Lung Cancer Assistant: An Ontology-Driven, Online Decision Support Prototype for Lung Cancer Treatment Selection

M. Berkan Sesen, MSc[1], Rene Banares-Alcantara, PhD[1], John Fox, PhD[1], Timor Kadir, PhD[1], J. Michael Brady, PhD[1]
[1]University of Oxford, UK

Abstract

This paper describes the modelling of the LUCADA lung cancer ontology in OWL 2 and how this ontology is utilised by the online clinical decision support application Lung Cancer Assistant (LCA) for categorising patients and producing guideline-based treatment recommendations with the help of ontological inference. LCA is aimed to assist clinicians by interpreting existing patient data and use the results of this analysis to make meaningful predictions and facilitate the implementation of clinical guideline rules into daily practice.

1. Introduction

Lung cancer is the most common type of cancer and constitutes 21% of all cancer-related deaths. The two key elements in improving the survival rates for lung cancer have been reported as earlier diagnosis and referral to specialist multidisciplinary teams (MDTs) [1]. The main challenge for MDTs is that no single member of the team can sketch a comprehensive picture of the whole care pathway for a cancer patient on their own due to the complex and interdisciplinary nature of the treatment selection process.

In order to achieve the best outcomes, MDTs must keep abreast of an ever-growing flood of data from various disciplines and sources. This poses a significant informatics challenge, which can be addressed by using a clinical decision support (CDS) system that can assist clinicians by interpreting existing patient data for making meaningful predictions and facilitating the implementation of clinical guideline rules into daily practice.

With this motivation, we are building *Lung Cancer Assistant (LCA)*, an online, ontology-driven CDS prototype that aims to assist lung cancer experts in developing patient-specific and evidence-based treatment decisions that would maximise the benefits for the patient. The project is a joint effort of the Institute of Biomedical Engineering of Oxford University and the National Lung Cancer Audit (NLCA) of the NHS. *LCA* is being built and tested on the English lung cancer database (LUCADA), which consists of approximately 115,000 patient records.

In this paper, we describe how we have designed the LUCADA ontology, which is a domain-specific OWL 2 EL module of SNOMED-CT and how we make use of this ontology to formalise clinical guideline rules and to categorise patient entries.

2. Background

The adoption of clinical guidelines improves the overall quality of patient care and helps reduce practice variability and care expenses [2]. CIG formalisms have emerged

over the last two decades with the aim of improving the acceptance, maintenance and daily application of clinical guidelines by providing guideline-based recommendations for clinicians and recording the decisions and actions taken.

To date, the Guideline Interchange Format (GLIF3) [3], [4], Asbru [5], EON [6], SAGE [7] and PROforma [8] have been the most prominent of CIG models in terms of the number of clinical applications and scientific publications [2], [9], [10],[11]. All of these CIG formalisms come with proprietary expression languages to specify the decision criteria and to regulate the propagation of rules. In this work, we propose OWL-2 [12] as an alternative expression language that can be utilised for these purposes. This proposition is motivated by various reasons. First of all, none of the CIG formalisms above is suitable for the combination of guideline decision support with probabilistic techniques, which is a part of *LCA* that is being developed at the moment but is not discussed in this paper. In addition, some of the aforementioned CIG formalisms are discontinued or not maintained anymore, some do not have an execution engine and some just do not support the integration of an external standardised knowledge base, such as an ontology.

Researchers increasingly focus their efforts on the standardisation and the consolidation of the widening net of terminology and relations used in the multidisciplinary domain of cancer. From a CDS application's perspective, interoperability and adoption of a standardised vocabulary for representing domain knowledge is of utmost importance. SNOMED-CT is the most-comprehensive clinical ontology and has been approved by the NHS Information Standards Board as the full fundamental standard for all medical information applications with the National Health Service (NHS) of the UK. Therefore, we have chosen to model the domain knowledge for *LCA* in SNOMED-CT. Another major advantage of using OWL-2 for rule representation is the convenience of seamlessly extending the ontology with additional class expressions. Due to its size, the adoption of SNOMED-CT for a domain specific application entails the extraction of a module.

3. Modelling of LUCADA Ontology

The LUCADA database is maintained by NLCA with the aim of improving the outcome for people diagnosed with lung cancer and mesothelioma. It currently consists of 114 data items and around 115,000 patients. Detailed descriptions of all data items can be found in the LUCADA Data Manual v3.1.3 on the web [13].

In order to create the ontological representation of the tabular LUCADA database structure, we have manually extracted a domain-specific OWL 2 EL module of SNOMED-CT, which only includes concepts that are immediately relevant to represent the 114 LUCADA data items. This module constitutes the clinical domain of the LUCADA ontology and it covers all data items in the database, enabling a semantically accurate mapping of patients between the database and the ontology. In addition to the clinical domain, LUCADA ontology contains a small argumentation domain, which holds the classes that are used for guideline rule inference and argumentation. Figure 1 depicts the higher level classes and properties in the LUCADA ontology.

For the clinical domain, we have identified some key SNOMED-CT concepts that were essential for representing a patient record in LUCADA. Based on this, we modelled a data item either as an object property or a data property in the ontology. The object properties and data properties, added to the ontology for this purpose, are shown in blue and green respectively in Figure 1. In its current version, the LUCADA ontology T-Box consists of 396 classes, 35 object properties and 60 datatype properties.
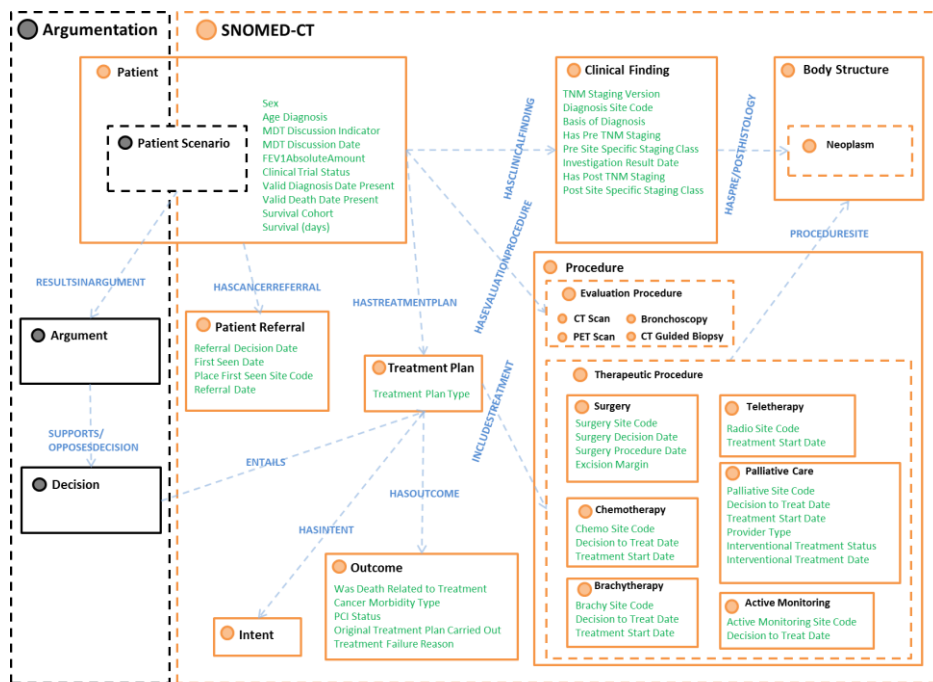

Figure 1. The LUCADA Ontology: The orange circles represent SNOMED-CT classes (concepts) and the black circles represent non-SNOMED-CT argumentation classes. The blue arrows represent object properties between the classes and the green list items represent the datatype properties belonging to the respective classes.

In order to keep the size of the module small and data abstraction concise, we have prioritised modelling data items as data properties of relevant classes. Representation in the form of an object property often entailed the inclusion of a corresponding new class in addition to the object property representing the data item and therefore was only chosen in the following situations:

1) When the creation of a distinctive class for a data item could be beneficial to model other data items through the newly created class' properties. An example of this is the 'hasTreatmentPlan' object property which has domain Patient and range Treatment Plan. Here, the inclusion of Treatment Plan as a separate concept enables connecting it to constituent treatment options under Procedure with the

'includesTreatment' object property. This structure enables modelling a patient with a suggested treatment plan which includes various distinct treatment types.

2) When more than one data item could be grouped under a common parent concept. An example of this is the introduction of the SNOMED-CT concept Clinical Finding, subsuming all diseases such as Primary (Cancer) Diagnosis, Dementia, Cardiovascular Disease and other comorbidities. This allows use of a single object property, 'hasClinicalFinding', to connect a Patient individual to all (taxonomically) disease-related concepts taken from the database.

3) When suitable object properties already existed within SNOMED-CT. While building the LUCADA ontology, precedence was given to using the properties inherently defined in SNOMED-CT. These are originally called defining attributes but translate into object properties in the OWL-2 representation of SNOMED-CT.

Some data items did not have one to one mappings with a SNOMED-CT concept. In such cases, the data items were modelled as compound concept definitions. For instance, the 'Severe Weight Loss' data item, which is one of the comorbidity types in the database, was modelled as "Weight Loss Finding 'Severity' Severe", where Weight Loss Finding and Severe are classes and 'Severity' is an object property in SNOMED-CT. Furthermore some data items were just too complex to be represented within the vocabulary of SNOMED-CT. As a result, 13 new classes had to be manually added to the LUCADA ontology. An example of such a class is Induction Chemotherapy to downstage before surgery, which appears as a treatment plan type in LUCADA. Since modelling this as a compound structure was not feasible, a concept with the same name was added to the LUCADA ontology. It is worthwhile to mention that for this very concept, a concept addition request was made to the SNOMED-CT UK Terminology Centre. After their review, the request was accepted on July 8 2011 and the concept has been added in the October 2011 SNOMED-CT release.

4. Patient and Guideline Rule Representation in LUCADA Ontology

Once the ontology design was complete, we have written a program using Java v.6 and Java OWL API v.3.2.3 [14] for the automatic transfer of patients from the database to LUCADA ontology. Figure 2 depicts how the fictitious patient, Jenny_Sesen is represented in the ontology. As can be seen, we create corresponding Patient, Cancer and Histology individuals (depicted as purple diamonds) to represent Jenny_Sesen's unique database record given in Figure 2a. In cases where the data item to be represented is not patient-specific, such as Tumour Laterality: Right, we use a reference individual to represent that standard value rather than creating a new individual for every patient entry. This implementation of reference individuals is mostly used for standard valued concepts such as Severity, Side, Decision, and Treatment Plan.
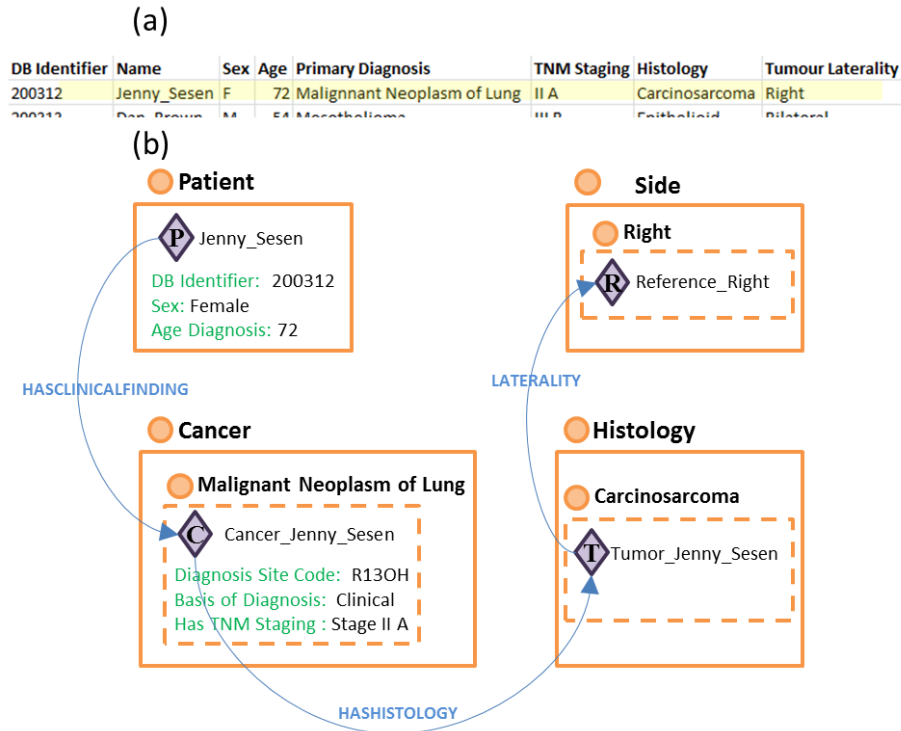
Figure 2. (a) The tabular database representation of the patient record 'Jenny_Sesen' (b) The ontological representation of 'Jenny_Sesen' in the LUCADA ontology, where the purple diamonds are individuals and orange circles are classes, blue arrows are object properties and green list items are data properties in the ontology.

Following the transfer of patients to the ontology, we have expanded the Java program to include a guideline rule inference framework, which is used to represent guideline rule criteria as Patient Scenario class expressions in the ontology. The hybrid Patient Scenario class is a subclass of both SNOMED-CT's Patient class and the proprietary Argumentation class. It can conceptually be regarded as a hypothetical patient cohort that fulfils a guideline's rule criteria. We can demonstrate this by analysing a guideline rule taken from the NICE Lung Cancer 2011 Guidelines [1].

"Offer chemotherapy to patients with stage III or IV NSCLC and good performance status (WHO 0, 1 or a Karnofsky score of 80–100)." **(R1)**

We can break this rule down into two functional components: 1) the head, which specifies the patients for whom the rule is applicable; 2) the body, which specifies the action(s) to take when the specific conditions in the head are satisfied [15]. The head of R1 can be written in the OWL-2 syntax as a compound logical expression given in (E1):

"(hasClinicalFinding **some** (NeoplasticDisease **and** ((hasPreTNMStaging **value** "III") **or** (hasPreTNMStaging **value** "IV")) **and** (hasPreHistology **some** NonsmallCellCarcinoma))) **and** (hasPerformanceStatus **some** (WHOPerfStatusGrade0 **or** WHOPerfStatusGrade1))"   **(E1)**

(E1) can be translated into plain English as: *(Individuals) whose performance status is either 0 or 1 and who have Neoplastic Disease with TNM staging of either 3 or 4 and histology finding type non-small cell cancer*. In order to formalise the guideline rules in the ontology, we create a Patient Scenario subclass for each such rule and then add a corresponding equivalence class expression as given in (E1). Adoption of the OWL 2 EL profile for the LUCADA ontology allowed us to make use of the existential quantification restrictions employed in Patient Scenario class expressions. When an ontology reasoner is run, all patients that fit the equivalent class expression, i.e. guideline rule criteria, of a specific rule are automatically inferred to be members of this particular Patient Scenario.
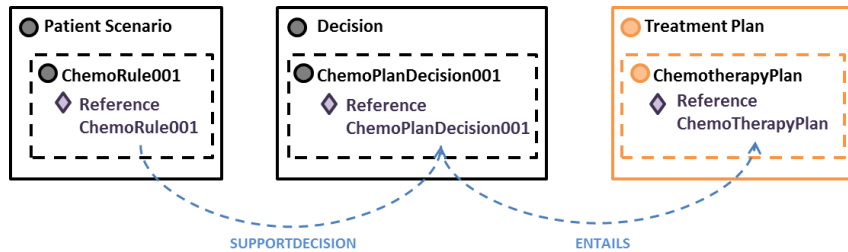

Figure 3. A reference-individual-level connection between a Patient Scenario, a Decision and a Treatment Plan, which represent the rule body statement "Offer Chemotherapy"

What action(s) to take for the members of a particular Patient Scenario, i.e. the rule body, are then explicated through an object property relation between the reference individuals of the Patient Scenario, Decision and Treatment Plan classes. This is depicted in Figure 3 for the body of rule R1.

5. Lung Cancer Assistant

The LUCADA ontology and the guideline inference framework have been brought together within the *Lung Cancer Assistant* (*LCA*) web-based decision support application, which is developed using the GWT Software Development Kit (SDK) for Java. As with all web-based software, the Lung Assistant architecture is separated into two components: 1) Client-side code that runs on an end user's local computer and connects to a server as necessary; 2) Server-side code that runs on a remote server and is reachable from an end user's local computer on demand.

The LUCADA ontology and the database are kept on the server-side and are utilised through event calls that are triggered from the client-side when a user interacts with the application. The patient interactions are handled by various ontology editing classes, database to ontology converting classes and the rule inference framework classes on the server side. All ontology-related classes were written using the OWL API v3.2.3.

The current version of LCA allows creating, updating and saving patients records. Following any of these user events, the ontology reasoner is triggered to 'reclassify' the ontology on the server-side, which ensures that newly entered data are reflected in the patient-specific guideline recommendations immediately. Due to the number of patients in the database, storing patient-record individuals in the ontology was not feasible in terms of classification. When all 115,000 patient records were added, the ontology approximated to 1 GB in size with over 700,000 individuals. None of the commonly used reasoners, i.e. HermiT, Fact++, Pellet, could manage to finish classifying this ontology. Therefore, we have designed the architecture, in which upon creation each patient is temporarily added to the ontology for classification and then removed. The inferred Patient Scenario memberships are then stored in a separate table in the database to record what rules apply to each patient.
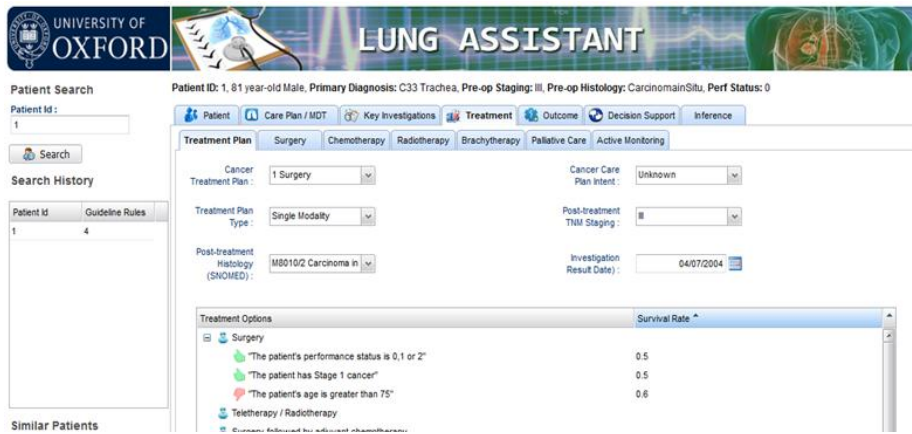

Figure 4. A screenshot of the Treatment tab of LCA: The patient specific treatment arguments are displayed within the Treatment Options box which is automatically generated as a result of the rule inference framework explained

The data item fields in the Lung Assistant UI are distributed into different tabs in the order of the corresponding LUCADA sections given in [13]. A screenshot of the Treatment Tab of the application is given in Figure 4. When a new patient is saved or an existing patient is loaded, the treatment recommendations are displayed in the scrollable Treatment Options box.

6. Discussion and Future Work

In the current version of LCA, ontological inference is used to create patient-specific treatment arguments by automatically grouping patients based on guideline rules in the ontology. So far, we have input all rules that concern treatment selection in the British Thoracic Society guidelines into *LCA*. Our next task will be extracting and formalising rules from the National Institute for Clinical Excellence guideline documents into the system.

In addition, we are currently benchmarking various Bayesian techniques to integrate with the current guideline-based decision support functionality. The probabilistic inference, provided by the suitable Bayesian classifier, is intended to reinforce the

qualitative patient-specific arguments by quantifying their reliability and introducing degrees of support based on the argument claim's significance on survival.

7. References

[1] NICE, "Quick Reference Guide Lung cancer," 2011.

[2] P. D. E. Clercq, K. Kaiser, and A. Hasman, "Computer-interpretable Guideline Formalisms," Studies In Health Technology And Informatics, vol. 139, pp. 22-43, 2008.

[3] A. Boxwala et al., "GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines.," Journal of biomedical informatics, vol. 37, no. 3, pp. 147-61, Jun. 2004.

[4] D. Wang et al., "Design and implementation of the GLIF3 guideline execution engine.," Journal of biomedical informatics, vol. 37, no. 5, pp. 305-18, Oct. 2004.

[5] S. Miksch, Y. Shahar, and P. Johnson, "ASBRU: A task Specific, Intention-based, and Time Oriented Language for Representing Skeletal Plans," Aids, pp. 1-25, 1997.

[6] M. Musen, S. TU, and K. Das, "EON: a component-based approach to automation of protocol-directed therapy," Journal of the American Medical Informatics Association, vol. 3, no. 6, 1996.

[7] P. Ram et al., "Executing clinical practice guidelines using the SAGE execution engine.," Studies in health technology and informatics, vol. 107, no. Pt 1, pp. 251-5, Jan. 2004.

[8] D. Sutton and J. Fox, "The Syntax and Semantics of the PRO forma Guideline Modeling Language," Journal of the American Medical Informatics Association, vol. 10, no. 5, pp. 433-443, 2003.

[9] A. Boxwala et al., "Toward a representation format for sharable clinical guidelines.," Journal of biomedical informatics, vol. 34, no. 3, pp. 157-69, Jun. 2001.

[10] D. Isern and A. Moreno, "Computer-based execution of clinical guidelines: a review.," International journal of medical informatics, vol. 77, no. 12, pp. 787-808, Dec. 2008.

[11] M. Peleg, S. Tu, and J. Bury, "Comparing Guideline Models : A Case-study Approach," Journal of the American Medical Informatics Association, pp. 52-68, 2003.

[12] W3C, "OWL 2," 2009. [Online]. Available: http://www.w3.org/TR/2009/PR-owl2-overview-20090922/.

[13] The National Lung Cancer Audit, "The National Clinical Lung Cancer Audit (LUCADA) Data Manual," 2010.

[14] M. Horridge and S. Bechhofer, "OWL API Version 3.2.3," 2011. [Online]. Available: http://owlapi.sourceforge.net/.

[15] W3C, "SWRL: A Semantic Web Rule Language Combining OWL and RuleML," 2011. [Online]. Available: http://www.w3.org/Submission/SWRL/.