by a different part of our study [9], suggesting that in forming the OSC, the main requirement is an understanding of the target ontology (and not the mark-up tools), whereas in creating the mark-up, the main requirement is knowledge of the mark-up tools.

Finally, most clinical editors were completely inexperienced in knowledge engineering, and thus we expect the results to improve when better trained clinical editors are used. Thus, if sufficient training in the semantics of the target ontology is provided and the methodology introduced in this study is used, the mark-up process might well be performed by medical students, interns, or general practitioners, working together with a knowledge engineer.

6.3. Support of quality control during guideline specification

After a mark-up of the GL has been performed by a clinical editor, evaluating the quality of the mark-up using the MAT should be done by a knowledge engineer with the assistance of a senior expert physician, who would act as a "referee" for measuring the completeness and correctness measures. However, in addition to its post-mark-up use in a full evaluation study, a version of the MAT would be potentially useful for practical, quality-assessment purposes of sample mark-ups *during* the specification and implementation of new GLs in a particular clinical site. The MAT could also be used as a knowledge-visualization and browsing tool to examine each portion of the specified clinical procedural knowledge.

6.4. Need for a graphical specification tool

As part of this study, we found that the usability of the Uruz specification tool was ranked by the clinical editors as rather low on a standard usability scale, despite its high functionality [9]. Thus, we feel that research should be devoted to the use of graphical authoring tools. Indeed, given this initial insight, a prototype graphical interface has already been developed by members of our group and will be evaluated in the near future [45].

In conclusion, we trust that our study—and the methodologies proposed in it—will serve as a benchmark for other ongoing evaluations and that its insights can be used to increase the methodological use and the rigorous evaluation of knowledge acquisition tools in various medical domains, both for operational and for educational purposes.

Acknowledgments

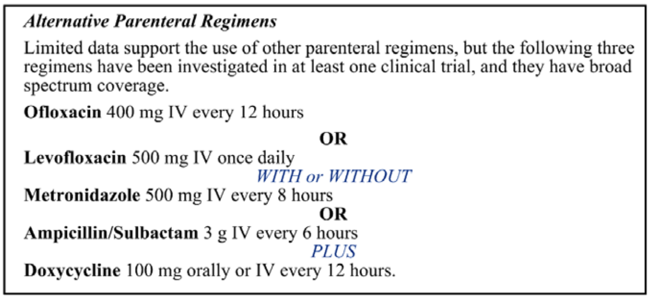
We thank Dr. L. Basso and Dr. H. Kaizer for their efforts and contribution to this research in the early evaluations of the mark-up tools. This research was supported in part by NIH award number R01-LM-06806, by an IBM faculty fellowship (Y.S.), and by Deutsche Telekom Laboratories (DT-Labs) at Ben-Gurion University of the Negev. The views expressed are those of the authors and not necessarily those of the Department of Veterans Affairs.

Appendix A. The Asbru knowledge role (KR) classes and subclasses

See Table A.1.

Appendix B. Mark-up example

See Figs. B.1–B.3.



Alternative Parenteral Regimens

Limited data support the use of other parenteral regimens, but the following three regimens have been investigated in at least one clinical trial, and they have broad spectrum coverage.

Ofloxacin 400 mg IV every 12 hours

OR

Levofloxacin 500 mg IV once daily

WITH or WITHOUT

Metronidazole 500 mg IV every 8 hours

OR

Ampicillin/Sulbactam 3 g IV every 6 hours

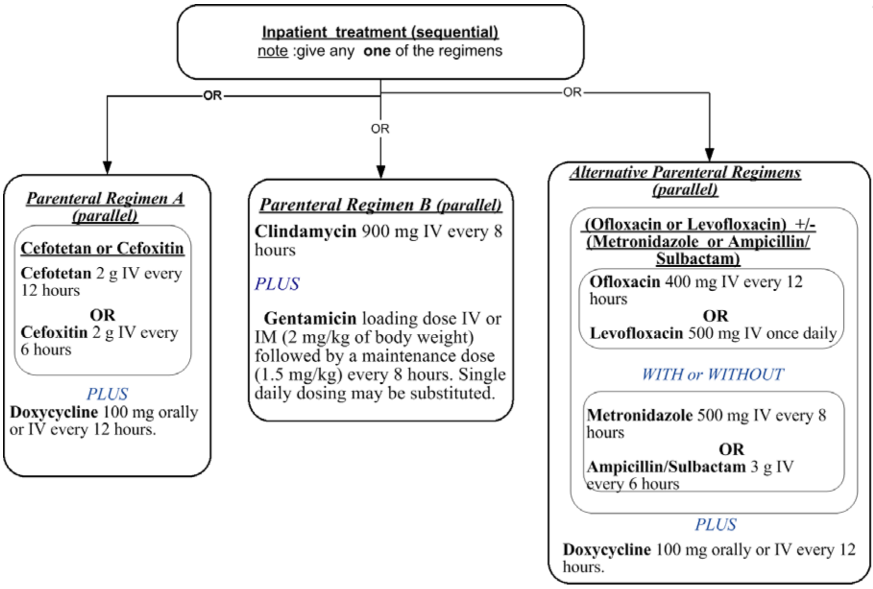
PLUS

Doxycycline 100 mg orally or IV every 12 hours.

Fig. B.1. A small sample from a guideline (original) textual source.

Table A.1
The Asbru knowledge role (KR) classes and the relevant KR subclasses in each KR class

KR class	KR subclass	Description
Context	Actors	Specifies who is responsible or taking part in performing the guideline actions
	Clinical context	Specifies where in the clinical setting the patient is being examined
Intentions	Overall—outcome	The state(s) that should achieved, or maintained, or avoided after finishing the plan
	Overall—process	The action(s) that should take place after finishing the plan
	Intermediate outcome	The state(s) that should be achieved, or maintained, or avoided during the process of the plan
	Intermediate—process	The action(s) that should take place during the process of the plan
Conditions	Filter condition	Specifies the exclusion/inclusion criteria of the guideline
	Setup condition	Specifies the additional criteria that should be achieved through actions of the physician prior to the start of plan application
	Abort condition	Specifies when a plan must end unsuccessfully
	Suspend condition	Specifies when a plan must be put on a hold
	Reactivate condition	Specifies when a plan can be reactivated after being suspended
	Complete condition	Specifies when a plan can end successfully
Plan-Body	Simple action	An atomic plan with simple semantics; suitable for defining plans with a single action
	If-then-else	A condition between two plans
	Repeating plan	A plan that should be repeated more than on time once in periods
	Subplans—parallel order	There are two or more subplans that overlap
	Subplans—sequential order	At any moment in time, only one subplan is performed
	Plan-activation	A plan defined in DeGeL
	Switch case	The criteria have some possible values
	To be defined	This plan is not in the scope of this guideline, or needs to be defined later



Inpatient treatment (sequential)
note :give any one of the regimens

Parenteral Regimen A (parallel)

Cefotetan or Cefoxitin

Cefotetan 2 g IV every 12 hours

OR

Cefoxitin 2 g IV every 6 hours

PLUS

Doxycycline 100 mg orally or IV every 12 hours.

Parenteral Regimen B (parallel)

Clindamycin 900 mg IV every 8 hours

PLUS

Gentamicin loading dose IV or IM (2 mg/kg of body weight) followed by a maintenance dose (1.5 mg/kg) every 8 hours. Single daily dosing may be substituted.

Alternative Parenteral Regimens (parallel)

(Ofloxacin or Levofloxacin) +/- (Metronidazole or Ampicillin/ Sulbactam)

Ofloxacin 400 mg IV every 12 hours

OR

Levofloxacin 500 mg IV once daily

WITH or WITHOUT

Metronidazole 500 mg IV every 8 hours

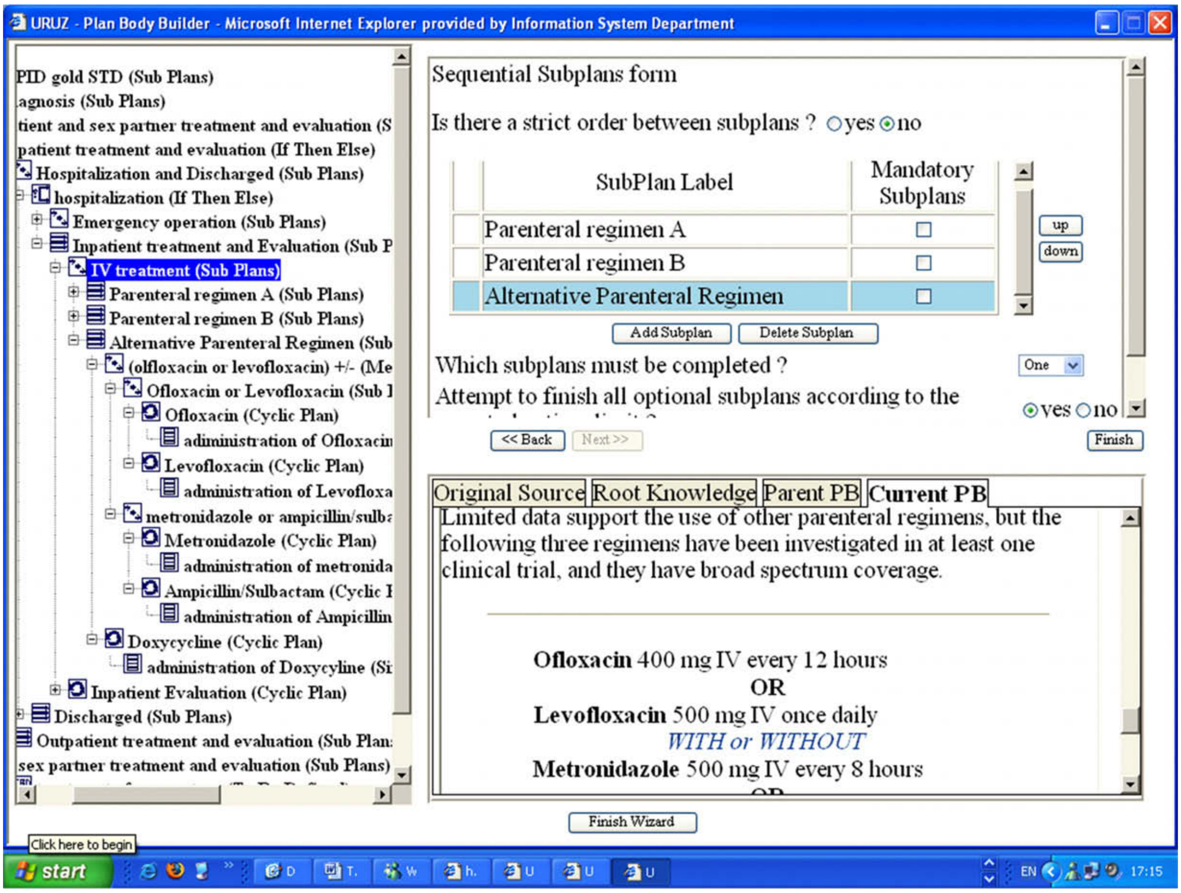
OR

Ampicillin/Sulbactam 3 g IV every 6 hours

PLUS

Doxycycline 100 mg orally or IV every 12 hours.

Fig. B.2. An example of an ontology-specific consensus. In this case, the consensus is specific to the Asbru ontology and is a small part of the consensus document that refers to the guideline shown in Fig. B.1.



URUZ - Plan Body Builder - Microsoft Internet Explorer provided by Information System Department

PID gold STD (Sub Plans)
agnosis (Sub Plans)
tient and sex partner treatment and evaluation (S
patient treatment and evaluation (If Then Else)
Hospitalization and Discharged (Sub Plans)
hospitalization (If Then Else)
Emergency operation (Sub Plans)
Inpatient treatment and Evaluation (Sub P
IV treatment (Sub Plans)
Parenteral regimen A (Sub Plans)
Parenteral regimen B (Sub Plans)
Alternative Parenteral Regimen (Sub
(olfloxacin or levofloxacin) +/- (Me
Ofloxacin or Levofloxacin (Sub I
Ofloxacin (Cyclic Plan)
administration of Ofloxacin
Levofloxacin (Cyclic Plan)
administration of Levofloxa
metronidazole or ampicillin/sulba
Metronidazole (Cyclic Plan)
administration of metronida
Ampicillin/Sulbactam (Cyclic I
administration of Ampicillin
Doxycycline (Cyclic Plan)
administration of Doxycyline (Si
Inpatient Evaluation (Cyclic Plan)
Discharged (Sub Plans)
Outpatient treatment and evaluation (Sub Plan:
sex partner treatment and evaluation (Sub Plans)

Sequential Subplans form

Is there a strict order between subplans ? ○ yes ● no

SubPlan Label	Mandatory Subplans
Parenteral regimen A	□
Parenteral regimen B	□
Alternative Parenteral Regimen	□

Which subplans must be completed ? One

Attempt to finish all optional subplans according to the ● yes ○ no

Original Source | Root Knowledge | Parent PB | Current PB

Limited data support the use of other parenteral regimens, but the following three regimens have been investigated in at least one clinical trial, and they have broad spectrum coverage.

Ofloxacin 400 mg IV every 12 hours
OR
Levofloxacin 500 mg IV once daily
WITH or WITHOUT
Metronidazole 500 mg IV every 8 hours

Fig. B.3. An example of use of the Uruz guideline-specification tool. In this case, the ontology-specific consensus shown in Fig. B.2 is marked up by a clinical editor and is being converted into a semi-formal representation, as a hierarchical set of plans and subplans with various control-structure specifications (e.g., "in sequence" and "in parallel").

References

[1] Grimshaw JM, Russel IT. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. Lancet 1993;342:1317–22.

[2] Grimshaw JM, Eccles MP. Is evidence-based implementation of evidence-based care possible? Med J Aust 2004;180(6 Suppl.):S50–1.

[3] Wang D, Peleg M, Tu S, Boxwala A, Greenes R, Patel V, et al. Representation primitives, process models and patient data in computer-interpretable clinical practice guidelines: a literature review of guideline representation models. Int J Med Inform 2002;68(1–3):59–70.

[4] De Clercq P, Blom J, Korsten H, Hasman A. Approaches for creating computer-interpretable guidelines that facilitate decision support. Artif Intell Med 2004;31(1):1–27.

[5] Shahar Y, Young O, Shalom E, Galperin M, Mayaffit A, Moskovitch R, et al. A framework for a distributed, hybrid, multiple-ontology clinical-guideline library and automated guideline-support tools. J Biomed Inform 2004;37(5):325–44.

[6] Ohno-Machado L, Gennari JH, Murphy SN, Jain NL, Tu SW, Oliver DE, et al. The guideline interchange format: a model for representing guidelines. J Am Med Inform Assoc 1998;5(4):357–72.

[7] Patel VL, Allen VG, Arocha JF, Shortliffe EH. Representing clinical guidelines in GLIF: individual and collaborative expertise. J Am Med Inform Assoc 1998;5(5):467–83.

[8] Peleg M, Tu S, Bury J, Ciccarese P, Fox J, Greenes RA, et al. Comparing computer-interpretable guideline models: a case-study approach. J Am Med Inform Assoc 2003;10(1):52–68.

[9] Shalom E. An evaluation of a methodology for specification of clinical guidelines at multiple representation levels. Master Thesis. Ben-Gurion University of the Negev, Beer-Sheva, Israel; 2006.

[10] Ruzicka M, Svatek V. Markup based analysis of narrative guidelines with the Stepper tool. Stud Health Technol Inform 2004;101:132–6.

[11] Svatek V, Ruzicka M. Step-by-step markup of medical guideline documents. Int Med Inform 2003;70:329–35.

[12] Musen MA, Tu SW, Das AK, Shahar Y. EON: a component-based approach to automation of protocol-directed therapy. J Am Med Inform Assoc 1996;3:367–88.

[13] Tu SW, Musen MA, Shankar R, Campbell J, Hrabak K, McClay J, et al. Modeling guidelines for integration into clinical workflow. Medinfo 2004:174–8.

[14] Johnson P, Tu S, Jones N. Achieving reuse of computable guideline systems. Medinfo 2001;10(Pt. 1):99–103.

[15] Boxwala AA, Peleg M, Tu S, Ogunyemi O, Zeng QT, Wang D, et al. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. J Biomed Inform 2004;37(3):147–61.

[16] Noy NF, Crubezy M, Fergerson RW, Knublauch H, Tu SW, Vendetti J, et al. Protégé-2000: an open-source ontology-development and knowledge-acquisition environment. In: AMIA Annu Symp Proc; 2003, p. 953.

[17] Sutton DR, Fox J. The syntax and semantics of the PROforma guideline modelling language. J Am Med Inform Assoc 2003;10(5):433–43.

[18] Terenziani, P, Montani S, Bottrighi A, Torchio M, Molino G. Supporting physicians in taking decisions in clinical guidelines: the GLARE "What if" facility. In: Proc AMIA Symp; 2002, pp. 772–76.

[19] Quaglini S, Stefanelli M, Lanzola G, Caporusso V, Panzarasa S. Flexible guideline-based patient careflow systems. Artif Intell Med 2001;22:65–80.

[20] Gordon C, Herbert I, Johnson P, Nicklin P, Pitty D, Reeves P. Telematics for clinical guidelines: a conceptual modelling approach. Stud Health Technol Inform 1997;43 Pt. A:314–8.

[21] Shahar Y, Miksch S, Johnson P. The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. Artif Intell Med 1998(14):29–51.

[22] Miksch S, Kosara R, Shahar Y, Johnson P. AsbruView: visualization of time-oriented, skeletal plans. In: Proceedings of the 4th international conference on artificial intelligence planning systems (AIPS-98); 1998.

[23] De Clercq P, Hasman A. Experiences with the development, implementation and evaluation of automated decision support systems. Medinfo 2004;11(Pt. 2):1033–7.

[24] Haschler I, Skonetzki S, Gausepohl HJ, Linderkamp O, Wetter T. Evolution of the HELEN representation for managing clinical practice guidelines. Methods Inform Med, submitted for publication, <http://www.klinikum.uni-heidelberg.de/fileadmin/inst_med_biometrie/med_Informatik/helen3/evolution_helen_methods.pdf/> [accessed 30 June 2007].

[25] Hagerty CG, Pickens D, Kulikowski C, Sonnenberg F. HGML: a hypertext guideline markup language. Proc AMIA Symp 2000:325–9.

[26] Shiffman RN, Karras BT, Agrawal A, Chen R, Marenco L, Nath S. GEM: a proposal for a more comprehensive guideline document model using XML. J Am Med Inform Assoc 2000;7(5):488–98.

[27] Votruba P, Miksch S, Kosara R. Facilitating knowledge maintenance of clinical guidelines and protocols. Medinfo 2004;11(Pt. 1):57–61.

[28] Seyfang A, Miksch S, Marcos M, Wittenberg J, Polo-Conde C, Rosenbrand K. Bridging the gap between informal and formal guideline representations. In: 17th European conference on artificial intelligence (ECAI-06); 2006.

[29] Moskovitch R, Hessing A, Shahar Y. Vaidurya—a concept-based, context-sensitive search engine for clinical guidelines. Medinfo 2004;11(Pt. 1):140–4.

[30] Young O, Shahar Y, Lunenfeld E, Liel Y, Bar G, Shalom E, et al. Runtime application of hybrid-Asbru clinical guidelines. J Biomed Inform 2007;40:507–26.

[31] Brooke, J. SUS—a quick and dirty usability scale. Available from: <www.usabilitynet.org/trump/documents/Suschapt.doc/>; 1996.

[32] Shadbolt N, O'Hara K, Crow L. The experimental evaluation of knowledge acquisition techniques and methods: history, problems and new directions. Int Human–Computer Studies [Special issue on evaluation of KA techniques] 1999;51(4):729–55.

[33] Noy NF, Grosso W, Musen MA. Knowledge-acquisition interfaces for domain experts: an empirical evaluation of protege-2000. In: 12th international conference on software engineering and knowledge engineering (SEKE2000), Chicago, IL; 2000.

[34] Patel VL, Branch T, Wang D, Peleg M, Boxwala AA. Analysis of the process of encoding guidelines: an evaluation of GLIF3. Methods Inf Med 2002;41(2):105–13.

[35] SD, Shiffman RN. A preliminary evaluation of guideline content markup using GEM—an XML guideline elements model. In: Proc AMIA Symp; 2000, p. 413–17.

[36] Peleg M, Gutnik L, Snow V, Patel VL. Interpreting procedures from descriptive guidelines. J Biomed Inform 2006;39(2):184–95.

[37] Gershkovich P, Shiffman RN. An implementation framework for GEM encoded guidelines. In: Bakken S, editor. Proc AMIA Symp; 2001, p. 204–08.

[38] Miller RA, Waitman LR, Chen S, Rosenbloom ST. The anatomy of decision support during inpatient care provider order entry (CPOE): empirical observations from a decade of CPOE experience at Vanderbilt. J Biomed Inform 2005;38(6):469–85.

[39] Shalom E, Shahar Y, Lunenfeld E, Taieb-Maimon M, Young O, Bar G, et al. The importance of creating an ontology-specific consensus before a markup-based specification of clinical guidelines. In: 17th European conference on artificial intelligence (ECAI-06); 2006.

[40] Centers for Disease Control and Prevention (CDC) web site. Sexually transmitted diseases treatment guidelines. Available from: <http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5106a1.htm/>; 2002.

[41] Inpatient Management of COPD: Emergency Room and Hospital Ward Management (B1). Veterans Affairs (VA) web site. Available from: <http://www.oqp.med.va.gov/cpg/COPD/copd_cpg/content/b1/annoB1.htm/>; 2005.

[42] The American Association of Clinical Endocrinologists (AACE) Web site. American Association of Clinical Endocrinologists medical guidelines for clinical practice for the evaluation and treatment of hyperthyroidism and hypothyroidism. Available from: <http://www.aace.com/pub/pdf/guidelines/hypo_hyper.pdf/>; 2002.

[43] Siegel S, Castellan NJ. Nonparametric statistics. McGraw-Hill International Editions; 1988.

[44] Walpole RE, Myers HR. Probability and statistics for engineers and scientists. New York: Macmillan Publishing Co., Inc.; 1978.

[45] Shalom E, Shahar Y. A graphical framework for specification of clinical guidelines at multiple representation levels. In: Proc AMIA Annu Symp; 2005, p. 679–83.