Toward a Translational Medicine Approach for Hypertrophic Cardiomyopathy

Catia M Machado[1], Francisco M Couto[1], Alexandra R Fernandes[2,3,4], Susana Santos[2,3], and Ana T Freitas[5]

[1] LaSIGE, Departamento de Informática, Universidade de Lisboa, Lisboa, Portugal
cmachado@xldb.di.fc.ul.pt
[2] Universidade Lusófona de Humanidades e Tecnologias, Lisboa, Portugal
[3] Centro de Química Estrutural, Instituto Superior Técnico, Lisboa, Portugal
[4] Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal
[5] Instituto de Engenharia de Sistemas e Computadores, Instituto Superior Técnico, Lisboa, Portugal

Abstract. Hypertrophic cardiomyopathy (HCM) is a complex genetic disease characterized by a variable clinical presentation and onset, as well as a high number of associated mutations.
Therefore, this disease is a good candidate for a translational medicine approach to assist in its prognosis. For this purpose, we propose a framework containing two components: one for data integration, and another for data analysis based on clinical-genetic associations obtained with data mining techniques.
In this article we present the implementation of the first component. At its basis is a semantic data model developed in OWL representing the clinical and genetic data necessary for the characterization of HCM patients. This model follows a modular approach and includes mappings to controlled vocabularies such as the NCI Thesaurus and SNOMED-Clinical Terms.
The development of the model has been done in collaboration with biomedical experts, who are also the providers of the data to populate it.

Keywords: translational medicine, data integration, data mining, hypertrophic cardiomyopathy, clinical decision support systems

1 Introduction

Hypertrophic cardiomyopathy (HCM) is a genetic disease that may afflict as many as 1 in 500 individuals, and is the most frequent cause of sudden cardiac death among apparently healthy young people and athletes [1, 2]. It is characterized by a variable clinical presentation and onset, which results in a difficult clinical diagnosis prior to the development of severe or even fatal symptoms [1, 2]. Moreover, its genetic diagnosis is complex, since there are approximately 900

mutations in more than 30 genes currently known to be associated with the disease [3].

In terms of prognosis, the task is by no means trivial since the severity of HCM varies even between direct relatives. It has been observed that the presence of a given mutation can correspond to a benign manifestation in one individual and result in sudden cardiac death in another [1, 2].

As a consequence of all these factors, HCM is an example of a disease that can benefit from a translational medicine approach to aid in its prognostic. Given the clinical manifestations of the disease and the mutations associated with it, it might be possible to identify a set of factors that will aid cardiologists in the task of risk assessment and management. This task is of paramount importance as it could enable the timely identification of patients prone to sudden cardiac death, and the regulation of their physical activities in order to minimize such risk.

A pivotal step toward the concretization of a translational medicine approach consists in the integration of data originating from different domains of knowledge. Ontologies, and controlled vocabularies in general, are important tools for data integration since they provide a standard way of representing knowledge. Ideally, these vocabularies are references accepted by the community, such as the Gene Ontology [4] and the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) [5].

Due to their importance in data integration, ontologies are a central piece in the implementation of the Semantic Web vision proposed by Tim Berners-Lee [6]. This vision is that of a Web of data, rather than the Web of documents that is the current standard. The general idea is that instead of links connecting Internet pages that are mostly designed to be interpreted by humans, we can have links connecting the data elements themselves. The Semantic Web can be seen as a framework for data integration at a Web-wide scale that is independent of the domain of knowledge, and focused on the meaning and on the context of the data.

In order to implement this vision of a new Web, a set of tools and technologies have been proposed as standards by the World Wide Web Consortium (W3C) [7], namely: RDF, a language for data representation and interchange [8]; and OWL, a language to formally define meaning in Web resources that supports reasoning [9].

Several translational medicine examples exist using Semantic Web standard technologies, namely the work developed by Gudivada and colleagues [10] in a task of gene prioritization, ASSIST (Association Studies aSsisted by Inference and Semantic Technologies) [11], and the Neuroweb system [12]. In the first example, the integrated resources are locally maintained in relational format and instantly converted to RDF upon need, whereas in the other two they are maintained in their original format and location. In all three examples, data integration is mediated by an ontology developed in OWL. None of the systems reuse existing ontologies, although the developers of Neuroweb considered the use of resources such as SNOMED-CT and the Disease Ontology. However,

the authors verified that SNOMED-CT did not provide a suitable formulation of concepts for their purpose, and that the taxonomy adopted by the Disease Ontology was different from the one used by the clinicians participating in the Neuroweb network.

In our translational medicine approach for HCM, we are interested in the prognosis of the disease. Our goal is the identification of associations between clinical and genetic factors that can be used to aid medical doctors in the prediction of the outcome of the disease, for every individual patient, particularly in respect to the occurrence of sudden cardiac death. For this purpose, we propose a framework that integrates clinical and genetic data mediated by a semantic data model representing the disease, and that explores data mining models depicting the clinical-genetic associations (Figure 1).

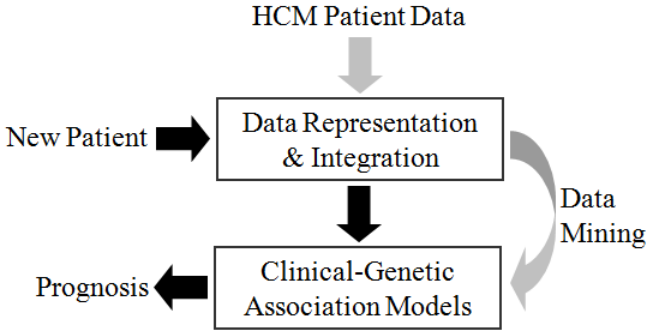

Fig. 1. High-level schematic representation of the translational medicine framework we are developing for the disease HCM. Data of patients with known prognosis (*grey arrows*) will be represented and integrated according to a semantic data model developed for the disease, and will be explored with data mining techniques to obtain clinical-genetic association models. Data of new patients, with unknown prognosis (*black arrows*), will be represented according to the semantic model, and will be evaluated based on the clinical-genetic association models to obtain a prediction of the prognosis.

In this article we present the implementation of the data representation and integration component. The semantic data model at the core of this component provides a useful framework for the integration of data from two different domains of knowledge, clinical and genetic, and from different institutions. The concepts modeled were identified and defined with the help of medical doctors, geneticists and molecular biologists based on the data elements collected during their activities. The model is currently being populated with data from four medical institutions and two research centers.

The rest of this document is organized as follows: Section 2 describes the development of the semantic model; Section 3 presents the semantic model; Sections 4 and 5 contain a Discussion and the Conclusions, respectively.

2 Semantic Model Development

The development of the HCM semantic model followed the guidelines presented by Noy and McGuinness [13] for ontology development. The first step was the definition of the domain (i.e. the disease, HCM) and the scope (i.e. the representation of the data necessary for the diagnosis and the prognosis of HCM), followed by the enumeration of relevant concepts and the reuse of existing controlled vocabularies.

The following steps describe our approach:

1. An initial set of concepts was identified in collaboration with biomedical experts.
2. The concepts were represented in OWL Lite, including hierarchical and non-hierarchical relations.
3. Existing controlled vocabularies of interest were searched.
4. New concepts to consider were identified in these controlled vocabularies.
5. The concepts and relations represented were continuously validated by the biomedical experts.
6. The consistency of the model (i.e. the absence of syntactic or semantic errors) was evaluated periodically.

The model was developed in the Protégé-OWL editor (version 3.4.2) [14] following a modular approach. The consistency evaluations were performed by running the reasoner HermiT [15] available in Protégé.

OWL was the language of choice to comply with the Semantic Web standards and to take advantage of external resources published in the Semantic Web.

The identification of controlled vocabularies of interested was performed using BioPortal, from the National Center for Biomedical Ontology [16]. The concepts initially identified in collaboration with the biomedical experts were used as search terms, namely *clinical history*, *angina*, *hypertrophic cardiomyopathy*, *resuscitated sudden death*, and *electrocardiography*. We searched for vocabularies referring to the medical and molecular biology domains that contained the concepts of interest, and that represented these concepts in a hierarchical organization in accordance with the vision of the HCM domain conveyed by the experts. The adequacy of the vocabularies was evaluated based on their scope. The list initially compiled was narrowed down based on the number of concepts of interest the vocabulary contained.

As previously published [17], three vocabularies were initially identified and considered for the HCM model: SNOMED CT (version 2010_01_31), the National Cancer Institute Thesaurus (NCIt) (version 10.03)[18], and the Ontology of Clinical Research (OCRe) (version 0.95) [19].

We opted to use more than one vocabulary for each module for two reasons: (*i*) none of the vocabularies contained a complete list of the concepts of interest; (*ii*) the provided representation of the concepts was not always the most suitable for our purposes.

We did not reuse entire modules of any of the vocabularies since our goal was not to convey the most complete representation of the disease. We rather wanted

to represent the concepts necessary for its diagnosis and prognosis, as well as include a minimum set of concepts that would facilitate the mapping between the HCM model and the vocabularies. In addition, one of the concerns during the development of the model was to maintain it as simple as possible, in order to avoid overwhelming the biomedical end-users with superfluous information.

Our approach to the use of these vocabularies is summarized in the following steps:

1. The regions of interest in each vocabulary were identified.
2. The hierarchical structure of the HCM model was refined in accordance with the vocabulary considered.
3. The concepts in the model were renamed in accordance to the vocabulary.
4. The concepts in the model were manually mapped to the equivalent concept in the controlled vocabulary, through a *hasDbXRef* property [6].
5. When the vocabulary provided a definition for the mapped concept, it was added to the model.

Considering that the controlled vocabularies were also exploited to identify new concepts to include in the model, they served the dual purpose of aiding in the development of the model and providing mappings.

Since its preliminary version [17], the HCM model has been extended in number of concepts and mappings. One of the previously considered vocabularies, OCRe, was eliminated due to the deprecation of the concepts we had reused (e.g. *Health Care Site*). The model is currently mapped to four controlled vocabularies: SNOMED CT and the NCIt as before, and also to the Gene Regulation Ontology (version 0.5, released on 04_20_2010) [20] and to the Sequence Ontology (released on 11_22_2011) [21].

In addition to the two major alterations that resulted in the conversion of the model from one to three modules and in the incorporation of the knowledge from the controlled vocabularies, the model suffered several rounds of adjustments.

3 HCM Semantic Model

The resultant HCM model is composed by three modules:

- *Clinical Evaluation* - containing administrative concepts and clinical data elements that play a role in the diagnosis and the prognosis of HCM patients.
- *Genotype Analysis* - containing concepts associated with the genetic testing of biological samples.
- *Medical Classifications* - an auxiliary module containing medical standards used in the characterization of clinical elements such as patient symptoms.

Table 1 shows the composition of the three modules, both in number of concepts and properties. *Clinical Evaluation* is the largest, with a total of 63 concepts and approximately 60 object and data properties (Figures 2 and 3).

[6] `http://www.geneontology.org/formats/oboInOwl\#hasDbXref`

Table 1. Composition of the *Clinical Evaluation*, *Genotype Analysis*, and *Medical Classifications* modules in terms of: number of top-level concepts, total number of concepts, and number of data and object properties.

Module	Top-level concepts	Total concepts	Properties
Clinical Evaluation	5	63	60
Genotype Analysis	7	19	39
Medical Classifications	2	4	2

Genotype Analysis contains 19 concepts and approximately 39 properties (Figure 4). Finally, *Medical Classifications* contains two high-level concepts (*Angina Classification* and *Heart Failure Classification*), each with one sub-concept, and a total of ten instances. As an example of this last module, Figure 5 shows the concept *Heart Failure Classification* and the data properties for one of its instances, *NYHA_Class2*.

The *Clinical Evaluation* (ce:) module imports the other two, *Genotype Analysis* (ga:) and *Medical Classifications* (mc:). The bridge between modules is made through the following non-hierarchical relationships (here represented as triples, where the central elements are object properties):

- ce:*Patient* ce:*hasBiologicalSample* ga:*Biological Sample*
- ce:*Biomarker Analysis* ce:*performedInBiologicalSample* ga:*Biological Sample*
- ce:*Angina* ce:*hasAnginaClassification* mc:*Angina Classification*
- ce:*Congestive Heart Failure* ce:*hasHeartFailureClassification* mc:*Heart Failure Classification*

Patients' mutations can be identified through this relationship between *Clinical Evaluation* and *Genotype Analysis* since in the latter module a *Biological Sample* is connected with the mutations identified therein.

In terms of mappings to controlled vocabularies, SNOMED CT was used in the *Clinical Evaluation* module, the NCIt in the *Clinical Evaluation* and *Genotype Analysis* modules, the Gene Regulation Ontology and the Sequence Ontology in the *Genotype Analysis* module (see Table 2). More precisely, each vocabulary was considered in the following top-level concepts:

- SNOMED CT: *Clinical Finding* and *Observable Entity*
- NCIt: *Health Care Site*, *Person* and *Procedure* (from *Clinical Evaluation*); *Biological Sample*, *Gene*, *Mutation* and *Protein* (from *Genotype Analysis*)
- Gene Regulation Ontology: *Nucleic Acid Molecule*
- Sequence Ontology: *Primer*

Although the *Medical Classifications* module does not contain mappings to controlled vocabularies, its concepts are nonetheless linked to Web pages where their definition can be found.

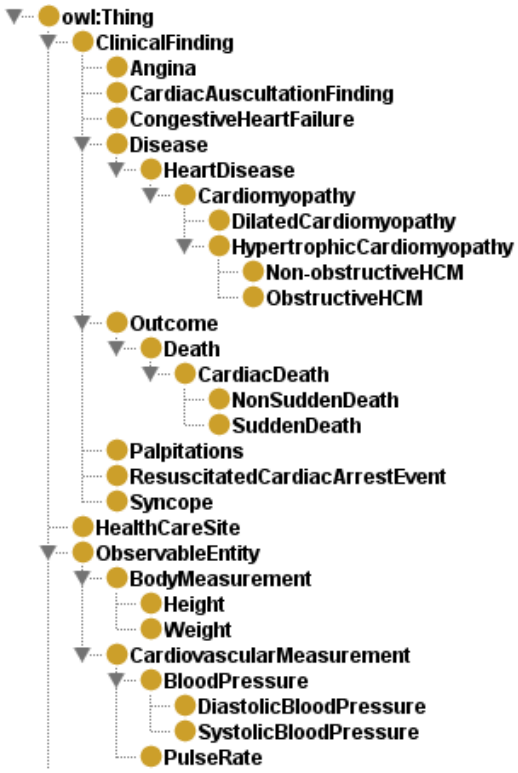

Fig. 2. Hierarchical structure of the *Clinical Evaluation* module, showing three of the top-level concepts (*Clinical Finding*, *Health Care Site* and *Observable Entity*) with all sub-concepts visible.

4 Discussion

The decision to divide the model in modules was motivated by the observation that we wanted to represent two conceptually different types of knowledge: the knowledge related to patients, such as symptoms and treatments (represented in the *Clinical Evaluation* module); and the knowledge related to the analysis of biological samples collected from patients, such as amplification fragments and mutations (represented in the *Genotype Analysis* module). The third module, *Medical Classifications*, was added to represent standard medical classifications of any type, since these are independent from both patients and sample analysis. In terms of the use of the model in the final prognosis framework, this also means that we can easily provide two different views of the data: one centered on the patient, which is of interest for the medical doctors; and one centered on the biological samples, which is of interest for the molecular biologists.

Additionally, the modular development of the HCM model also facilitates its extension and reutilization. While the *Clinical Evaluation* module is the most

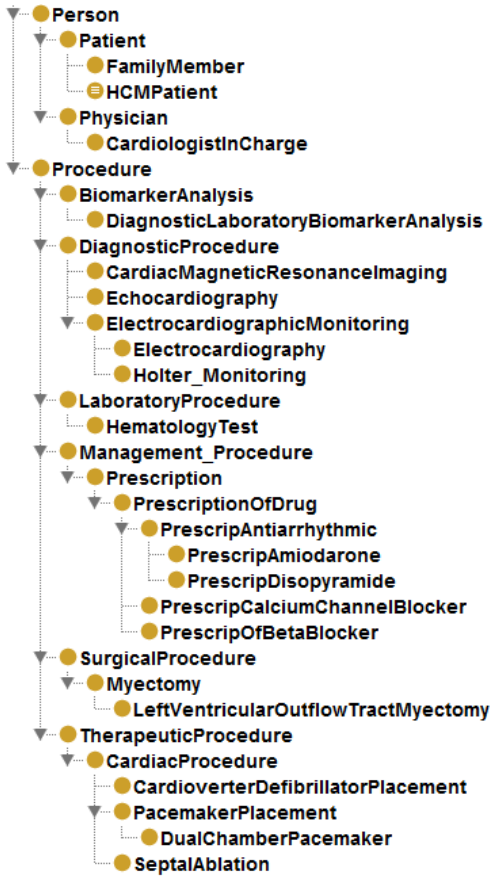

Fig. 3. Hierarchical structure of the *Clinical Evaluation* module, showing two of the top-level concepts (*Person* and *Procedure*) with all sub-concepts visible. *Defined* classes, i.e. containing necessary and sufficient conditions, are indicated by a symbol with three horizontal lines.

specific and is best suited for the characterization of heart diseases, *Genotype Analysis* can be used in the context of any disease. In the case of *Medical Classifications*, although presently containing only two classes representing the classifications used by the medical experts, it can be expanded to include any standard or set of guidelines that refer to the medical aspects of HCM characterization or any other disease.

The use of controlled vocabularies proved to be advantageous on several levels: it saved us the work of creating a completely new model; it assisted us in identifying additional concepts and relations of interest; and it will facilitate the future addition of concepts since they can be searched in the vocabularies and easily integrated in their hierarchy. Nonetheless, the process was far from trivial.

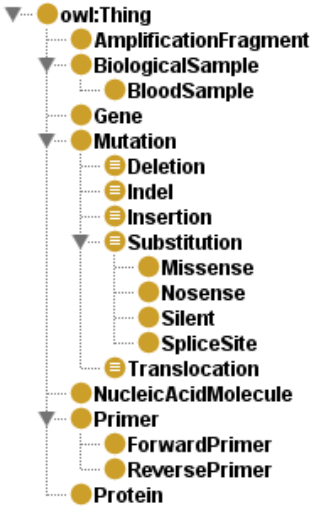

Fig. 4. Hierarchical structure of the *Genotype Analysis* module, showing all seven top-level concepts with all sub-concepts visible. *Defined* classes, i.e. containing necessary and sufficient conditions, are indicated by a symbol with three horizontal lines.

Table 2. Percentage of concepts from the HCM semantic model mapped to the following controlled vocabularies: SNOMED - Clinical Terms, NCI Thesaurus, Sequence Ontology and Gene Regulation Ontology. The percentages are indicated for the modules *Clinical Evaluation* and *Genotype Analysis*.

Module	Vocabulary (%)				Total (%)
	SNOMED CT	NCIt	SO	GRO	
Clinical Evaluation	42.9	42.9	-	-	85.8
Genotype Analysis	-	63.2	26.3	5.3	94.8

First of all, for the identification of the vocabularies we searched for concepts of interest on all the vocabularies available from BioPortal. This was a challenging task, since several vocabularies exist that fulfilled the requirement. After evaluating the most promising options, the initial list was progressively narrowed down until only those indicated remained. When this process was first started, we were not aware of the existence of the Biomedical Ontology Recommender service [22] available from BioPortal. However, we tested it afterwards and concluded that the vocabularies chosen coincided with the recommendations provided by the service. Additionally, the use of this service would have expedited considerably our work since it provides recommendations based on several concepts at the same time, and our searches were executed for one concept at a time.

Secondly, several issues came to light during the development of the model related to:

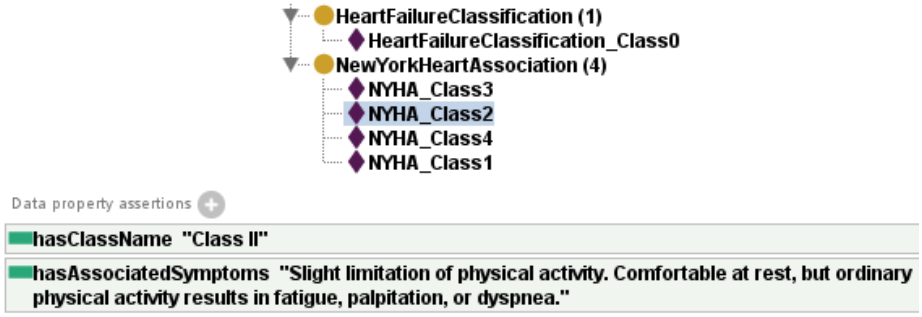

Fig. 5. Representation of the concept *Heart Failure Classification* (from the *Medical Classifications* module) with: one instance; and the sub-concept *New York Heart Association* with its own four instances and the data properties of the instance *NYHA_Class2*.

- **Absent concepts**: Inexistence of a concept of interest in the vocabulary.
- **Complexity**: Excess of concepts and of level of detail in general.
- **Placement**: Different possibilities concerning the placement of a concept in the hierarchy of the model.
- **Overlapping regions**: Existence of overlapping concepts/regions of interest on different vocabularies.
- **Absent textual definitions**: Inexistence of textual definitions for concepts of interest.

The **Absent concepts** issue occurred both in the *Clinical Evaluation* and the *Genotype Analysis* modules. In the former module, we needed a concept *Cardiologist in charge* to represent the cardiologist that is primarily responsible for the HCM patient. According to the specifications of the biomedical experts guiding the development of the model, this cardiologist is the only medical doctor associated with the patient for this disease, and is responsible for every data element and evaluation represented in the model. Neither SNOMED CT nor NCIt provide such a representation, and the notions of "Physician" and of specific medical specialties such as "Cardiologist" are represented under *Occupation*, which can be interpreted as a label rather than a representation of a person. In this situation, we opted to use the concept *Person* from NCIt to aggregate *Patient* and *Physician*, and added *Cardiologist in Charge* as a sub-concept of *Physician*. In the *Genotype Analysis* module we needed to represent the *Translocation* and *Indel* sub-concepts of *Mutation*, as shown in Figure 4. While *Mutation* was mapped to the NCIt, this vocabulary does not include the indicated sub-concepts, and thus we mapped them to the Sequence Ontology.

The solution followed to deal with the **Complexity** of the controlled vocabularies, both in the form of number of concepts and detail of representation, was to consider only the concepts necessary for the description of the disease and for the structure of the model. The structure is particularly important for the mapping of the HCM model to external resources and for the future addition of

concepts. An example of the complexity issue occurred with the concept *Procedure*. This concept is mapped to *Intervention or Procedure* from the NCIt, which contains thirteen sub-concepts, but we were interested in only five of them. If all thirteen were considered, the level of complexity of the model would be increased without any benefit for the end-users.

The **Placement** issue derived from our decision of not representing more than one parent per concept (i.e. multiparenting), even at the expense of a possible loss of detail. This decision was motivated by our intention of creating a model that would provide a straightforward experience to the biomedical experts when inputing or retrieving data (during the utilization of the prognosis framework), and thus avoid possible uncertainties due to multiple options. As such, we were occasionally forced to evaluate different possibilities for the placement of a concept in the hierarchy of the model. This occurred with concepts from SNOMED CT, in which situations we recurred to the NCIt to help us identify a solution common to both vocabularies. One such case occurred with *Syncope*, a *Clinical Finding* that is represented in SNOMED CT as a sub-concept of three different concepts: *Clinical history and observation finding*, *Finding by site* and *Disease*. In the HCM model we consider the concept *Clinical Finding* and its sub-concept *Disease*, and the decision was whether to place *Syncope* directly under the first-level *Clinical Finding* or the second-level *Disease*. In NCIt the concept is represented directly under the concept *Finding* and not under its sibling *Disease or disorder*, and consequently we chose to place it under *Clinical Finding* in the HCM model. Similar decisions were made for the concepts *Angina* and *Congestive heart failure*, which are sub-concepts of *Finding by site* and *Disease* in SNOMED CT, and of *Finding* in NCIt.

The **Overlapping regions** issue results from the existence of more than one vocabulary describing the same domain of knowledge. According to the accepted OBO Foundry [23] principle named "clearly delineated content" (FP005[7]), ontologies should be orthogonal to each other in order to enable the utilization of two different ontologies to define complementary perspectives on the same entities. In essence, we agree with this principle since the existence of a single ontology for a given domain would mean that anyone wanting to reuse it in an application semantic model would just have to follow it and consider the necessary knowledge. On the other hand, in light of our experience with the development of the HCM model, we consider that the availability of more than one vocabulary can be positive when no vocabulary is accepted as the single reference by the community.

An example of the overlapping regions in the *Clinical Evaluation* module occurred with the concept *Outcome*, a *Clinical Finding* with possible examples of outcomes being decreased pain and death. *Clinical Finding* and its sub-concepts are mapped to SNOMED CT, but this vocabulary represents *Death* in a high-level class *Event*, which is not necessary for the HCM model. Moreover, NCIt has a concept *Outcome* under *Finding*, which also includes several sub-concepts relevant for the HCM model: *Death*, *Cardiac death*, *Sudden cardiac death* and

[7] `http://www.obofoundry.org/wiki/index.php/FP_005_delineated_content`

Non sudden cardiac death. In this situation the decision was to consider *Outcome* and its sub-concepts from NCIt in the *Clinical Finding* concept, which is otherwise mapped to SNOMED CT.

Two other examples of the overlapping regions issue in the *Genotype Analysis* module involved the concepts *Primer* and *Nucleic acid molecule*. *Primer* is represented in the NCIt under *Drug, Food, Chemical or Biomedical Material* and without sub-classes. However, in the Sequence Ontology, a *Primer* is a *Sequence feature* with the two sub-classes *Forward Primer* and *Reverse Primer*, which were included in the HCM model. In the second situation, the concept *Nucleic Acid* was intended to represent actual nucleic acid molecules extracted from biological samples. While both the NCIt and the Sequence Ontology include the concept *Nucleic Acid*, neither define it suitably for our purposes: the former defines *Nucleic acids* as "A family of macromolecules", whereas the latter defines *Nucleic acid* as "An attribute describing a sequence consisting of nucleobases bound to repeating units". Consequently, we opted to use the Gene Regulation Ontology exclusively for its concept *Nucleic acid molecule*, witch is more suitably defined as a "A complex, high-molecular-weight biochemical macromolecule composed of nucleotide chains that convey genetic information".

The overlapping regions issue is of particular importance given that using a domain representation that is unfamiliar to the end-users of the HCM prognosis framework may hinder significantly their acceptance of the framework.

The **Absent textual definitions** issue was perceived as a significant burden to the reuse of the affected concepts, since there where situations in which their intended use was not readily understandable. This was a common problem when using SNOMED CT, as this vocabulary lacks definitions for most of its concepts. For example when representing the concept *Cardiologist in Charge*, it was only possible to interpret the intended use of the concept *Cardiologist* based on the hierarchical organization of the vocabulary. By contrast, the NCIt has available detailed descriptions for the majority of its concepts, which provides a greater assistance when more complex decisions have to be made. This issue is not new, and has already been the subject of an OBO Foundry principle (FP 006 textual definitions[8]).

An issue particular to the development of the *Genotype Analysis* module occurred with the concepts *Nucleic acid molecule*, *Gene* and *Protein*. As represented in the Gene Regulation Ontology, these concepts are related with each other: *Gene* is represented under *DNA*, which in turn is a *Nucleic acid*; and *Nucleic acid* and *Protein* are both *Information biopolymer*(s) ("macromolecules that harbor biological information in their structures"). However, these relationships could not be conveyed in the HCM because what we want to represent under each concept is conceptually different: *Nucleic Acid Molecule*, the physical molecules; *Gene*, the list of genes associated with HCM (not the physical genes); and *Protein*, the list of proteins encoded by the genes associated with HCM (not the physical proteins).

[8] `http://www.obofoundry.org/wiki/index.php/FP_006_textual_definitions`

5 Conclusions

Hypertrophic cardiomyopathy (HCM) is a complex genetic disease both in terms of diagnosis and prognosis, due to a great variability in terms of clinical manifestations and associated mutations. Furthermore, the presence of the same mutation in different individuals can result in very different clinical manifestations. Consequently, this disease is a good candidate for a translational medicine approach.

In this article we present a semantic data model that is the core element of a component of data representation and integration in our proposed prognosis framework for HCM. The data integrated with this component will be explored with data mining techniques to identify associations between clinical and genetic data. Our aim is that these associations might be used as guidelines to assist cardiologists in the prediction of the outcome of the disease for individual patients, in addition to existing guidelines [24]. In particular, we are interested in predicting the occurrence of sudden cardiac death.

The first step in the development of the model was the identification of the clinical and genetic data elements considered in the actual assessment of patients, in accordance with the practice of the medical and molecular biology experts with whom we collaborate.

The model was developed in OWL following a modular approach to facilitate its extension and reutilization. The concepts in all of the three modules that compose it (*Clinical Evaluation*, *Genotype Analysis* and *Medical Classifications*) are also mapped to external controlled vocabularies to facilitate the interaction with other systems. The current version of the semantic model includes mappings to the following four vocabularies: SNOMED CT, NCI Thesaurus, the Gene Regulation Ontology, and the Sequence Ontology. The use of these vocabularies was advantageous at various levels, but was not challenge-free. The solutions found resulted in a model that contains: mappings to more than one vocabulary; new concepts, not previously represented in any of the vocabularies; a minimum set of concepts necessary to describe the disease and to map it to external vocabularies; a hierarchical organization that results from more than one vocabulary. The solutions devised resulted from a compromise between the representations provided by the vocabularies and the vision of the domain conveyed by the biomedical experts that assisted in the development of the semantic model.

The model has been continuously evaluated in terms of correct representation of the domain of knowledge (performed by the biomedical experts) and in terms of consistency, with checks performed periodically normally after important alterations.

The semantic model is currently being populated with data from six Portuguese institutions: the *Hospitais da Universidade de Coimbra* (Coimbra), the *Centro de Cardiologia da Universidade de Lisboa*, the *Hospital da Luz* and the *Hospital de Sta. Cruz* (all three in Lisbon) provide the clinical data; the *Centro de Química Estrutural* of the *Instituto Superior Técnico of the Universidade Técnica de Lisboa* and the *Universidade Lusófona de Humanidades e Tecnologias*

provide the genetic data. Future work includes the assessment of the effectiveness of the model to deal with real data.

Acknowledgments. This work was supported by the FCT through the Multi-annual Funding Program, the doctoral grant SFRH/BD/65257/2009, the post-doctoral grant SFRH/BPD/20996/2004 and the SOMER project (PTDC/EIA-EIA/119119/2010).

References

1. Maron, B J, Maron, M S, Wigle, E D, E Braunwald: The 50-Year History, Controversy, and Clinical Implications of Left Ventricular Outflow Tract Obstruction in Hypertrophic Cardiomyopathy: from Idiopathic Hypertrophic Subaortic Stenosis to Hypertrophic Cardiomyopathy. J. Am. Coll. Cardiol. 54, 191–200 (2009)
2. Alcalai, R., Seidman, J.G., Seidman, C.E.: Genetic Basis of Hypertrophic Cardiomyopathy: from Bench to the Clinics. J. Cardiovasc. Electrophysiol. 19, 104–110 (2008)
3. Harvard Sarcomere Mutation Database, `http://genepath.med.harvard.edu/~seidman/cg3/`
4. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: Tool for the Unification of Biology. Nat. Genet. 25, 25–29 (2000)
5. Systematized Nomenclature of Medicine-Clinical Terms (SNOMED), `http://www.ihtsdo.org/snomed-ct/`
6. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Sci. Am., 29–37 (2001)
7. World Wide Web Consortium, `http://www.w3.org/`
8. RDF Primer, `http://www.w3.org/TR/2004/REC-rdf-primer-20040210/`
9. OWL Web Ontology Language Current Status, `http://www.w3.org/standards/techs/owl#w3call`
10. Gudivada, R.C., Qu, X.A., Chen, J., Jegga, A.G., Neumann, E.K., Aronow, B.J.: Identifying Disease-Causal Genes Using Semantic Web-based Representation of Integrated Genomic and Phenomic Knowledge. J. Biomed. Inform. 41, 717–729 (2008)
11. Agorastos, T., Koutkias, V., Falelakis, M., Lekka, I., Mikos, T., Delopoulos, A., Mitkas, P.A., Tantsis, A., Weyers, S., Coorevits, P., Kaufmann, A.M., Kurzeja, R., Maglaveras, N.: Semantic Integration of Cervical Cancer Data Repositories to Facilitate Multicenter Association Studies: the ASSIST Approach. Cancer Inform. 8, 31–44 (2009)
12. Colombo, G., Merico, D., Boncoraglio, G., Paoli, F.D., Ellul, J., Frisoni, G., Nagy, Z., van der Lugt, A., Vassanyi, I., Antoniotti, M.: An Ontological Modeling Approach to Cerebrovascular Disease Studies: the NEUROWEB Case. J. Biomed. Inform. 43, 469–484 (2010)
13. Noy, N. F., McGuinness, D. L.: Ontology Development 101: A Guide to Creating Your First Ontology. Technical report number KSL-01-05, Knowledge Systems, AI Laboratory, Stanford University (2001)
14. Protégé Ontology Editor, `http://protege.stanford.edu`

15. Hermit OWL Reasoner, http://hermit-reasoner.com/
16. Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey M-A, Chute, C. G., and Musen, M. A.: BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse. Nucleic Acids Res. 37, W170-W173 (2009)
17. Machado, C.M., Couto, F., Fernandes, A.R., Santos, S., Cardim, N., Freitas, A.T.: Semantic Characterization of Hypertrophic Cardiomyopathy Diseases. In: First Workshop on Knowledge Engineering, Discovery and Dissemination in Health (KEDDH10) (2010)
18. Sioutos, N., Coronado, S., Haber, M.W., Hartel, F.W., Shaiu, W.L., Wright, L.W.: NCI Thesaurus: a Semantic Model Integrating Cancer-Related Clinical and Molecular Information. J. Biomed. Inform. 40, 30–43 (2007)
19. The Ontology of Clinical Research (OCRe), http://rctbank.ucsf.edu/home/ocre.html
20. Beisswanger, E., Lee, V., Kim, J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S., Hahn, U.: Gene Regulation Ontology (GRO): Design Principles and Use Cases. St. Heal. T. 136, 9–14 (2008)
21. Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R., Ashburner, M.: The Sequence Ontology: a Tool for the Unification of Genome Annotations. Genome Biol 6, R44 (2005)
22. Jonquet, Clement, Musen, Mark A., Shah, Nigam H.: Building a Biomedical Ontology Recommender Web Service. J. Biomed. Semantics 1(Suppl 1), S1 (2010)
23. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. Nat. Biotech. 25, 1251–1255 (2007)
24. Gersh, B.J., Maron, B.J., Bonow, R.O., Dearani, J.A., Fifer, M.A., Link, M.S., Naidu, S.S., Nishimura, R.A., Ommen, S.R., Rakowski, H., Seidman, C.E., Towbin, J.A., Udelson, J.E., Yancy, C.W.: 2011 ACCF/AHA Guideline for the Diagnosis and Treatment of Hypertrophic Cardiomyopathy: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation 124, e783–e831 (2011)