4

Information Extraction Approach for Clinical Practice Guidelines Representation in a Medical Decision Support System

Fernando Pech-May[1], Ivan Lopez-Arevalo[2] and Victor J. Sosa-Sosa[2]
[1]*Instituto Tecnológico Superior de los Ríos*
Balancan, Tabasco
[2]*Information Technology Laboratory, Cinvestav – Tamaulipas*
Victoria, Tamaulipas
Mexico

1. Introduction

Errors in healthcare are a leading cause of death and injury. Kohn et al. (Kohn et al., 2000) mention that, for example, preventable adverse events are a leading cause of death in the United States. In their studies they state that at least 44,000 and perhaps as many as 98,000 americans die in hospitals each year as result of medical errors. Similar scenarios are for other countries. This situation has motivated the usage of Clinical Practice Guidelines (CPGs) to reduce the uncertainty of the clinical professional (nurses and physicians) when making decisions about the patient illness.

Clinical Practice Guidelines (CPGs) are documents containing guidelines and structured recommendations that are defined by domain experts based on medical and scientific evidence (Teije et al., 2006; Twaddle, 2005). Thus CPGs provide guides and scientific evidence to clinical professional to make flexible recommendations about specific health circumstances (Field & Lohr, 1990).

The main objective of CPGs is to offer to clinical staff a set of recommendations that are focused on helping in the diagnosis, prognosis, and treatment of specific illness. The goal is to enhance the medical attention to patients. Furthermore, a CPG is an important support for the patient itself and his/her family on understanding the efficiency of a treatment and an important tool to improve the quality in medical care. Because CPGs are largely documents in narrative form, sometimes are ambiguous and lack of a defined structure and internal consistency, which make them too complicated for being understood directly by a computer. The information usually contained in a CPG is plain texts, lists, diagrams, tables, and annotations in HTML, XHTML or PDF format. To make this information understandable for a computer, it is required the usage of CPGs formal representation languages (Clercq et al., 2004; Votruba et al., 2004). In this sense, many researchers have proposed different frameworks, approaches, representation languages, and tools for CPGs modelling, which can be interpreted by computers (Hripcsak et al., 2005; Isern & Moreno, 2008). Some of these tools and approaches provide orientation for a specific representation of CPGs, others are intended for a more general use in several representation languages.

However, nowadays the formalization process is still carried out manually. Although there exist several formal languages, their usage represents a complex and time-consuming work for a manual formalization of CPGs. The usage of tools for the formalization of CPGs requires not only knowledge about formal methods, but also about the medical domain.

This paper describes a basic Information Extraction (IE) approach to enhance the knowledge acquisition on Clinical Practice Guidelines. The aim is to support the CPG modeller during the formalization process facilitating the CPG interpretation by computers, becoming an important module in every Medical Decision Support System. The output of this approach can be used for a better understanding by non-clinical medical people of CPGs. The starting point of this work was motivated for a preliminary effort wherein a CPG interpreter was developed (Pech-May, 2010).

The paper is organised as follows. In section 2 the contextualisation background about Clinical Practice Guidelines and Information Extraction is given. Section 3 describes the proposed Information Extraction approach. Section 4 shows the experiments carried out and the obtained results of a first prototype for the proposed approach. Finally, the section 5 presents some conclusions, remarks, and further work.

2. Background

2.1 Clinical Practice Guidelines

For interpretation issues, the medical procedures within a CPG are translated into algorithms that describe such procedures for diagnosis, prognosis, and treatment. The representation of a CPG as algorithms allows the organization of the relevant information in a directly applicable manner. In consequence this representation can enhance and support the decisions making process (Patel et al., 2001; Lyng et al., 2008).

Several languages for CPGs representation have been developed for different purposes, users, and applications. Shiffman et al. (Shifman et al., 2000) described the requirements for modelling the knowledge of CPGs taking into account issues as completeness, expressivity, usability, and reuse.

Most of the representation languages use the XML format as readable-machine language. Some of the most used languages for representing CPGs are:

- **Asbru** (Young et al., 2007) is a task-specific and intention-based plan representation language. It was designed specifically for a set of management-task plans. Some tools that help the formalization of GPCs in Asbru are AsbruView[1] and DELT/A[2].
- **GLIF** (Wang et al., 2004) (the Guideline Interchange Format) defines an ontology for the representation of CPGs, as well as a medical ontology for representing medical data and concepts. GLIF (on its third version) includes a formal expression language for specifying decision criteria and patient state. A tool that allows modeling CPGs in GLIF is Protégé[3].
- **GEM** (Ciccarese et al., 2004) (the Guideline Elements Model) is an XML-based guideline document model that can store and organize the heterogeneous information contained in practice guideline documents. A tool that supports the formalization of CPGs in GEM is GEM Cutter[4].

[1] http://www.asgaard.tuwien.ac.at/asbruview/
[2] http://gem.med.yale.edu/default.htm
[3] http://protege.stanford.edu
[4] http://gem.med.yale.edu/default.htm

- **EON** (Tu et al., 2001) is a guideline modeling and execution system that is part of the EON architecture, a component-based suite of models and software components for the creation of guideline-based applications.
- **PROforma** (Sutton and Fox, 2003) allows the guideline to be modeled as a set of tasks and data items, it is designed to support the management of medical procedures and clinical decision making at the point of care. The PROforma task model divides a generic task (keystone) into four types: plans, decisions, actions, and enquiries.

The following list includes some remarkable aspects that are considered in the most typical formal languages such as Asbru, PROforma, and GLIF.

- **Organization of plans:** Asbru as well PROforma use an isolate generic object class for modelling plans: *the plan object*. GLIF uses two types of plans: *guides and macros*. The guides cover the direction and flow control decisions. The macros are used to specify in a declarative way the procedure patterns for specific purposes by means of a set of implementation steps, which appear as a one block in the CPGs.
- **Specification of goals/intentions:** GLIF specifies goals as strings; Asbru represents the intentions of plans as temporal patterns depending on the context.
- **Action model:** Actions are the primitives for the modeling used to represent tasks in a CPG (for instance, prescription, clinical research). All the languages allow specifying medical actions, but just GLIF has special structured classes to do it. This modeling method has an efficient mechanism to map instances of medical actions to terms of a restricted vocabulary. Regarding to the effect of actions, Asbru and PROforma, unlike GLIF, support express effects allowing to reason about actions based on its effects. In Asbru the effects of a plan can be used to select between different alternative plans and express causal relations. In turn in PROforma the effect of actions are modelled as postconditions, which are semantically different to the effects on Asbru because they represent assertions when an action is completed.
- **Representation of medical knowledge and patient data:** PROforma model medical knowledge by means of relations between concepts (indications, conindications, interaction between drugs, etc.). Such relations are included as arguments in alternative decisions. In GLIF the medical knowledge is represented as instances of concept-relation. Asbru has not a explicit representation for this kind of knowledge as part of the CPG model; nevertheless, this knowledge can be accessed by means of functions calls.

The use of any of the formal languages involves several additional tasks that depend on the particular language. To tackle this issue, several assistant tools have been developed to support the formalization process. They range from markup-based tools, such as DELT/A, Stepper, and GEM-Cutter, to graphical tools using symbols to model diagrams, such as Protégé or the plan body wizard of the DeGel framework. A brief description of these tools is given below.

- **Stepper** (Růzicka and Svatek, 2004) is a markup tool for the formalization of narrative CPGs. The formalization of the CPGs is done through user-defined stages and each stage transformed to XML.
- **GEM Cutter** (Karras et al., 2000) transforms CPGs into GEM format. The GEM Cutter interface shows the textual CPG and its XML representation, thereby facilitating the user interaction in the transformation of the CPG.
- **DELT/A** (Votruba et al., 2004) support the translation of HTML documents to XML. DELT/A provides two main features: (1) linking between a textual guideline and its formal representation, and (2) applying design patterns as macros forms.

- **Uruz** is part of the Degel framework (Shalom et al., 2003), it uses a markup mechanism that allows the user to introduce medical terms in the CPG. Such terms can come from some vocabulary as ICD-9-CM (ICD-9-CM, 2010)[5].
- **Protégé** (Gennari et al., 2002) is a general purpose tool for knowledge acquisition. It is broadly used in several knowledge domain fields. This tool allows modelling CPGs in different representation formal languages.
- **AsbruView** (Kosara et al., 2002) is a graphical user interface for Asbru to support the development of CPGs and medical protocols. AsbruView is focused on visualising data and plans during the design and execution.

According to the specialized literature, important work has been done trying to translate CPG documents into a readable-machine presentation (Pech-May, 2010). Dart et al. (Dart et al., 2001) proposed a generic model to represent any CPG in XML format. Moreover, they proved that CPGs can be modeled in a generic XML file. Bosse (Bosse, 2001) developed an interpreter capable of simulating CPGs written in Asbru language for one CPG. Geldof (Geldof, 2002) presented a methodology to formalize CPGs in several languages, from his understanding until computerizing in XML. Aguirre-Junco et al. (Aguirre-Junco, et al., 2004) described a knowledge specification method based on a structured and systematic analysis of text allowing a detailed specification of a decision tree for CPGs. Fuchsberger and Miksch (Fuchsberger & Miksch, 2002) presented an execution unit tailored for a particular CPG representation in Asbru plans.

The main aim for all the above work is the reasoning with the extracted medical knowledge. A reasoning process over such knowledge is desirable by nurses and physicians. Following this tendency, some efforts have been done to develop Medical Decision Support Systems (Kaiser & Miksch, 2005). But, like in similar works about Decision Support Systems, the knowledge acquisition becomes a bottleneck, which is the main limitation for this kind of systems. In this sense, we are introducing an approach to extract knowledge from textual CPGs that integrates an innovative Information Extraction module that facilitates knowledge acquisition.

2.2 Information Extraction

The Information Extraction (IE) is responsible for structuring information contained in plain texts, which can be relevant for a particular domain (called extraction domain) (Karras, et al., 2000; Lehnert et al., 1994). The IE is a research subject that covers many areas. The goal of an IE system is finding and linking relevant information while ignoring the strange and irrelevant information. Peshkin and Pfeffer (Peshkin & Pfeffer, 2003) define the Information Extraction as the task of filling template information from previously unseen text which belongs to a pre-defined domain.

One of the main reasons to use IE is its role in the evaluation and comparison of different Natural Language Processing technologies in domains highly influenced by human interactions, like the medical domain.

The IE systems can be classified based on two approaches:

[5] International Classification of Diseases. This is a classification of diseases and procedures used in the coding of clinical information derived from medical assistance, mainly in the hospital environment and specialized medical care centers.

- **Knowledge Engineering (KE):** This is focused on an empiric method or based on a domain corpus to develop efficient and robust Natural Language Processing systems (Kasabov 2006).
- **Machine Learning (ML):** This has a well-known set of documents and outputs and uses a set of patterns to extract knowledge by means of Machine Learning techniques (Ethem, 2004).

Based on the ML approach, the IE can be seen as useful technique to extract information from Clinical Practice Guidelines (CPGs) with the aim of enhancing its formalization. Particularly one of the tools used in the medical domain is the Badger system (Soderland, et al., 1995), which is a text analysis tool to summarize medical patient records by extracting diagnoses, symptoms, physical findings, test results, and therapeutic treatments based on linguistic concepts.

There are other IE systems based on Machine Learning techniques such as SRV -Sequence Rules with Validation- (Freitag, 1998), which transforms the patterns learning problem into a classification problem; RAPIER (Califf, 1998) that uses pairs of test documents and fills templates; and WHISK (Sonderland, 1999) that uses learning rules to extract a set of text styles. These tools can be adapted to several domains.

Some authors consider the Information Extraction (IE) as a later stage in the Information Retrieval (IR) process (Marie-Francine, 2006), the main difference between both is that IE provides the exactly desired information, while IR is in charge of finding the documents wherein the desired information should appear. Some new technologies try to merge advantages from both, such as some web wrappers (XWRAP (Liu et al, 2000) or (Baumgartner et al., 2001)) that extract information from HTML documents and search answers (automatic response over punctual queries). In this sense, a wrapper is a program that retrieves information from different repositories, merging, and unifying them. The aim of a wrapper is to locate relevant information in a semi-structured data and put it into a self-described representation for further processing (Kushmerick et al., 1997).

3. Approach

Most representation languages for CPG are very powerful and complex. They can contain many different types of information and data. The main goal for the application of Information Extraction on CPG documents is to obtain the relevant text by means of natural language patterns which can be used in the formalization of the CPG. This approach is illustrated in Figure 1.

The approach facilitates the formalization process by using several intermediate representations that are obtained by stepwise procedures. The idea is to obtain an intermediate representation of a CPG in XML format for reasoning. Such intermediate representation takes into account all the most important pieces from the CPG (such as actions, processes, sequences, etc.). The final output is a XML representation. This approach is an extension and adaptation of the work carried out by Cem Akkaya (Akkaya, 2005), which is a basic method for IE. The initial idea is to enhance the performance of a preliminary prototype to match patient data against CPGs (Pech-May et al., 2009) within a more general Medical Decision Support System.

To make the extraction, some specific templates have been generated, which are filled by the desired information. To detect such information, a heuristic method is applied. The filled templates are later processed.

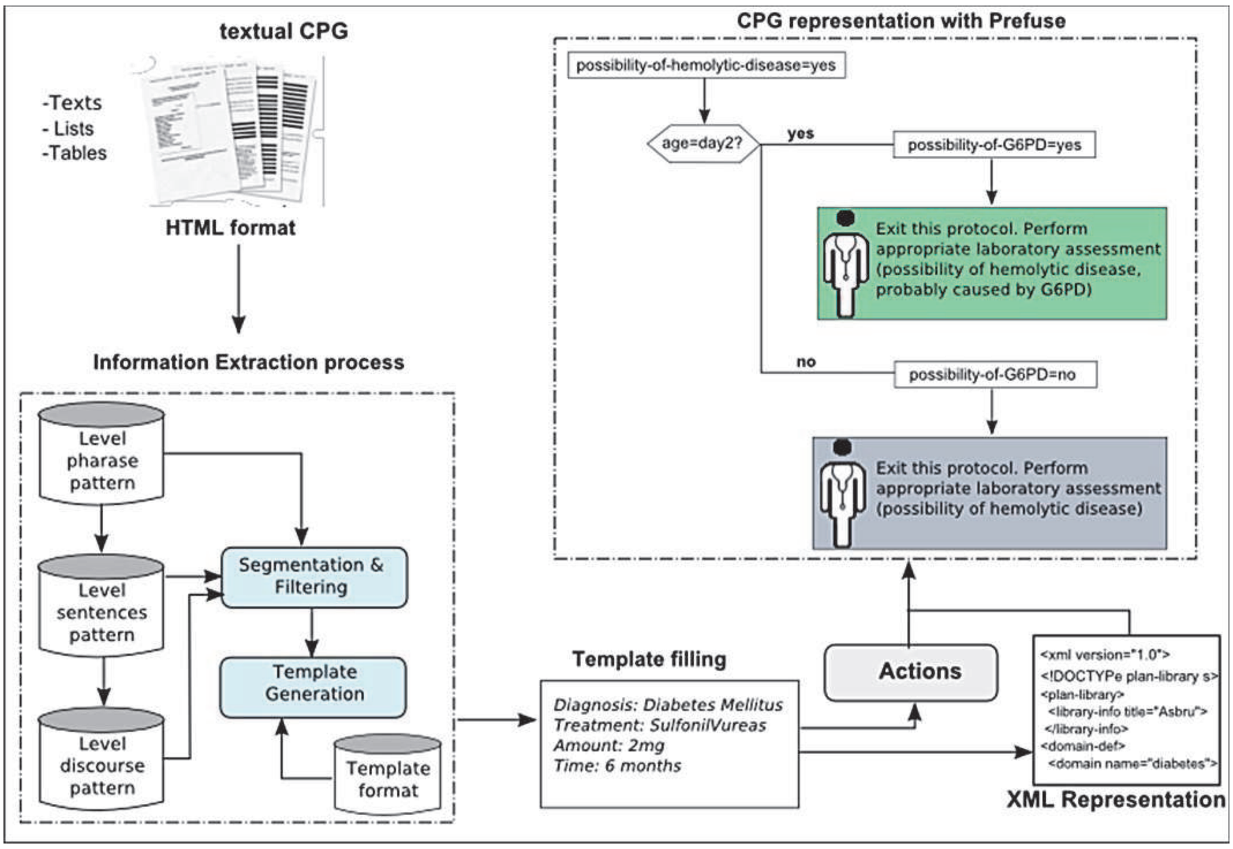
textual CPG
-Texts
- Lists
-Tables
HTML format
Information Extraction process
Level pharase pattern
Level sentences pattern
Level discourse pattern
Segmentation & Filtering
Template Generation
Template format
Template filling
Diagnosis: Diabetes Mellitus
Treatment: SulfonilVureas
Amount: 2mg
Time: 6 months
Actions
<xml version="1.0">
<!DOCTYPe plan-library s>
<plan-library>
<library-info title="Asbru">
</library-info>
<domain-def>
<domain name="diabetes">
XML Representation
CPG representation with Prefuse
possibility-of-hemolytic-disease=yes
age=day2?
yes
possibility-of-G6PD=yes
Exit this protocol. Perform appropriate laboratory assessment (possibility of hemolytic disease, probably caused by G6PD)
no
possibility-of-G6PD=no
Exit this protocol. Perform appropriate laboratory assessment (possibility of hemolytic disease)

Fig. 1. Proposed IE approach.

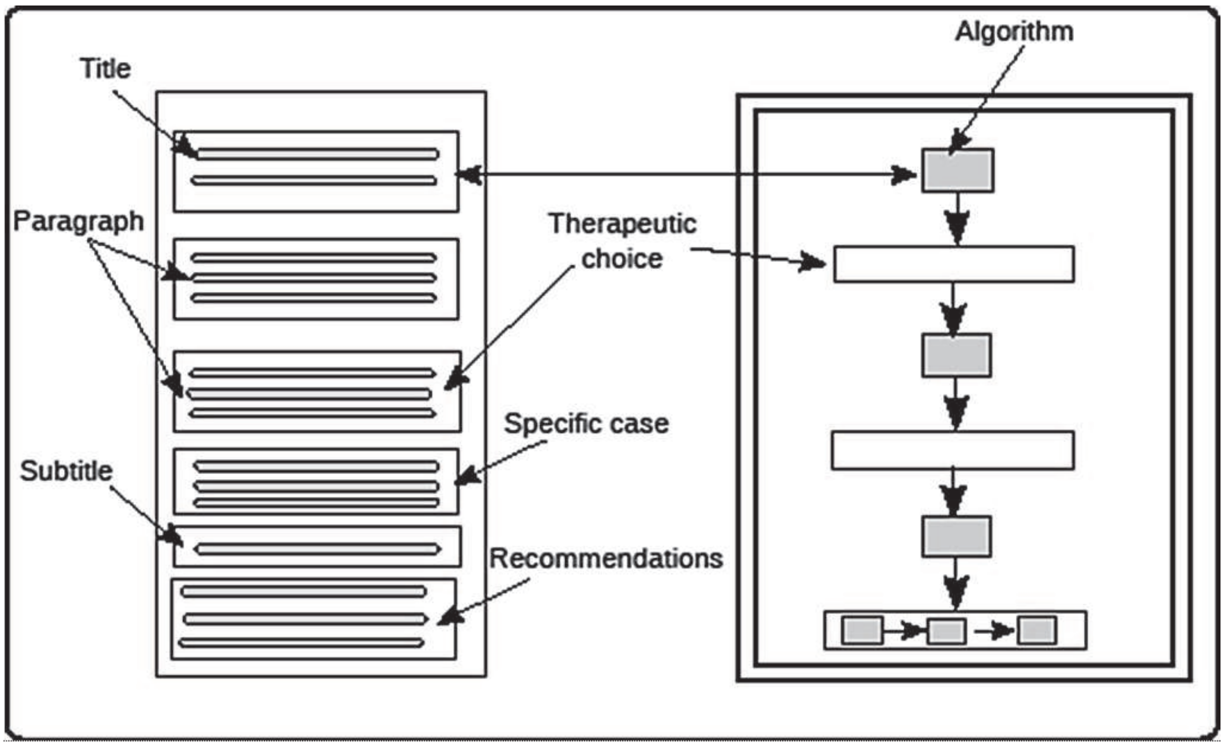

Fig. 2. General structure for CPGs.

The input CPGs are chosen from the National Guideline Clearinghouse[6] (NGC) repository in XHTML format. Then, XHTML documents are analyzed to extract relevant information, and subsequently to obtain the intermediate representation. Such representation is displayed through templates in form of views by means of the Prefuse[7] tool (Jeffrey et al., 2005). The approach considers that tested CPGs follow the structure of the NGC repository since these CPGs have a predefined structure. This approach works only for textual CPGs because in graphical or chart representation few text is included, relevant text (from medical experts) is mandatory for the approach. In general, a GPC has a structure consisting of separated sections for the treatment of a disease. For example, Figure 2 shows the general sections for diagnosis and treatment of a disease.

A general flowchart for a CPG is depicted in Figure 3, based on the general structure of a CPG.

In order to obtain information from CPGs, this approach is based on the transformation of multiple processes following three heuristic patterns for Information Extraction:

a. Phrase pattern level (lexical level)
b. Sentence pattern level (syntactic level)
c. Speech pattern level (semantic level)

It is necessary a parsing process to obtain the extraction rules. This is based on a knowledge engineering approach considering syntactic and semantic restrictions, and taking into account delimiters.

In order for processing a large amount of documents and information, it is necessary specific heuristics for each type of information required, for example:

a. Different types of information, in which each type of information needs specific methods for its processing (e.g. processes, parameters).
b. Different representations of information, in which it should be taken into account that the information could be represented in different ways (structured, semi-structured, or plain text).
c. Different types of guidelines, in which there may be CPGs for different diseases, diverse user groups, and several organizations that may contain similar CPGs.

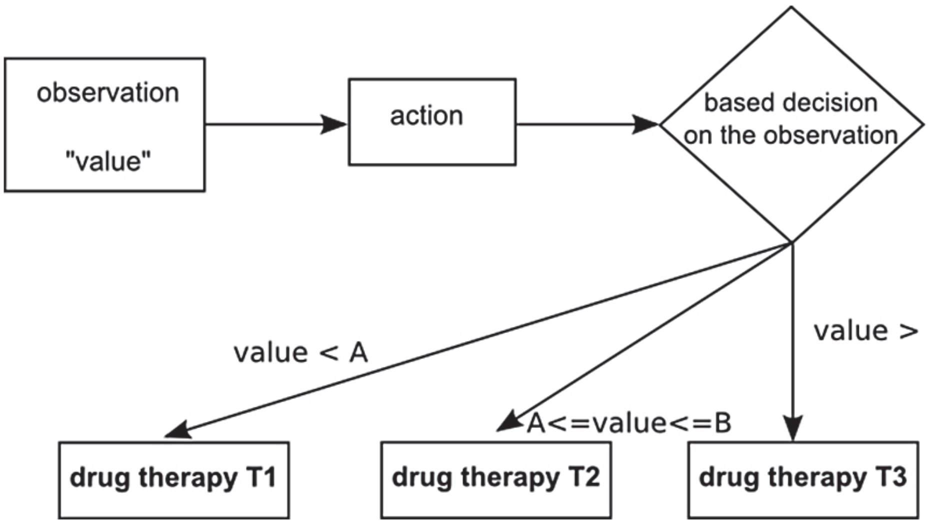

Fig. 3. General flowchart for CPGs

[6] www.guideline.gov

[7] The Prefuse toolkit is a set of software tools for creating interactive data visualizations for Java language.

The core of the approach is based on the atomic approach (Appelt & Israel, 1999), which basic idea is to assume that every noun phrase and verb of the right type, independently of the syntactic relations obtained among them, indicate an event/relationship of interest. It does not take into account the accuracy of the data extraction. Subsequently to the extracted data, a segmentation and filtering process is performed for its depuration. In this way, only the data concerning to the information of interest (diagnosis, treatment, drugs, etc.) is obtained. These data is stored in specific templates for further processing. The medical terms used in our prototype come from the Medical Subject Headings (MeSH)[8] of the National Library of Medicine from United States. Next, each heuristic pattern is briefly described.

3.1 Phrase pattern level

In this stage a lexical parser is used. It has the responsibility of splitting the text in paragraphs, tokens, and identifies important phrases in the CPG (e.g. administration of a drug, surgical procedures, dose of a drug, etc.). The lexical analyzer function is to identify the relevant information and then extract important data from the CPG. The lexical analyzer is in charge of filtering, for the second level of IE (sentence pattern level), the information that can be used by the syntactic level. They are defined by regular expressions as:

- Action terms (mainly verbs; e.g., “activate”, “perform”, “prescribe”, “treat”, “integrate”, “receive”, etc).
- Condition terms (regular expressions describing a condition, such as “if [, : \.]+”, “in case(s)? [, : \.]+”, “if [_,2 weeks]+”, etc.).
- Time Annotations (e.g. [ESS, LSS], [EFS, LFS], [MinDu, MaxDu], REFERENCE)
- Dose unit terms (e.g. “(m|d|c)?(l|g)(/kg/day)?”, “drop(s)?”, “teaspoon(s)?”, “tsp”).

3.2 Sentence pattern level

In this level (syntactic level), the entire document is parsed and split into sentences. Then every sentence is processed with regard to its context within the document and its group affiliation. Thereby, the context is obtained by captions (e.g. “Acute Pharyngitis in children Algorithm Annotations | Treatment | Recommendations:”) and a group contains sentences from the same paragraph or the same list, if there are no sublists. Thus, each sentence is now checked for relevance. Useful medical terms and keywords to identify medical actions can be found. The words or groups of words are mainly verbs indicating the application of a therapy, administration of a drug or a surgical procedure. This level considers two groups of patterns:

- **Free text pattern**. It is used to identify paragraphs from a list of items. The pattern indicates therapy instruments (surgical procedures) combined with key terms (e.g. prescribe, indicate, execute, etc).
- **Concise text pattern**. It is used to detect specific defined patterns such as lists of items with incorrect grammar. In general, it denotes the right therapy to apply, instruments for the therapy or drugs. These can be merged with other detected labels in the sentence or phrase pattern level.

Detecting relevant sentences is a challenging task, which is undertaken in two steps:

1. detecting irrelevant sentences to exclude them from further processing, and
2. detecting relevant sentences.

[8] http://www.nlm.nih.gov/mesh/

In both steps, special keywords are used to detect whether a sentence is irrelevant or relevant. Keywords describing irrelevant sentences are “history”, “diagnosis”, “criteria”, “symptom”, “clinical assessment”, “risk factor”, “complicating factor”, “etiology”, and so on. These terms point out that the following paragraph does not describe treatment processes, but that it describes symptoms, demonstration of diagnoses, and so on. If such a term appears within a caption the corresponding section is removed.

3.3 Speech pattern level

In this level semantic aspects are solved and the design and the structure of the final document XML are improved. In addition, this level is used to categorize sentences, actions, and to find their relationships. To accomplish the later task the following processes in the CPG are identified:

a. processes with temporal dependencies (processes at some point depending on another process),
b. sequential processes (the processes that are required to run with the authorization of other),
c. processes containing a thread,
d. selection process, and
e. recurring processes.

The application of the extraction rules gives as a result a well-structured XML document which can be represented by using specific templates or graphical forms (by using the Prefuse tool):

- **Templates**. It is the final representation of the Information Extraction module; these can be filled with specific CPG data. After collected all the relevant phrases from the CPG, the document is generated by using the representation of sentences, actions, relationships, and hierarchical structure. The representation is done through a document markup listing all relevant sentences and an identification of MeSH terms. Thus the document contains information for a dose, duration, actions, administration of a disease, etc.
- **Graphical representation**. The generated document is represented visually using Prefuse. In this way, it provides data to optimize table structures, tree graph design, visual encoding techniques, dynamic queries, integrated search, and database connectivity.

4. Experiments

A first implementation of this approach was developed by using Java language. For a performance analysis, different CPGs corresponding to different specific diseases were employed, which are:

- Diagnosis and treatment of otitis media in children
- Diagnosis and treatment of diabetes-mellitus (type 2)
- Acute pharyngitis in children
- Diagnostics and treatment of jaundice
- Management and treatment of dengue hemorrhagic fever at first and second attention level
- Treatment of breast cancer
- Chronic cough in a child

The CPGs were divided into two groups:

1. Clinical guidelines to develop and improve the heuristics
2. Clinical guidelines to test the obtained heuristics

The choice of these groups is not a trivial task because the organizations that develop CPGs do not regularly take care of following the same hierarchical structure. In this experiment, complex hierarchical structures were used as selection criteria, and distributed evenly to each group. Before applying the heuristics, some pre treatment was carried out (to verify that XHTML documents satisfy the structuring elements). This is achieved through the conversion of paragraphs/sections from the CPG and their corresponding items (according to the three pattern levels). Our test considered the following two tasks:

- **Task 1**: detection of relevant sentences, and
- **Task 2**: summarization of the detection types of sentence and the relationship between processes.

The performance of the prototype was evaluated by using the precision and recall measures. The recall score measures the ratio of correct information extracted from the texts against all the available information present in the text. The precision score measures the ratio of correct information that was extracted against all the information that was extracted (Lehnert et al., 1994). The following summarizes the obtained results by task:

- **Task 1:** It obtains promising results (Table 1), even if it means lowering the precision punctuation. The lower recall score implies that detecting relevant sentences has to be improved. The high accuracy on precision score shows that irrelevant sentences were classified as relevant.
- **Task 2**: The entry for task 2 (Table 2) consists of sentences identified with very high punctuation in the previous task. The recall score is very high, which means that only few sentences were falsely not detected. The precision score implies that some slots were filled out incorrectly. The reason for this is that they do not always detect the correct type of sentence and specially when assigning annotations to their particular actions, situation that has to be improved.

Table 3 presents an overall evaluation. For all the tables, the nomenclature for columns is:

COR –Number of correct slots that were identified by our IE system
MAT –Total number of slots that match a CPG template in the CPGs group
IDE –Total number of slots that were identified by our IE system
REC: Represents our system recall that is given by COR/IDE
PRE: Represents our system precision that is given by COR/MAT

At the phrase pattern level several regular expressions were necessary. Figure 4 shows a fragment for a pattern in this level. At the sentence pattern level the text free patterns were identified, such as <p> and </p> (to identify de paragraphs), <li> and </li> (to identify lists of items), and some additional labels. These labels are combined with the labels from de phrase pattern level like <dosage> or </dosage>, <dose> or </dose>, etc.

After each CPG was processed in the three analysis stages (phrase level, sentence level, and speech level), an intermediate representation was obtained. For this, two files were generated, the first one containing the list of relevant sentences (see Table 5) and a second one which is a mark-up document (see Table 6).

The intermediate representation shown in Table 6 contains a set of actions and relations. An action contains sentences describing the action and annotation assigned by means of the DELT/A tool. It also contains the instrument for the treatment and an identifier within the MeSH dictionary. If the information is about a dose, duration of treatment or drug management, then a corresponding MeSH identifier is assigned to it. Table 7 partially shows the actions and its assigned MeSH identifiers for the CPG "Diagnosis and treatment of otitis media in children".

CPG	COR	MAT	IDE	REC	PRE
Diagnosis and treatment of otitis media in children	23	24	26	0.958	0.884
Diagnosis and treatment of diabetes-mellitus (type 2)	45	57	53	0.789	0.849
Acute pharyngitis in children	9	12	10	0.75	0.9
Diagnostics and treatment of jaundice	53	56	53	0.946	1
Management and treatment of dengue hemorrhagic fever at first and second attention level	65	68	75	0.955	0.866
Treatment of breast cancer	56	59	74	0.946	0.756
Chronic cough in a child	23	31	26	0.741	0.884

Table 1. Evaluation of Task 1 for each CPG.

CPG	COR	MAT	IDE	REC	PRE
Diagnosis and treatment of otitis media in children	27	32	27	0.843	1
Diagnosis and treatment of diabetes-mellitus (type 2)	32	34	40	0.941	0.8
Acute pharyngitis in children	56	58	67	0.965	0.835
Diagnostics and treatment of jaundice	36	42	45	0.857	0.8
Management and treatment of dengue hemorrhagic fever at first and second attention level	73	75	75	0.973	0.973
Treatment of breast cancer	86	116	98	0.741	0.877
Chronic cough in a child	14	18	20	0.777	0.7

Table 2. Evaluation of Task 2 for each CPG.

Task	COR	MAT	IDE	REC	PRE
Task 1	355	443	432	0.801	0.821
Task 2	849	865	978	0.981	0.868

Table 3. Overall evaluation results

<table>
<tr><td><number></td><td>([\d]+(([\.]([\d]+)) | ((\s*[\d]+)?/[\d]+))?)</td></tr>
<tr><td><numberOrRange></td><td><number>(((_to_) | (\s*-\s*))<number>)?</td></tr>
<tr><td><time-unit></td><td>m(illi)?)?sec(ond)?(s)? | min(ute)?(s)? | hour(s)? |
day(s)? | week(s)? ...</td></tr>
<tr><td><dose-unit></td><td>(m | c | d)?(l | g)(/kg(/<time-unit>)?)? | drop(s)? | tab(s)? ...</td></tr>
<tr><td><dosage></td><td><numberOrRange>[\s]*<dose-unit></td></tr>
<tr><td><time></td><td><numberOrRange>[\s]*<time-unit></td></tr>
<tr><td><iteration></td><td>TID | BID | QD | (Q | every) <time> |
<numberOrRange> _(times | doses)_(per | a)_<time-unit></td></tr>
<tr><td><person></td><td>those | patient(s)? | person(s)? | child(ren)? ...</td></tr>
<tr><td><condition></td><td>(in_(case(s)? | areas) | if | unless | who(m)?)_[ˆ,:]+ |
In_.*allergic [ˆ,\.:]+ | (for | in)_(a_)?(<person>) [ˆ,\.:]+</td></tr>
</table>

Table 4. Examples of phrase level patterns.

```
<sentence>
  <delta-link link-id="8"/>
  <description>In children with risk factors for Streptococcus pneumoniae,
           it is recommended that Amoxicillin, high dose (80 to 90
           mg/kg/day) or Augmenting (with high dose amoxicillin component)
           be utilized as first-line therapy (Nash and Wald, 2001 [S];
           Wald, Chiponis, and Ledesma-Medina, 1986 [B]; Nelson, Mason,
           and Kaplan, 1994 [C]; Dowell et al., 1999 [E]; Dowell, 1-1998
           [E]; Friedland and McCracken, 1994 [E]; Local Expert Consensus
           [E]).
  </description>
</sentence>
<sentence>
  <delta-link link-id="9"/>
  <description>Note: Failure with amoxicillin is likely to be due to resistant
                Streptococcus pneumoniae, Haemophilus influenzae, or Moraxella
                catarrhalis.
  </description>
</sentence>
<sentence>
  <delta-link link-id="10"/>
  <description>High dose amoxicillin will overcome Streptococcus pneumoniae
                resistance (changes in penicillin-binding proteins)
                (Dowell et al., 1999 [E]; Whitney et al., 2000 [D]).
  </description>
</sentence>
```

Table 5. Fragment of the relevant sentence file corresponding to the GPC "Diagnosis and treatment of otitis media in children".

```
<li>
  <a id="delta:8">In children with risk factors for  Streptococcus
             pneumoniae, it is recommended that Amoxicillin, high dose
             (80 to 90  mg/kg/day) or Augmenting (with high dose
             amoxicillin component) be utilized as first-line therapy
             (Nash and Wald, 2001 [S]; Wald, Chiponis, and Ledesma-
             Medina, 1986 [B]; Nelson, Mason, and Kaplan, 1994 [C];
             Dowell et al., 1999 [E]; Dowell, 1 -1998 [E]; Friedland
             and McCracken, 1994 [E]; Local Expert Consensus [E]).
  </a>
  <ul type="disc">
    <li>
      <a id="delta:9">Note: Failure with amoxicillin is likely to be due
             to resistant <Streptococcus pneumoniae, Haemophilus
             influenzae, or Moraxella catarrhalis.
      </a>
      <a id="delta:10">High dose amoxicillin will overcome Streptococcus
              pneumoniae resistance (changes in penicillin-
              binding proteins) (Dowell et al., 1999 [E]; Whitney
              et al., 2000 [D]).
      </a>
      The clavulanic acid component of Augmentin is active against
      Resistant Haemophilus influenzae and Moraxella catarrhalis (B-
      lactamase enzyme) (Wald, Chiponis, and Ledesma-Medina, 1986 [B];
      Dagan et al., 2000 [A]).
    </li>
  </ul>
</li>
```

Table 6. Fragment of the mark-up document file corresponding to the GPC "Diagnosis and treatment of otitis media in children".

With the obtained actions from a CPG, it is possible transform the CPG into an Asbru document. At this moment this step is carried out manually.

In Asbru, a plan is represented by means of plans definitions. A plan contains a *plan name, arguments, knowledge role,* and a *plan body.* Table 8 shows an example for a fictitious plan following the Asbru specification. In Table 9 can be seen fragment of sentences, actions and plans for the CPG *Diagnosis and treatment of otitis media in children* in Asbru.

```
<action id="8" parent="5" group="18" selection="0">
  <delta-link link-id="8"/>
  <description>In the child with no risk factors for penicillin-resistant Streptococcus
                pneumoniae standard dose amoxicillin or Augmentin (with standard
                dose Amoxicillin component) may be considered as initial therapy.
  </description>
  <agents>
```

```
    <agent MeSH="D000658" name="amoxicillin"/>
    <agent MeSH="D019980" name="Augmentin"/>
  </agents>
  <condition>
    <item>In the child with no risk factors for penicillin-resistant Streptococcus
            pneumoniae
    </item>
  </condition>
  <annotations>
    <annotation>Note: Forty-six percent of isolates at Children’s Hospital Medical
                   Center of Cincinnati, Ohio have intermediate or high
                    Penicillin-resistant Streptococcus pneumoniae and local data
                   supports that 15% of children locally may fail initial therapy
                   with standard dose amoxicillin.
       <delta-link link-id="9"/>
    </annotation>
  </annotations>
  <context>
    <item>Antibiotic Treatment</item>
  </context>
</action>
```

Table 7. Partial actions corresponding to the GPC “Diagnosis and treatment of otitis media in children”

Plan	Plan-1
TIME ANNOTATION	([,], [,24 hours], [,], *NOW*)
PREFERENCES	Select-method: exact-fit
INTENTIONS	Avoid intermediate state: (glucose-level = high)
CONDITIONS	Abort-condition: (glucose-level = high) Filter-condition: ((patient-age > 60) AND (patient-age < 80))
EFFECTS	Plan-effect: Parameter="glucose-level" Relationship="decrement" Likelihood 0.65
PLAN_BODY	Parallel subplans: Continuation spaci_cation: (treatment-1 OR treatment-2) Diagnosis Treatment-1 Treatment-2

Table 8. Example of a fictitious plan in Asbru

<**treatment** title="Diagnosis and treatment	<**treatment** title="Diagnosis and treatment
of otitis media in children.">	of otitis media in children.">
<**sentences**>	<**actions**>
<**sentence**>	<**action group**="3" **id**="1" **parent**="0">
…	…
<**sentence**>	</**action**>
<**sentence**>	<**action group**="9" **id**="14" **parent**="12">
<**delta-link link-id**="14"/>	<**delta-link link-id**="14"/>
<**description**>Therapeutic (10 day)	<**description**>Therapeutic (10 day)
course of antibiotics.</**description**>	course of
</**sentence**>	antibiotics.</**description**>
<**sentence**>	<**agents**>
<**delta-link link-id**="15"/>	<**agent MeSH**="D000900"
<**description**>Consideration may be	**name**="antibiotic">
given to **a** shortened course of antibiotics (5	<**duration** term="10 day"/>
days) for children who are at low risk (i.e.,	</**agent**>
age > 2 years, no history of chronic or	<**agent MeSH**="D000900"
recurrent otitis media and intact tympanic	**name**="antibiotic">
membranes).</**description**>	<**duration** term="5 days"/>
</**sentence**>	</**agent**>
<**sentence**>	</**agents**>
<**delta-link link-id**="16"/>	<**annotations**>
<**description**>First-Line	<**annotation**>Consideration may be
Medications</**description**>	given to **a** shortened course of
</**sentence**>	antibiotics (5 days) for children
<**sentence**>	who are at low risk
<**delta-link link-id**="17"/>	(i.e., age & > 2 years, no history
<**description**>amoxicillin (40	of
mg/kg/day) if	chronic or recurrent otitis media
low risk (> 2 years, no day care, and	and intact tympanic
no antibiotics for the past three	membranes).
months).</**description**>	<**delta-link link-id**="15"/>
</**sentence**>	</**annotation**>
<**sentence**>	<**annotation**>The use of nasal
<**delta-link link-id**="18"/>	decongestants and corticosteroids
<**description**>80 mg/kg/day if not	is not supported in the literature.
low risk or for resistant AOM	<**delta-link link-id**="34"/>
if the lower dose	</**annotation**>
was used initially .</**description**>	</**annotations**>
</**sentence**>	<**context**>
<**sentence**>	…
…	</**context**>
<**sentences**>	</**action**>
a) SENTENCES	**b) ACTIONS**

Table 9. Fragments for the GPC ″ Diagnosis and treatment of otitis media in children"; a) fragment of sentences, b) fragment of extracted actions; *continue in the Table 10.*

```
<plan name="PLAN_PARENT_2"
     title="Therapeutic (10 day) course of antibiotics.">
    <conditions>
      <setup-precondition confirmation- required="yes">  <none/>
      </setup-precondition>
    </conditions>
    <plan-body>
       <plan-activation>   <plan-schema name="PLAN_PARENT_1">
                               <delta-link link-id="1"/>
                           </plan-schema>
       </plan-activation>
     </plan-body>
</plan>
<plan name="PLAN_14" title="Therapeutic (10 day) course of antibiotics.">
  <delta-link link-id="14"/>    <delta-link link-id="15"/>
  <delta-link link-id="34"/>
  <explanation text="Consideration may be given to a shortened  course of antibiotics (5 days) for
children who are at low risk (i.e., age & > 2 years,   no history of chronic or recurrent otitis media and
intact tympanic membranes). The use of nasal  decongestants and corticosteroids is not supported in
the literature."/>
  <conditions/>
     <plan-body>
       <subplans type="unordered">  <wait-for>  <all/>  </wait-for>
         <plan-activation>  <plan-schema name="PLAN_16">
                                 <delta-link link-id="16"/>
                            </plan-schema>
        </plan-activation>
        <plan-activation>
          …
        <plan-activation>
    …
</plan>
```

c) ASBRU PLANS

Table 10. Continuation from Table 9, c) fragment from the transformation of actions to the Asbru format.

The above actions can be seen graphically as a tree graph by using the Prefuse tool. This view enhance the support to the clinical staff about identifying, in a easy way, what are the symptoms in the patient to decide a dose for a drug or the right therapy. Figure 4 shows a small fragment for a visual plan of the CPG *Diagnosis and treatment of otitis media in children*.

5. Conclusions

This paper describes a basic Information Extracting approach applied to obtain knowledge from Clinical Practice Guidelines. The final objective of this work is to obtain an intermediate representation of actions from a textual CPG in XML format by means of an

Information Extraction module. The approach applies three heuristics using specific expression patterns over the structure of CPG documents. Through the application of generic Information Extraction heuristic rules, a single formatted document is obtained, which contain the lists, sub-lists, and paragraphs from the original CPG. This document is an intermediate knowledge representation in XML format. The result of the extracted information is used to fill individual slots templates, which represent processes and their relationships in a CPG document. It can be translated into two formal representations: 1) Asbru language (although other languages can be used) and 2) A graph representation by using the Prefuse tool. The aim of the second option is to show the hierarchical structure of the CPG; thus a physician can see in a graphical view the symptoms and routes on the CPG where the patient can be directed for an action or therapy.

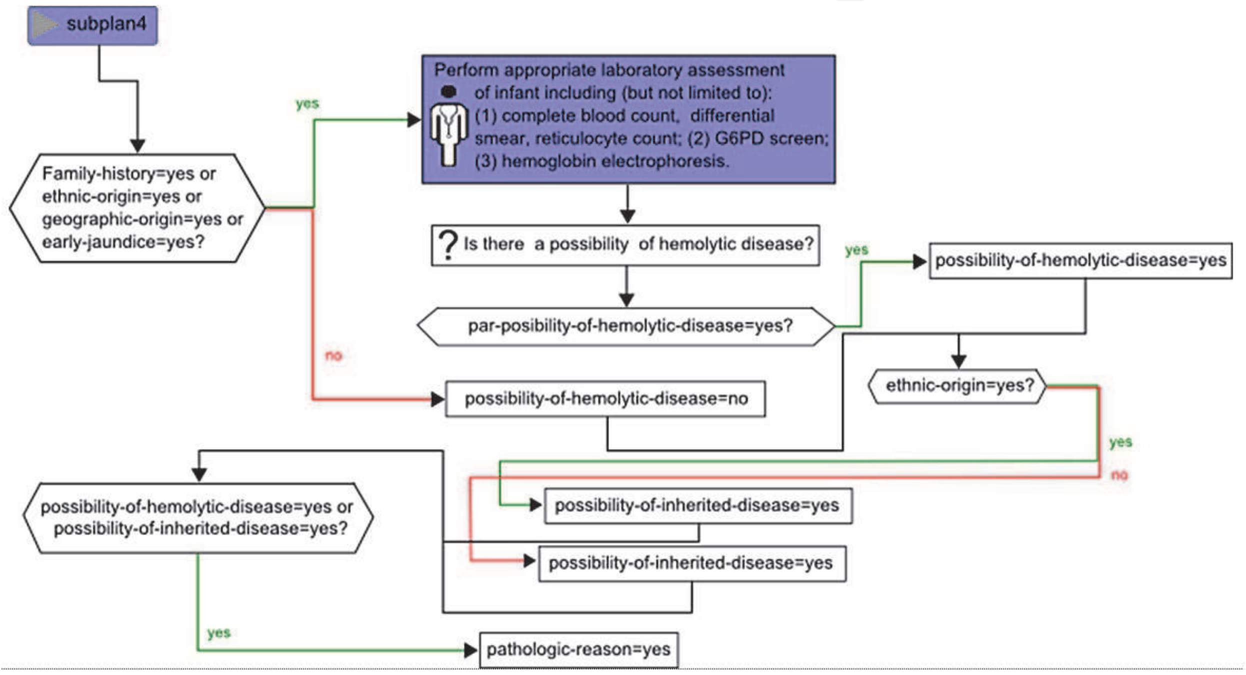
subplan4
Family-history=yes or ethnic-origin=yes or geographic-origin=yes or early-jaundice=yes?
yes
Perform appropriate laboratory assessment of infant including (but not limited to): (1) complete blood count, differential smear, reticulocyte count; (2) G6PD screen; (3) hemoglobin electrophoresis.
? Is there a possibility of hemolytic disease?
yes
possibility-of-hemolytic-disease=yes
par-posibility-of-hemolytic-disease=yes?
no
possibility-of-hemolytic-disease=no
ethnic-origin=yes?
yes
no
possibility-of-inherited-disease=yes
possibility-of-hemolytic-disease=yes or possibility-of-inherited-disease=yes?
possibility-of-inherited-disease=yes
yes
pathologic-reason=yes

Fig. 4. Example of a visual plan for the CPG Diagnosis and treatment of otitis media in children.

To obtain the actions for a CPG, three stages are necessary for phrases (lexical), sentences (syntactic) and semantic. The first stage is the phrase pattern level wherein a CPG document is lexically analysed; the document is tokenised, relevant phrases are identified and important data is detected. This level filters, to the sentence pattern level, only information used within the syntactic level. The second stage is basically a syntactic analyser, it uses the relevant phrases or identified tokens in the lexical level. At this level medical terms and keywords, to identify medical actions, are identified. The set of terms consist mainly of verbs denoting the application of a therapy, administration of drugs or surgical procedure. This level is divided in two groups of patterns: text free pattern and concise text pattern. The third stage is the speech pattern level where the design and structure of the document is enhanced. It categorises sentences and finds their relations. The approach has been implemented in a first prototype. The experiments show that the proposed heuristic-based approach can achieve good results, especially for CPG with a major portion of semi-structured text. The obtained intermediate representation may be used in a next stage for a better formalisation of the CPG.

As a future work the rules for processing CPGs containing complex information will be improved. Another goal is to create a support model with the ability for evaluating plans that are contained in CPGs.

6. Acknowledges

This research was partially funded by project number 153880 from “Fondo Mixto Conacyt-Gobierno del Estado de Tamaulipas” and project number TAB-2010-C19-144199 from "Fondo Mixto Conacyt-Gobierno del Estado de Tabasco".

7. References

Aguirre-Junco, A.R., Colombet I., Zunino, S., Jaulent, M.C., Leneveut, L. & Chatellier G. (2004). Computerization of guidelines: A knowledge specification method to convert text to detailed decision tree for electronic implementation, *Stud Health Technol Inform*.107(1):115-9.

Akkaya, C. (2005). Extracting process information from clinical practice guidelines. Master's thesis, Vienna University of Technology.

Appelt, D. E. and Israel D. J. (1999). Introduction to information extraction technology, A tutorial prepared for IJCAI-99, Stockholm,scheweden.

Baumgartner, R., Flesca, S. & Gottlob G. (2001). VisualWeb Information Extraction with Lixto, *Proceedings of the 27th International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 119-128.

Bosse, T. (2001). An interpreter for clinical guidelines in asbru. Master's thesis, Department of Artificial Intelligence, Faculty of Sciences. Vrije Universiteit Amsterdam. Amsterdam, The Netherlands.

Califf, M. (1998). Relational Learning Techniques for Natural Language Information extraction. Ph.D. thesis, Department of Computer Sciences, University of Texas, Austin, TX, USA

Ciccarese, P., Caffi, E., Boiocchi, L., Quaglini, S. & Stefanelli, M. (2004). A guideline management system, *Proceedings of 11th World Congress of the International Medical Informatics Association*, IOS Press, San Francisco, USA, pp. 28–32.

CIE-9-CM (2010) International Classification of Diseases, Ninth Revision, Clinical Modification. National Center for Health Statistics of USA. USA.

Clercq, P. A., Blom, J.A., Korsten, H.H. & Hasman, A. (2004). Approaches for creating computer-interpretable guidelines that facilitate decision support, *Artificial Intelligence in Medicine,* 31(1):1–27.

Dart, T., Xu, Y., Chatellier, G. & Degoulet, P. (2001). Computerization of guidelines: Towards a “guideline markup language”, pp 186-190.

URL: http://www.metapress.com/content/mhd4p271n0653xx2

Ethem, A. (2004). *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press.

Field, M. J. & Lohr, K. (1992). *Guidelines for Clinical Practice: From development to use, Institute of Medicine*, National Academy Press.

Freitag, D. (1998). Machine Learning for Information Extraction in Informal Domains. Ph.D. thesis, Computer Science Department, Carnegie Mellon University. Pittsburgh, PA, USA

Fuchsberger, C. & Miksch, S. (2002). Asbru's execution engine: Utilizing guidelines for artificial ventilation of newborn infants. *Technical report*, Vienna University of Technology, Institute of Software Technology and Interactive Systems.

Geldof, M. (2002). The formalization of medical protocols: easier said than done. Master's thesis, Department of Artificial Intelligence, Faculty of Sciences. Vrije Universiteit Amsterdam. Amsterdam, The Netherlands.

Gennari, J. H., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubézy, M., Eriksson, H., Noy, N. F. & Samson, W.T. (2002). The evolution of protégé: An environment for knowledge-based systems development, *International Journal of Human Computer Studies*, 58(1):89-123.

Hripcsak, G., Clayton, P. B., Pryor, T. A., Haug, P. & Wigertz, O. B. (2005). The arden syntax for medical logic modules, *International Journal of Clinical Monitoring and Computing,* 10(4): 215–224.

Isern, D. & Moreno, A. (2008). Computer-based execution of clinical guidelines: A review, *International Journal Medical Informatics*, 77(12):787-808.

Jeffrey, H., Stuart, K. C. & James, A. L. (2005). Prefuse: a toolkit for interactive information visualization, *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, Portland, Oregon, USA, pp. 421-430.

Kaiser, K. & Miksch, S. (2005). Modeling computer-supported clinical guidelines and protocols. Technical report, Vienna University of Technology, Institute of Software Technology and Interactive Systems, Vienna.

Karras, B., Deshpande, A., Polvani, K., Agrawal, A. & Shiffman, R.N. (2000). *Gem cutter manual*. Yale Center for Medical Informatics.

Kasabov, N. (2006). *Evolving Connectionist Systems: The Knowledge Engineering Approach*, Springer-Verlag New York, Inc.

Kohn, L. T., Corrigan, J. M. & Molla, S. (2000). *To Err Is Human: Building a Safer Health System*, National Academy Press, Washington, D.C.

Kosara, R., Miksch, S., Seyfang, A. & Votruba. P. (2002). Tools for acquiring clinical guidelines in asbru, *Proceedings of the 6th World Conference on Integrate Design and Process Technology (IDPT'02)*, Society for Design and Process Science, New York, pp. 22-27.

Kushmerick, N. (1997). Wrapper induction for information extraction, Ph.D. thesis, Department of Computer Science and Engineering, University of Washington.

Lehnert, W. & Cowie, J. (1994). Evaluating an information extraction system, *Commun.* ACM, 39(1):80-91.

Liu, L., Pu, C. & Han, W. (2000). XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources, *Proceedings 16th International Conference on Data Engineering (ICDE),* IEEE Computer Society, Washington, DC, USA, pp. 611–621.

Lyng, K. M., Hildebrandt, T. & Mukkamala, R. R. (2008). From paper based clinical practice guidelines to declarative workflow management, *Business Process Management Workshops*, Springer, Milano, Italy, pp. 336-347.

Marie-Francine, M. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context (The Information Retrieval Series)*. Springer-Verlag.

Patward, S. & Riloff E. (2006). Learning Domain-Specific Information Extraction Patterns from the Web, *Proceedings of the ACL 2006 Workshop on Information Extraction Beyond the Document*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 66-73.

Pech-May, F., Lopez-Arevalo, I. & Sosa-Sosa, V. (2009) Toward the validation of patient data for clinical practice guidelines. In Proceeding of the 6th International Conference on Electrical Engineering, Computing Science and Automatic Control. Toluca, Mexico, pp. 467-472.

Pech-May F. Validator for Clinical Practice Guidelines in Patients. (2010). Master's Thesis, Laboratory of Information Technology, Cinvestav, Tamaulipas, Mexico.

Peshkin, L. & Pfeffer, A. (2003). Bayesian information extraction network, *Proceedings Of the 18th International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann Publishers Inc.*, Acapulco, Mexico, pp. 421-426.

Patel, V. L., Arocha J., Diermeier, M., How, J. & Mottur-Pilson C. (2001). Cognitive psychological studies of representation and use of clinical practice guidelines, *International Journal of Medical Informatics,* 63(3):147–167.

Růzicka, M. & Svatek V. (2004). Mark-up based analysis of narrative guidelines with the stepper tool, *Journal of Studies in health technology and informatics,* 101(1):132–136.

Soderland, S., Aronow, D., Fisher, D., Aseltine, J. & Lehnert, W. (1995). Machine learning of text analysis rules for clinical records, Tr 39, Center for Intelligent Information Retrieval.

Sonderland, S. (1999). Learning information extraction rules for semi-structured and free text. Machine Learning, *Kluwer Academic Publishers*, 34(1):233-272.

Sutton, D. R. & Fox, J. (2003). The syntax and semantics of the proforma guideline modeling language, Journal *of the American Medical Informatics Association (JAMIA)*, 10(5):433-443.

Teije, A. T., Marcos, M., Balser, J., Croonenborg, V., Duelli C., Harmelen, F. V., Lucas, P. J., Miksch, S., Reif, W., Rosenbrand, K. & Seyfang, A. (2006). Improving medical protocols by formal methods, *Artificial Intelligence in Medicine,* 36(1):193-209.

Terenziani, P., Montani, S., Bottrighi, A., Molino, G. & Torchio, M. (2005). Clinical guidelines adaptation: managing authoring and versioning issues, in: S. Miksch, J. Hunter, E. Keravnou (Eds.), *Proceedings of 10th Conference on Artificial Intelligence in Medicine (AIME 2005),* Springer-Verlag, Aberdeen, Scotland, pp. 151–155.

Tu S. W. & Musen, M. (2001). Modeling data and knowledge in the EON guideline architecture, *Proceedings of 10th Triennial Congress of the International Medical Informatics Association (MEDINFO 2001),* Studies in Health Technology and Informatics, IOS Press, London, UK, pp. 280–284.

Twaddle, S. (2005). Clinical practice guidelines, *Singapore Medical Journal* 46(12):681-687.

Votruba, P., Miksch, S. & Kosara, R. (2004). Facilitating knowledge maintenance of clinical guidelinesand protocols, *Proceeding of 11th World Congress Of Medical Informatics*, Studies in health technology and informatics, IOS Press, Amsterdam, Netherlands pp. 57–61.

Wang, D., Peleg, M., Tu, S.W., Boxwala, A.A., Ogunyemi, O., Zeng, Q., Greenes, R. A., Patel, V. L. & Shortliffe, E.H. (2004). Design and implementation of the GLIF3 guideline execution engine, *Journal of Biomedical Informatics*, 37(1):305–318.

Young, O., Shahar, Y., Liel, Y, Lunenfeld, E., Bar, G., Shalom, E., Martins, S. B., Vaszar, L. T., Marom, T. & Goldstein, M. K. (2007). Runtime application of Hybrid-Asbru clinical guidelines, *Journal of Biomedical Informatics,* 40(1):507–526.