LEVERAGING ELECTRONIC HEALTHCARE RECORD STANDARDS AND SEMANTIC WEB TECHNOLOGIES FOR THE IDENTIFICATION OF PATIENT COHORTS

Jesualdo Tomás Fernández-Breis[1+*], José Alberto Maldonado[2+], Mar Marcos[3+], María del Carmen Legaz-García[1], David Moner[2], Joaquín Torres-Sospedra[3], Angel Esteban-Gil[4], Begoña Martínez-Salvador[3], Montserrat Robles[2]

+ These authors have contributed equally to this work

[1]Departamento de Informática y Sistemas, Universidad de Murcia, 30100, Murcia, Spain

[2] Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia

[3] Dept. of Computer Engineering and Science, Universitat Jaume I, Av. de Vicent Sos Baynat s/n, 12071 Castellón, Spain

[4]Fundación para la Formación e Investigación Sanitaria, C/ Luis Fontes Pagán nº 9 - 1ª planta, 30003 Murcia, Spain

* Corresponding author: Jesualdo Tomás Fernández Breis, Departamento de Informática y Sistemas, Universidad de Murcia, 30100 Murcia, Spain, phone: +34868884613, fax: +34868884151, email: jfernand@um.es

Keywords: Medical Informatics; Electronic Health Records/standards*; Semantics*, Decision Support Systems, Clinical;

DOI information: 10.1136/amiajnl-2013-001923

ABSTRACT

Introduction

The secondary use of Electronic Healthcare Records (EHRs) often requires the identification of patient cohorts. In this context, an important problem is the heterogeneity of clinical data sources, which can be overcome with the combined use of standardized information models, Virtual Health Records, and semantic technologies, since each of them contributes to solving aspects related to the semantic interoperability of EHR data. Our main objective is to develop methods allowing for a direct use of EHR data for the identification of patient cohorts leveraging current EHR standards and semantic web technologies.

Materials and Methods

We propose to take advantage of the best features of working with EHR standards and ontologies. Our proposal is based on our previous results and experience working with both technological infrastructures. Our main principle is to perform each activity at the abstraction level with the most appropriate technology available. This means that part of the processing will be performed using archetypes (i.e., data level) and the rest using ontologies (i.e., knowledge level). Our approach will start working with EHR data in proprietary format, which will be first normalized and elaborated using EHR standards and then transformed into a semantic representation, which will be exploited by automated reasoning.

Results

We have applied our approach to protocols for colorectal cancer screening. The results comprise the archetypes, ontologies and datasets developed for the standardization and semantic analysis of EHR data. Anonymized real data has been used and the patients have been successfully classified by the risk of developing colorectal cancer.

Conclusion

This work provides new insights in how archetypes and ontologies can be effectively combined for EHR-driven phenotyping. The methodological approach can be applied to other problems provided that suitable archetypes, ontologies and classification rules can be designed.

INTRODUCTION

Objective

Our main goal is providing methods allowing for a direct utilization of data from electronic health records (EHRs) in the process of identification of patient cohorts. Leveraging of current EHR standards and semantic web technologies is also regarded as an important objective in this work.

Background and significance

With the increasing adoption of EHRs there is a growing interest in methods to enable the secondary use of EHR data, e.g. in clinical research. This secondary use often involves the identification of patient cohorts from EHR data (or EHR-driven phenotyping), which is an expensive and time-consuming process. According to recent reviews [1], there are many publications reporting on automated systems to facilitate this task. Most of these works rely on proprietary formats for data integration, and rarely use EHR interoperability standards like HL7, openEHR, or ISO13606. In this context, an important problem is the heterogeneity of clinical data sources, which usually differ in the data models, naming conventions, and degree of detail for representing similar data [2]. Another problem related to clinical data sources is the "impedance mismatch" [3] that usually exists between EHR data and the data required by the EHR-driven phenotyping algorithms, at a rather high level of abstraction.

There is evidence that the use of standardized information models can help to solve the integration problem of clinical data sources. Some initiatives use a Virtual Health Record (VHR) over the set of local EHR systems to overcome the aforementioned problems [4,5,6]. The VHR includes a generic information model potentially capable of representing a wide range of clinical concepts, and a query language. Standardization of the VHR is regarded as an important issue. Consequently, several works have based their VHR on standard EHR architectures. However, the use of a VHR based on a standard EHR architecture is not sufficient for semantic interoperability in the context of EHR-driven phenotyping. The main problem is the partitioning of concepts between the information and domain models. To solve this problem it is necessary to make explicit all the assumptions about the representation of data. Thus, specific domain concept definitions are needed rather than the generic concepts provided by EHR architectures. Examples of such definitions are openEHR/ISO13606 archetypes [7], CDA templates, and detailed clinical models [8]. Currently, the Clinical Information Modeling Initiative (CIMI) [9] is working on providing a common format for the definition of health information content.

In addition, there is an increasing use of semantic web technologies for managing EHR information and knowledge. The reason for this is the potential of technologies like OWL [10], which enables a formal representation of the domain information entities and knowledge that can be exploited by automated means. In line with this, important international initiatives [11,12] consider that semantic web technologies are fundamental to achieve consistent and meaningful representation, access, interpretation and exchange of EHR data. To mention some examples, EHR standards have been represented by means of OWL ontologies with different purposes [13,14,15]. OWL technologies make automated reasoning possible, which has been exploited in the validation of clinical models [16,17], and for reasoning over EHR data [18,19]. There are also numerous studies making use of ontologies for biomedical data

integration [20]. One of the problems identified is the availability of ontologies corresponding to the needs of a specific application.

With the purpose of providing methods allowing for the smooth execution of EHR-driven phenotyping algorithms, in this work we propose to leverage domain concept definitions based on standard EHR architectures, on the one hand, and ontology-based descriptions of inclusion/exclusion criteria with the potential for automated reasoning, on the other hand. Our proposal is to use archetypes for the former and the OWL formalism for the latter. Essential to our proposal, a set of archetype-based concept models of the kind of a VHR will serve to solve the integration and mismatch problems inherent to the direct utilization of EHR data by phenotyping algorithms. Additionally, OWL ontologies will ensure a precise characterization of these algorithms, with the added value of automated support via classification reasoning, which can be of great help in the process of identification of patient cohorts.

MATERIAL AND METHODS

Our methodological approach takes advantage of the best features of EHR standards and ontologies, at data and knowledge levels, respectively. Our main principle is to perform each activity at the abstraction level with the most appropriate technology available. This means that part of the processing will be performed using archetypes (i.e., data level) and the rest using ontologies (i.e., knowledge level). Our approach (see Figure 1) assumes that the EHR data are stored in a proprietary format in the database of the clinical institution. EHR data undergo a transformation pipeline, where the first step is a pre-processing to convert the relational data instances into XML extracts that can be readily used in the next step.

Phenotyping algorithms usually require performing a series of arithmetic and logical operations (or abstractions) on the data, and subsequently reasoning using these more elaborated data. The second step of our approach deals with the former processing using archetypes. In this work, we will use openEHR archetypes [21], however the same approach could be applied to other EHR standards. In this step the XML extracts will be converted into normalized EHR data compliant with the underling EHR architecture, and then this data will be transformed to meet the requirements of the phenotyping algorithm. For instance, if the algorithm uses the number of adenomas of a patient, first the information about each adenoma finding would be normalized, and afterwards the count would be calculated. Accordingly, a series of archetypes will be necessary to abstract from the raw data to the normalized data to be processed by the phenotyping algorithm. The corresponding abstractions and transformations will be carried out via archetype mappings. The design of this archetype layer (or phenotyping archetype) is specific for the particular phenotyping algorithm, while it is the result of a trade-off between reusability and simplicity. In the phenotyping archetype, a distinction can be made between first-level archetypes and second (and so on) level archetypes, depending on whether their value can be obtained from the EHR data or, on the contrary, require data that are not directly available in the EHR. The outcome of this second step is a collection of archetype-compliant data instances suitable for being consumed by the next steps.

The objective of the third step is the transformation of the data instances into a semantic representation. So far, the clinical data of a patient are represented using archetype-compliant XML extracts. This representation has limitations with regard to expressing domain

knowledge, and therefore neither supports inference nor can be used to perform reasoning as required by phenotyping algorithms (see above). Our proposal is that the inclusion/exclusion criteria defined in the phenotyping algorithm are implemented in OWL, and that the classification is performed through automated reasoning. To achieve this, a mapping is required between the phenotyping archetype and a domain ontology, which needs to cover the concepts and properties of the domain in question. This ontology can be built on purpose or re-used in case quality ontologies for the domain exist. Depending on the scope of the study, extensions to existing ontologies are likely to be needed, and/or a network of ontologies rather than a single one can be required. Once this step is performed, the data will be available in a formalism for which automated classification and reasoning are natural tasks.

The fourth step of our approach requires enriching the domain ontology with appropriate classification rules, so that a reasoner can automatically compute the groups of patients as well as guarantee the consistency and logical correction of the ontology. The output of this step will be the clinical group associated with each patient. Key methods of the data transformation pipeline are summarized in Tables 1 and 2, and detailed in Annex I.

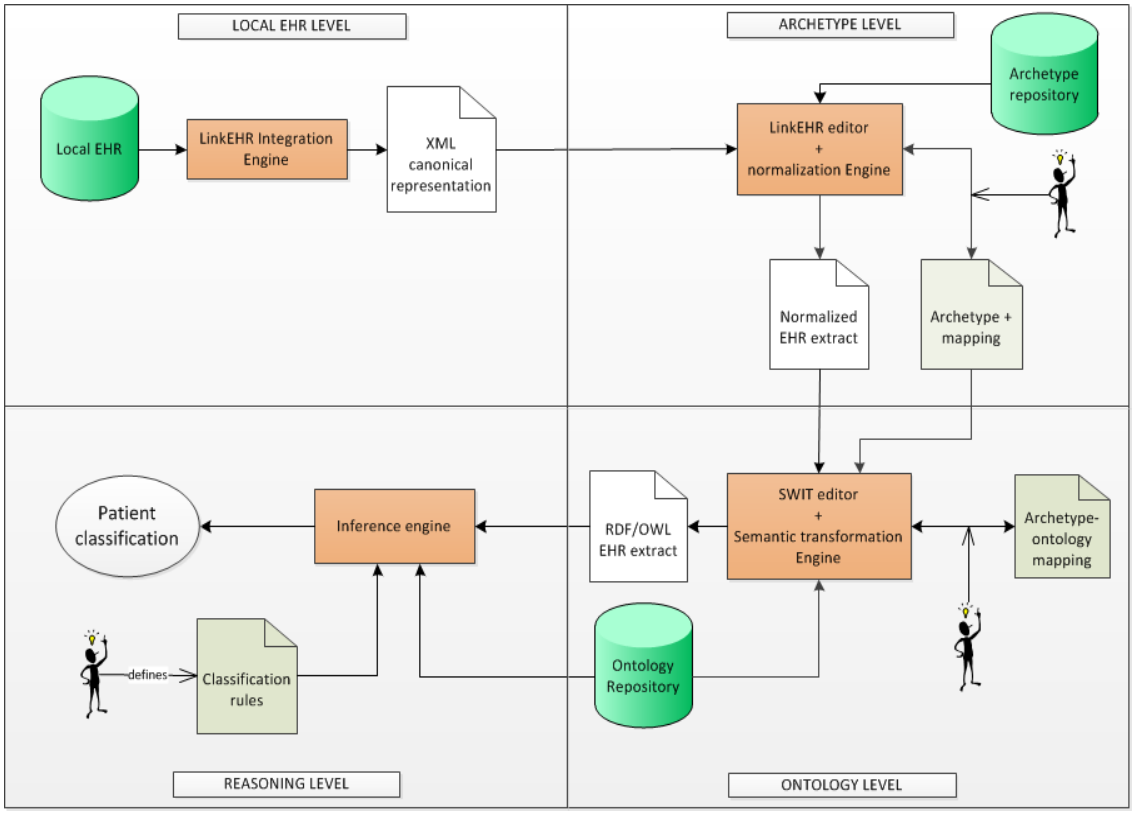

Figure 1. Overview of the methodological approach

Table 1. Summary of activities performed at archetype level

Clinical concept modeling using archetypes (archetype level)	
Goal	Development of the phenotyping archetype for the normalization and abstraction of the EHR data to be used in the phenotyping algorithm
Input	• Documentation about the domain and phenotyping algorithm (e.g. medical encyclopedias, clinical guidelines) • Terminological resources (e.g. SNOMED CT [22]) • Archetype repositories
Output	• (Semi)formal specification of domain concepts • Phenotyping archetype
Tasks	• Analysis and specification of domain concepts • Design and development of phenotyping archetype
Tools	• UMLS Terminology Services [23] • openEHR Clinical Knowledge Manager [24] • LinkEHR mapping module [7]
From EHR data to archetypes (archetype level)	
Goal	Transformation of EHR data into archetype-compliant normalized EHR extracts
Input	• Source EHR schemas and data • (Semi)formal specification of domain concepts • Phenotyping archetype
Output	• Specification of EHR data-archetype mapping • Set of XQuery scripts (one for each archetype) implementing the mappings • EHR data expressed as archetype-compliant XML documents
Tasks	• Definition of high-level mappings between source schemas (local schemas or archetypes) and phenotyping archetype. • Compilation of high-level mappings into XQuery scripts. • Execution of XQuery scripts on EHR data and/or archetype instances.
Tools	• LinkEHR archetype editor [25] • Saxon [26]

Table 2. Summary of the activities performed at ontology level

Domain knowledge modeling using ontologies (ontology level)	
Goal	Development of the ontologies for representing the domain knowledge
Input	• (Semi) formal specification of domain concepts • Repositories of ontologies
Output	• Set of ontologies representing the domain knowledge
Tasks	• Selection of existing ontologies appropriate for being reused • Development of the OWL ontologies by extension of selected ontologies or from scratch
Tools	• Bioportal [27] • Protégé [28]
From archetyped data to OWL (ontology level)	
Goal	Transformation of the normalized EHR extracts into a semantic representation to facilitate further processing and exploitation.
Input	• EHR data expressed as archetype-compliant XML documents • Set of ontologies representing the domain knowledge • Phenotyping archetype
Output	• Specification of archetype-ontology mapping • Set of OWL individuals representing the normalized EHR extracts
Tasks	• Definition of the mappings between the phenotyping archetype and the domain ontology • Application of the mappings to the normalized EHR extracts
Tools	• SWIT mapping and transformation modules [29]
OWL reasoning (ontology level)	
Goal	Design and application of the phenotyping algorithm to the OWL individuals
Input	• Set of OWL individuals representing the normalized EHR extracts • Set of ontologies representing the domain knowledge • Specification of the phenotyping algorithm
Output	• OWL ontology that implements the phenotyping algorithm • Classification of the OWL individuals according to the phenotyping algorithm
Tasks	• Implementation of the phenotyping algorithm in OWL • Application of the phenotyping algorithm through automated reasoning

	• Querying of the knowledge base to retrieve the classification of each subject
Tools	• Protégé [28] • Hermit [30] • OWLAPI [31]

RESULTS

Case study description

Colorectal cancer is one of the most important mortality causes in many developed countries according to Global Burden of Disease study, with an expected increase in incidence for the coming years [32]. Developing effective mechanisms for the early detection of colorectal cancer would contribute to a better management and control of this disease. Our case study focuses on the program for colorectal cancer screening in the Murcia region (Spain). To date, the physicians involved in this program apply the screening protocols drawn from the European and American guidelines (see [33]) to classify patients in levels of risk, and, according to such classifications, make clinical decisions. As result, a database recording clinical data and decisions taken has been compiled. This database, with data about more than 20,000 patients, is the source of the anonymized EHR data used in our study. Our hypothesis is that our approach can help physicians in their activity by suggesting the classification of the patients according to their risk of colorectal cancer.

Next, we summarize our main results. More information, including the archetypes, ontologies, mappings and datasets, is available at http://miuras.inf.um.es/colorectal.

Archetype infrastructure and mapping

In our case study, we started by analyzing the archetypes in the openEHR Clinical Knowledge Manager to identify suitable archetypes. The most suitable one was openEHR-EHR-OBSERVATION.lab_test-histopathology, that models a generic anatomical pathology or histopathology test. In order to accommodate the additional concepts required by our phenotyping algorithm, the archetype was specialized using the LinkEHR archetype editor. The specialization, named openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening, incorporates detailed information about adenoma findings, such as type, maximum size of the recorded dimensions (width, breadth and height), dysplasia grade, and whether they are sessile and/or advanced. Our case study requires concepts at different level of granularity: at finding and study levels. In order to represent study level concepts (maximum size of all adenomas and number of adenomas) we developed a second-level archetype (openEHR-EHR-EVALUATION.colorectal_screening.v1) from scratch.

For the generation of archetype instances we need to access the required EHR data and then to transform them into archetype instances. Firstly, the data access module of LinkEHR was used to generate a canonical XML view over the EHR system. This XML view was then used as source schema in the mapping of the openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening archetype, since their instances can be obtained directly from EHR data. When a local EHR is involved in a mapping scenario, the mapping specification requires a clear understanding of the source schema, thus we worked closely

with the database administrator. Figure 2 shows an example of value correspondence used to map the first-level archetype to the local EHR. Finally, we defined the mapping between the first- and second-level archetypes. In this case and due to the presence of aggregation functions we employed a structural mapping [7] to control the grouping context. An excerpt of this mapping is shown in Figure 3. Since the source path of the structural mapping is the root entity of the first-level archetype the counting of adenomas and the calculation of the maximum size of adenomas is done at study level. With this approach we were able to validate at each step the XQuery script [34] generated by the LinkEHR mapping tool. Figure 6 illustrates an example of the data transformations applied to the adenoma dimensions (length, width and depth) during the whole process, i.e. from local data to OWL instances. As can be observed, the canonical XML document is transformed into an XML instance of the finding (first-level) archetype. The mapping in Figure 2 is used in this transformation. Concretely it is employed to assign the value to the *maxsize* element (archetype_node_id=”at0.0.31”) with the maximum size of any recorded dimension of a particular finding. By contrast, the mapping between the first- and second-level archetypes calculates the maximum size of any recorded dimension only for adenomas and at study level. For this purpose, a mapping with a similar structure as the one displayed in Figure 3 was used.

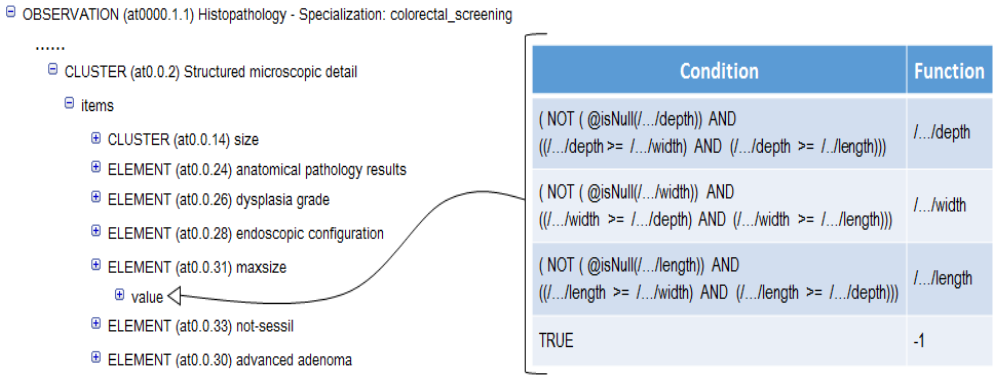

Figure 2. Value correspondence for calculating the maximum size (depth, width, or length) of an adenoma finding. As it can be observed, if none of the dimensions has been recorded, the value -1 is assigned

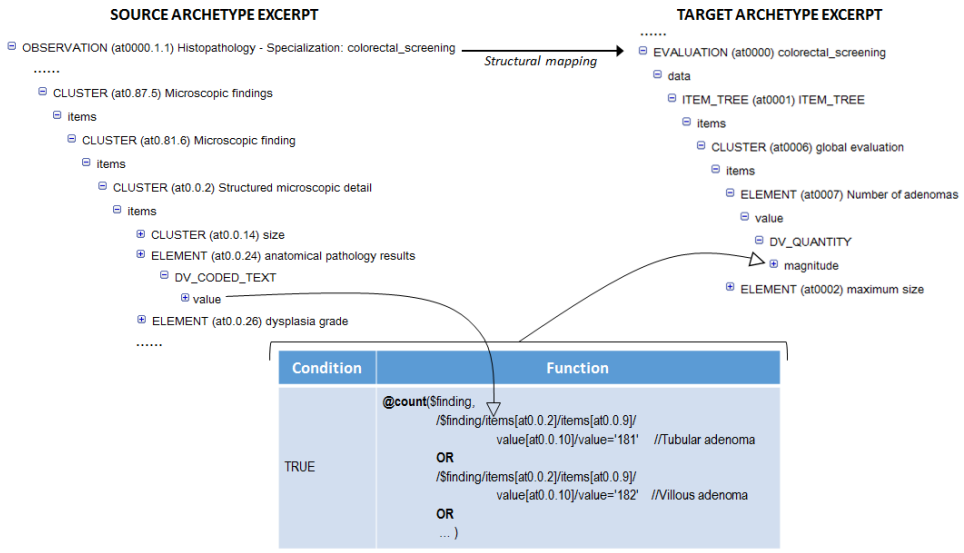

Figure 3. Calculation of the number of adenomas in a histopathology study. The variable $finding represents the path of microscopic findings (CLUSTER at0.81.6) in the source archetype. The complete source instance provides the context for grouping and counting

OWL infrastructure and reasoning

The starting point was a domain ontology developed by our clinical partners, called precol, which was designed for the data management activities they perform rather than for supporting automated reasoning. Then, we inspected Bioportal ontologies looking for more appropriate formalizations of the concepts and the inferencing capabilities required by the phenotyping algorithm. Given our goal, we decided that the best option was to re-engineer the classes of the precol ontology over which reasoning is to be performed using Protégé. The ontological infrastructure includes different ontologies for representing domain entities (colorectal-domain), the rules for determining the risk level (colorectalscreening-rules), and the data (colorectal-instances). The colorectal-domain ontology extends precol by adding some classes, properties and axioms oriented to reasoning. Figure 4 shows an excerpt of the inferred taxonomy of Finding. There, we can see different types of Adenoma, each of which is defined through sufficient conditions, which is an effective way of representing the axioms for automated reasoning. The reasoner exploits this taxonomic structure to answer the queries.

Next, we manually defined the mappings between the phenotyping archetype and the colorectal-domain ontology using SWIT [29]. An excerpt of the mapping rule for *Finding* is shown in Figure 5, where the lines show the correspondence between the archetype and ontology entities. Once defined, the mappings were automatically executed on the archetyped data instances to generate the OWL dataset. It should be noted that in this study, the mapping is defined between the first- and second-level archetypes and the domain ontology. Figure 6 illustrates how the data is transformed into OWL. There, an individual of the class Histopathology Report is created. This report has two findings, whose data come from the two ELEMENTs defined in the Histopathology colorectal screening archetype (top-right). Besides, this report has a max_size, whose value is obtained by executing the mapping with the colorectal screening archetype (bottom-left).

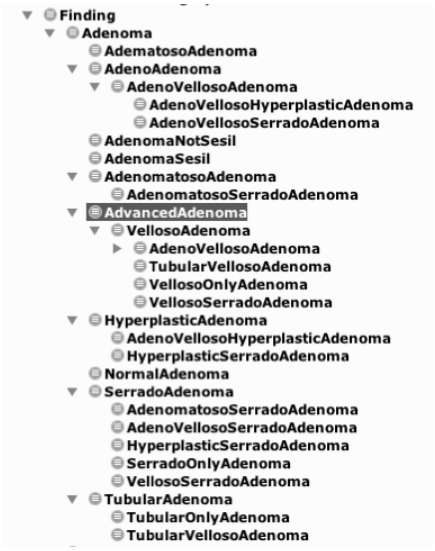

Figure 4. Excerpt of the domain ontology

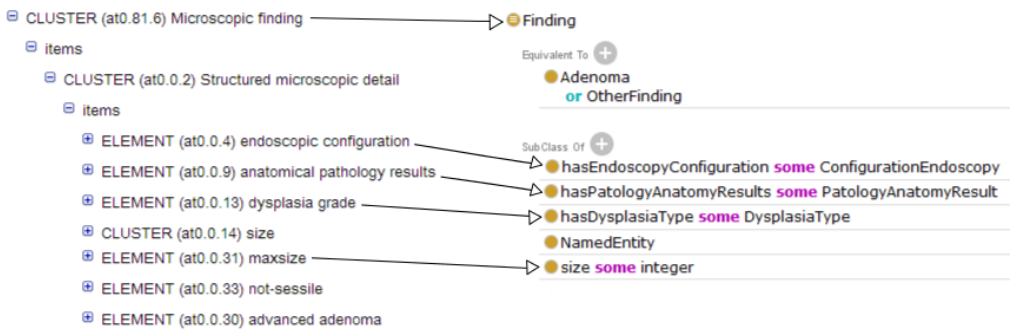

Figure 5. Partial view of the mapping for Finding

The European and American screening protocols have been implemented in the colorectal screening-rules ontology. Table 3 explains the rules defined in this ontology for low, intermediate and high risk according to the European and American protocols. Finally, OWL-DL reasoning over the OWL dataset generates the classifications according to each protocol, which are retrieved by using the OWLAPI or through DL queries [35].

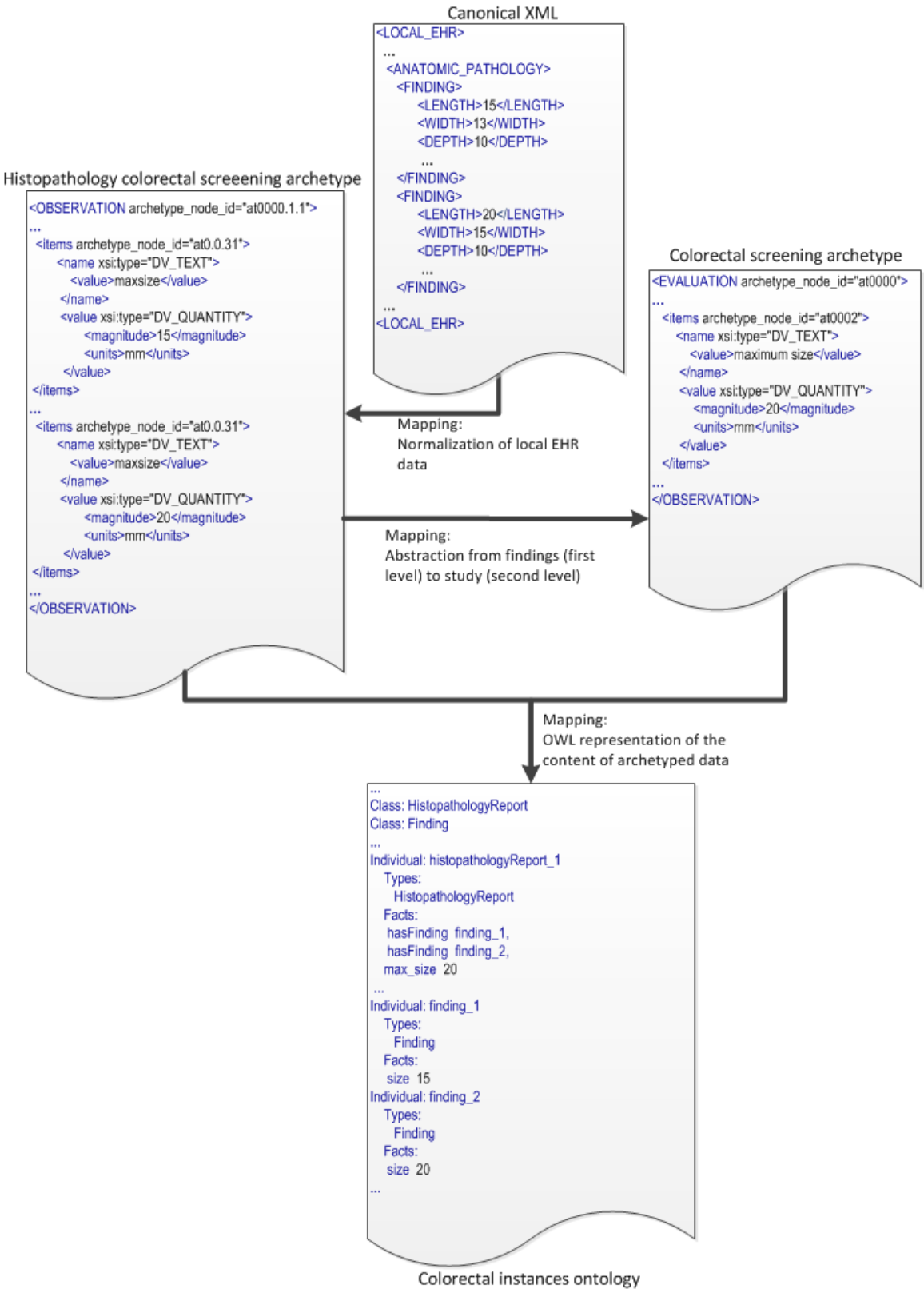

Figure 6. Example of a complete data transformation process

Table 3. The classification rules defined for the American and European protocols

Group	OWL rule and explanation
High risk American	**Rule:** *(HistopathologyReport and* *((hasAdenoma some AdvancedAdenoma) or (number some integer[>= 3])))* *or (HistopathologyReport and (max_size some integer[>= 20]))*
	Explanation: A histopathology report whose findings describe at least one advanced adenoma or at least 3 adenomas, or an histopathology report whose largest adenoma has at least 20mm The domain ontology contains the properties that an adenoma must meet to be classified as advanced by the reasoner Number represents the amount of adenomas described in the histopathology report. This value is calculated in the archetype layer. This could have been calculated in the ontology layer but that would require more time for reasoning max_size represents the size of the largest adenoma and this value is calculated in the archetype layer, since it cannot be easily calculated in the ontology layer
High risk European	**Rule:** *(HistopathologyReport* *and ((max_size some integer[>= 20])* *or (number some integer[>= 5])))*
	Explanation: A histopathology report whose findings describe an advanced adenoma of size equal or greater than 20mm or at least 5 adenomas For optimization purposes, the condition *has Adenoma some (AdvancedAdenoma and (size some integer[>=20]))* is not included, since it is guaranteed by *max_size some integer[>= 20]*
Intermediate risk European	**Rule:** *HistopathologyReport*

	and (((hasAdenoma some AdvancedAdenoma) *and (max_size some integer[< 20])* *and (number some integer[< 5]))* *or ((hasAdenoma only NormalAdenoma)* *and (number some integer[> 2])* *and (number some integer[< 5])))*
	Explanation: A histopathology report that meets one of the following conditions a) It contains less than 5 adenomas, at least 1 of which is advanced, and the size of the largest adenoma is less than 20mm b) It contains 3 or 4 normal adenomas It should be noted that normal adenoma and advanced adenoma are disjoint classes.
Low risk European /American	**Rule:** *HistopathologyReport* *and (hasAdenoma only NormalAdenoma)* *and (number some integer[< 3])*
	Explanation: A histopathology report that only contains at most 2 normal adenomas, but does not contain any advanced one. This rule is the same for both protocols.

Evaluation

We have carried out an overall evaluation using a small selection of 33 histopathology reports from the database. These reports were selected to cover a wide range of value combinations. Based on the mappings specifically designed for this purpose, appropriate archetype and OWL instances have been generated. OWL-DL reasoning was applied to classify the histopathology reports according to the European protocol. The classification results matched the results obtained by manually applying the protocol in 100% cases.

In addition, we have evaluated the performance of the main steps of our data transformation pipeline, namely the archetype level mapping, the ontology level mapping, and the OWL reasoning. In this evaluation we used a bigger number of histopathology reports (503 reports), randomly selected. Default values provided by our expert were used in case of missing data,

e.g. a missing value in the dysplasia type was interpreted as a low-grade dysplasia. To evaluate the performance of the mapping steps we tested the instance generation scripts to analyze the response time thereof. The mean time required for the archetype mapping step was 13 milliseconds per report, and about 150 milliseconds per report for the ontology mapping step. Regarding the performance of the reasoning step, the mean time to classify each report using Hermit was 2.1 seconds.

Table 4. Discrepancy matrix

		Database classification		
		High-risk	Intermediate-risk	Low-risk
OWL classification	High-risk	69	5	2
	Intermediate-risk	102	44	24
	Low-risk	26	17	206

Finally, we have compared the results of the OWL classification step with the classifications done by physicians as recorded in the original database (see Table 4). Despite the default values, the reasoner did not yield any classification result for a small number of reports (8 reports), due to data values not covered by the classification rules (e.g. dysplasia type with "could not be determined" as value). Focusing on the reports for which the reasoner yielded a classification, this classification matches the database one in a 64.4% of the cases. Among the discrepancies (35.6% of the cases), a 58% corresponds to reports classified as high-risk by physicians and as intermediate-risk by the OWL classifier. This suggests that physicians may tend to assign a higher risk level when compared to the protocol. Note that discrepancies with respect to the protocol do not necessarily imply a non-compliance, as physicians were not supposed to follow it. The list of discrepant cases and the corresponding data files are available in http://miuras.inf.um.es/colorectal. A sample of 17 cases from this list was presented to the physician for revision, selected among discrepancies involving low risk versus the other 2 classifications. The physician determined that the OWL classification was the correct one in 100% cases. Analyzing the reasons for the misclassifications in the database remains as future work.

DISCUSSION

In this work, we have presented a novel method to support EHR-driven phenotyping that combines EHR standards, archetypes, ontologies and reasoning. The novelty of the approach lies in how these technologies are combined for taking full advantage of their benefits. On the one hand, archetypes define semantically rich data structures based on EHR standards. They abstract from how the data are stored in a particular EHR system and, therefore, provide a meaningful way for exchanging healthcare data. Our approach considers the use of an archetype-based VHR to normalize and further elaborate EHR data (at the data level) so that the requirements of phenotyping algorithms can be met. On the other hand, ontologies serve

to provide a formal specification of domain knowledge for phenotyping purposes. Current languages like OWL make automated reasoning possible such that new knowledge can be inferred. Our approach uses OWL ontologies and classification reasoning (at the knowledge level) for the bulk of phenotyping algorithms.

Phenotyping algorithms require working at both data and knowledge levels, since the inclusion/exclusion criteria are calculated from raw EHR data but usually also need data not directly available in the EHR. E.g. in our case study, the classification of a patient in the "high risk" group requires to find either one advanced adenoma or at least three adenomas. In turn, the classification of an adenoma as "advanced" is done based on the specific value ranges for the size, histology and dysplasia of the adenoma. We have addressed the classification tasks at the knowledge level, and processing tasks such as counting and negation (e.g. if an adenoma is "not sessile") at the data level. The separation of concerns between data and knowledge levels is not a clear-cut issue. The decision will depend on the particular application as well as on the features of the representation language chosen for the knowledge level. To explore this issue, for our case study we have developed an archetype infrastructure which is able to determine directly at data level e.g. if an adenoma is advanced. This has been done to facilitate the reusability of the archetype infrastructure in platforms other than OWL. As criterion, it can be considered to include a domain concept definition in a particular level according to the expected reuse of the corresponding artifact (ontology or archetype).

Reuse is one of our main interests, of both archetypes and ontologies. The reusable archetype infrastructure provides standardized access to clinical data, possibly coming from different EHR systems, by just defining the necessary mappings between the source databases and the first-level archetypes (note that the mappings for second and higher-level archetypes can be fully reused). Coupled with the archetypes, the ontology infrastructure (including definitions and mappings) can also be reused to work with different EHR systems. The definitions in the ontology themselves can indeed be reused to a large extent, e.g. in the use case we have used them to define the rules for both the American and European protocols without modifying the archetype infrastructure.

We use OWL classes with sufficient conditions for defining the categories of interest related to the phenotyping algorithm. This means that those classes represent explicitly the knowledge on the categories as OWL content and, therefore, can be combined with additional knowledge for further studies and inference. Alternative representations like SPARQL [36] could have been used. In that case, however, the conditions of each category of interest would be embedded within queries, with no possibility of exploitation as separate knowledge. Besides, the SPARQL inference possibilities based on properties are limited. E.g. it would not be possible to identify which findings are advanced adenomas by just using SPARQL, unless the queries replicated all the conditions for a finding to be classified as such. In that case, the most reasonable option would be to reason over the OWL content first, to obtain all the finding classifications from the data, and then issue the SPARQL queries against the inferred knowledge base.

The use of OWL is appropriate for phenotyping algorithms where the classification of patients is based on the analysis of individual patient features, rather than e.g. features of relatives. Otherwise, some options are: (1) dealing with the problem at the data level; and (2) implementing those rules with languages such as SWRL[37] or SPARQL and combine them

with OWL reasoning. The decision must be made based on the particular problem. Examples of the first solution can be seen in our case study (see Table 3), since operations like count, maximum, and negation cannot be easily performed using OWL reasoning. Our solution was to perform such activities at the data level, and the time performance obtained suggests that our modeling decisions have been effective. Consequently, our recommendation is to carefully analyze the requirements of the phenotyping algorithm, taking also into account the reuse prospects of the archetypes/ontologies (see above).

Terminologies should play an important role to bridge the gap between EHR data and archetypes, and between archetypes and ontologies. In previous works we have managed SNOMED-CT content in both EHR-archetype [38,39] and archetype-ontology [17] transformation layers. In our view the exploitation of the terminological bindings defined in the archetypes and the increasing availability of terminologies in a processable form should be helpful not only to drive the transformation process but also for the semi-automatic generation of the required mappings. However, for simplicity, in this research work we have opted for disregarding terminological issues, focusing instead on the combination of EHR standards and semantic web technologies.

CONCLUSION

This work provides new insights in how archetypes and ontologies can be effectively combined for EHR-driven phenotyping. The main contribution of this work is the methodological proposal which describes how those technologies can be combined and how we can take full advantage of their benefits. This is a progress with respect to our previous works because, we had used ontologies for representing archetypes and data to facilitate clinical data and models interoperability [7] and used knowledge-rich clinical models based on archetypes to link clinical decision support systems with EHR systems [38,39], but never combined in an effective data analysis pipeline as proposed here.

With respect to related work, we are aware of the MobiGuide and EURECA projects [40,41], which focus on linking clinical decision support with EHR systems with the help of archetypes and/or semantic web technologies. However, to the best of our knowledge, none of these approaches has achieved the level of semantic interoperability and integration we demonstrate in this work. The approach of the SHARPn project [42] is similar to ours, and has proved to be effective in a platform for the secondary use of EHR data. The major differences are the use of Clinical Element Models [43] (instead of archetypes), the rule-based description of phenotyping algorithms, and the processing of textual EHR data. Compared to SHARPn, our approach does not cover the latter aspect. However it outstands for the clinical models used, allowing for the standardized representation of rather abstract clinical concepts, as opposed to raw EHR data.

The approach is rather generic and can be applied to other problems provided that suitable archetypes, ontologies and classification rules can be developed. Moreover, the approach promotes and emphasizes on the use of international standards and recommendations like openEHR/ISO13606 archetypes and OWL. As future work we envisage to perform a more principled clinical validation of our results. Furthermore, we plan to incorporate the issues of terminological knowledge and ontology alignment in our approach.

Acknowledgments

We thank the "Programa de Prevención del Cáncer de Colon y Recto de la Región de Murcia" for providing the data for performing this study.

Competing interests

We declare having no competing interests.

Contributorship Statement

JTFB, JAM and MM have conceived and designed the study, participated in the technical development and experimental validation, and have been the main contributors to the manuscript (these authors have contributed equally to this work). MCLG, AEG, DM, JTP, BMS, MR have contributed to the technical developments and experimental validation, and have critically revised the manuscript.

Funding

This work was supported by the Ministerio de Economía y Competitividad and the FEDER programme through grants TIN2010-21388-C01 and TIN2010-21388-C02. MCLG was supported by the Fundación Séneca through grant 15555/FPI/2010.

REFERENCES

1. Cuggia M, Besana P, Glasspool P. Comparing semi-automatic systems for recruitment of patients to clinical trials. International Journal of Medical Informatics 2011; **80(6)**: 371-388.
2. Sujansky W. Heterogeneous database integration in biomedicine. Journal of Biomedical Informatica. 2001; **34(4)**: 285-298.
3. Schadow G, Russler DC, Mead CN, McDonald CJ. Integrating medical information and knowledge in the HL7 RIM. Proceedings of the AMIA Symposium 2000: 764-768.
4. Johnson PD, Tu SW, Musen MA, Purves I. A virtual medical record for guideline-based decision support. Proceedings of the AMIA 2001Annual Symposium: 294–298.
5. German E, Leibowitz A, Shahar Y. An architecture for linking medical decision-support applications to clinical databases and its evaluation. Journal of Biomedical Informatics 2009;**42**:203–218.

6. Peleg M, Keren S, Denekamp Y. Mapping computerized clinical guidelines to electronic medical records: Knowledge-data ontological mapper (KDOM). Journal of Biomedical Informatics 2008; **41**:180–201.
7. Maldonado JA, Martínez-Costa C, Moner D, Menárguez-Tortosa M, Boscá D, Miñarro-Giménez JA, Fernández-Breis JT, Robles M. Using the ResearchEHR platform to facilitate the practical application of the EHR standards. Journal of Biomedical Informatics 2012;**45**:746-762.
8. Parker CG, Rocha RA, Campbell JR, Tu SW, Huff SM. Detailed clinical models for sharable, executable guidelines. Studies in Health Technology and Informatics 2004;**107**:145-148.
9. Clinical Information Modeling Initiative, http://informatics.mayo.edu/CIMI/index.php/Main_Page (accessed: June 2013)
10. W3C, OWL2 Web Ontology Language. http://www.w3.org/TR/owl2-overview/ (accessed: June 2013)
11. European Commission. Semantic interoperability for better health and safer healthcare. Deployment and research roadmap for europe. ISBN-13: 978-92- 79-11139-6; 2009.
12. SemanticHealthNet. http://www.semantichealthnet.eu/ (accessed: June 2013)
13. Martínez-Costa C, Menárguez-Tortosa M, Fernández-Breis JT, Maldonado JA. A model-driven approach for representing clinical archetypes for Semantic Web environments. Journal of Biomedical Informatics 2009;**42(1)**:150–64
14. Iqbal A. An OWL-DL ontology for the HL7 reference information model. Toward Useful Services for Elderly and People with Disabilities 2011:168–75
15. Tao C, Jiang G, Oniki TA, Freimuth RR, Zhu Q, Sharma D, Pathak J, Huff SM, Chute CG. A semantic-web oriented representation of the clinical element model for secondary use of electronic, Journal of the American Medical Informatics Association 2013;**20:3**;554-562
16. Heymans S, McKennirey M, Phillips J. Semantic validation of the use of SNOMED CT in HL7 clinical documents. Journal of Biomedical Semantics 2011;**2**:2.

17. Menárguez-Tortosa M, Fernández Breis JT. OWL-based Reasoning Methods for Validating Archetypes. Journal of Biomedical Informatics 2013;**46**:304-317
18. Lezcano L, Sicilia MA, Rodríguez-Solano C. Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules. Journal of Biomedical Informatics 2011;**44(2)**:343-353.
19. Tao C, Wongsuphasawat K, Clark K, Plaisant C, Shneiderman B, Chute CG. Towards event sequence representation, reasoning and visualization for EHR data. In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI '12). ACM, New York, NY, USA, 801-806.
20. Bodenreider O. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. IMIA Yearbook of Medical Informatics 2008; 67-79.
21. Beale T. Archetypes. Constraint-based Domain Models for Future-proof Information Systems.http://www.openehr.org/files/publications/archetypes/archetypes_beale_web_2000.pdf
22. SNOMED-CT http://www.ihtsdo.org/snomed-ct/ (accessed: June 2013)
23. UMLS Terminology Services. https://uts.nlm.nih.gov/home.html (accessed: June 2013)
24. The openEHR Foundation, openEHR Clinical Knowledge Manager. http://www.openehr.org/knowledge/ (accessed: June 2013)
25. Maldonado JA, Moner D, Boscá D, Fernández-Breis JT, Angulo C, Robles M. LinkEHR-Ed: A multi-reference model archetype editor based on formal semantics. International Journal of Medical Informatics 2009;**78**:559–570.
26. SAXON XSLT and XQuery processor. http://saxon.sourceforge.net/ (accessed: June 2013).
27. NCBO Bioportal. http://bioportal.bioontology.org/ (accessed: June 2013)
28. The Protégé Ontology Editor and Knowledge Acquisition System. http://protege.stanford.edu/ (accessed: June 2013)
29. Semantic Web Integration Tool. http://sele.inf.um.es/swit (accessed: June 2013)
30. Hermit Reasoner. http://www.hermit-reasoner.com/ (accessed: June 2013)

31. The OWLAPI. http://owlapi.sourceforge.net/ (accessed: June 2013)
32. Institute for Health Metrics and Evaluation. Global Burden of Disease http://www.healthmetricsandevaluation.org/gbd (accessed: June 2013)
33. Segnan N, Patnick J, von Karsa L. European guidelines for quality assurance in colorectal cancer screening and diagnosis 2010. First Edition. European Union. ISBN 978-92-79-16435-4
34. W3C. XQuery 1.0: An XML Query Language. http://www.w3.org/TR/xquery/ (accessed: June 2013)
35. DL Query. http://protegewiki.stanford.edu/wiki/DL_Query (accessed: June 2013)
36. SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/ (accessed: June 2013)
37. Semantic Web Rule Language, http://www.w3.org/Submission/SWRL/ (accessed: June 2013)
38. Marcos M, Maldonado JA, Martínez-Salvador B, Boscá D, Robles M. Interoperability of clinical decision-support systems and electronic health records using archetypes: a case study in clinical trial eligibility. Journal of Biomedical Informatics 2013, http://dx.doi.org/10.1016/j.jbi.2013.05.004
39. Marcos M, Maldonado JA, Martínez-Salvador B, Moner D, Boscá D, Robles M. An Archetype-Based Solution for the Interoperability of Computerised Guidelines and Electronic Health Records. Lecture Notes in Computer Science 2011; **6747**:276-285
40. MobiGuide: Guiding patients anytime everywhere. http://www.mobiguide-project.eu/ (accessed: June 2013).
41. EURECA: Enabling information re-Use by linking clinical RE search and Care". http://eurecaproject.eu/. (accessed: June 2013)
42. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, Tao C, Parker CG, Haug PJ, Huff SM, Chute CG. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. Journal of Biomedical Informatics 2012;45(4):763-71.

43. Clinical Element Models. http://informatics.mayo.edu/sharp/index.php/CEMS (accessed: June 2013)

ANNEX I

Clinical concept modeling using archetypes

A crucial task in our approach is the design of the archetype layer that will serve to abstract from the raw EHR data to the normalized data that the phenotyping algorithm requires. To achieve this, an analysis and specification of the domain concepts are carried out. Based on this analysis, the design and development of the phenotyping archetype are undertaken. This step necessitates a deep understanding of the domain in question, and also a good insight into available archetype models. Inputs to this step are documentation resources about the domain and phenotyping algorithm, such as medical encyclopedias and clinical guidelines, as well as terminological resources. Existing archetype resources are also used as an input in this step. The outputs comprise both a specification of the domain concepts and the phenotyping archetype itself. The tools to be used in this step are the UMLS Terminology Services (UTS) [1], the openEHR Clinical Knowledge Manager (CKM) [2] and the LinkEHR archetype editor [3].

The data used in the phenotyping algorithm often refer to more or less abstract clinical concepts that are not stored in the EHR but can be calculated using e.g. arithmetic operations, logical operations, or combinations thereof. Moreover, abstracted data might be in turn required for the calculation of other data at a higher level of abstraction. An example from a previous study [4] is the concept metastatic solid tumor, which depends equally on the existence of a primary solid tumor and a metastatic tumor, both in turn abstract concepts. To mirror this, in previous works [4,5] we proposed using a collection of interrelated archetypes representing the clinical concepts relevant to the domain (a sort of VHR), which can be referred to as knowledge-rich clinical models based on archetypes. Here, a distinction can be made between first-level archetypes, whose value can be obtained from the EHR data, directly or by means of rather simple expressions, and second (and so on) level archetypes, which require data that can be derived from the EHR but are not available as such.

Prior to the design of the set of archetypes we have performed a thorough analysis of the clinical concepts required by the phenotyping algorithm. This is important not only for the identification of suitable archetypes in existing repositories but also to obtain a precise characterization of the concepts themselves, e.g. including a logic description, if feasible. As far as possible, we select as a starting point the most suitable archetypes from the repository. If necessary, the selected archetypes are specialized to meet the needs of the particular algorithm. In this way we seek to increase the reuse chances of our archetype-based models, which could potentially be shared by other applications in the same domain. Also to increase reuse, we harness the above (logic) concept specifications as semantic annotations for the corresponding archetypes.

Both the domain analysis and the archetype design are labor-intensive and manual tasks requiring technical skills and, at the same time, a good understanding of the domain. In particular, the search for suitable archetypes is an exploration process which critically depends on the domain expertise of the modeler. Thus, if no archetypes are found using key domain concepts as search terms, alternative keys e.g. based on the clinical procedure should be tried. For instance, if the domain refers to findings identified after a colonoscopy, alternative search terms may comprise colonoscopy or histopathology. The archetype should not only have a meaningful name, but should also accommodate the information requirements (e.g. attributes) of the concept. Clearly, the search and browse facilities of the repository are

crucial in this task. Once suitable archetypes have been selected, the design of the archetype layer should be driven by the concept dependences identified as part of the domain analysis.

From EHR data to archetypes

Since the required data are held in local EHR systems we need to convert them into data structures compliant with the supporting archetypes. Therefore, it is necessary to spell out the relationships between the source schemas (EHR or archetypes) and target schemas (archetypes), i.e. their mappings. In complex mapping scenarios as such induced by archetype the common solution is to separate the specification of the relationships between the schemas from the implementation of the actual transformation [6].

This step has two main inputs, namely the source schema (either the local EHR schema or a set of archetypes) and the target archetype. We distinguish three main tasks: 1) mapping specification between source and target schemas; 2) compilation of the mapping specification into an executable program; and 3) execution of the resulting program over the source instances.

In our solution, we employ LinkEHR mapping module [3,6] in the first two tasks. In LinkEHR the relationships between source and target schemas are expressed by users in the form of high-level declarative schema mappings. Such mappings are automatically compiled by the tool into XQuery [7] scripts that perform the actual data transformation. Finally, the XQuery script is executed by an XQuery engine (in our case Saxon [8]) and applied to source instances to generate instances compliant with the target archetype. Although not strictly necessary, the mapping process was carried out starting with first-level archetypes and continuing with second-level ones, and so forth, i.e. following the flow of data. Thereby we were able to validate at each step the generated XQuery script.

Two types of high-level mappings are supported, namely value and structural correspondences. Value correspondences specify how to calculate the value of an atomic attribute in the target from a set of source values. They are defined by a set of pairs, each consisting of a function that specifies how to calculate the value of the target atomic attribute and a condition that must be satisfied to apply the function. Apart from the assignment of a fixed value, the simplest kind of transformation function is the identity function which copies a single source value into a target value. But quite often it is necessary to specify complex functions, for this purpose LinkEHR comes with a wide range of transformation functions covering mathematical, logical, string, type conversion, date and time and metadata (which allow access to archetype metadata such as descriptions or terminology bindings). For instance, Table 1 contains a simple value correspondence transforming gender codes. It transforms the local code in path /patient/gender into a normalized one. Note that the order is relevant and only the first applicable function is used.

Table 1. Example of value correspondence

Filter	Function
/patient/gender='M' OR /patient/gender='m'	0
/patient/gender='W' OR /patient/gender='w'	1
/patient/gender=0 OR /patient/gender=1	/patient/gender
true	9

A number of terminology functions are available to perform terminology abstraction by reasoning over the taxonomic (is-a) hierarchy of SNOMED CT [9], which can be exploited in handling EHR data encoded in this terminology. These allow e.g. calculating the data value for the concept metastatic tumor based on an expression involving the descendants of the SNOMED CT concept "Secondary malignant neoplastic disease". Aggregation functions are also available, which operate on a given grouping context and perform operations such as counting and adding.

Value correspondences can be also used to perform a mapping between terminologies, if needed. More importantly, as illustrated above, value correspondences with terminology functions can be used to perform the necessary abstractions to bridge the gap between EHR data and archetyped data. However, the LinkEHR support for terminology functions is limited to date. In case more advanced support is required external terminology servers might be used, since their functionality can be incorporated by extending the set of LinkEHR transformation functions.

The nested nature of archetypes makes the grouping semantics (how to group and nest data in the target) a key aspect. LinkEHR has a default grouping semantics that minimizes redundancy. In those scenarios where this default semantics is not suitable, structural mappings should be used. They are defined by a set of paths in the source and a filter. They control the creation of target instances, in such a way that a new target instance is constructed for each combination of source values that satisfies the filter. Structural mappings are very useful when using aggregation functions since they allow us to control the grouping context, for instance in the calculation the maximum size of adenomas at patient level instead of at study level.

Domain knowledge modeling using ontologies

Another crucial task in our approach is the design of the ontology layer that will support the implementation of the phenotyping algorithm and will obtain the classification of the patients by means of automated reasoning. OWL ontologies provide a formalization of domain knowledge and permit to represent such content independently of the source EHR schema. The representation without structural details facilitates its exploitation by phenotyping algorithms and its reuse for different tasks.

An input of this step is the result of the analysis done during the clinical concept modeling step. This analysis provides the relevant domain concepts that will have to be represented in the ontology. Best practices in ontology engineering recommend to reuse existing ontologies and to create modular ontologies [10]. Such recommendations would mean in practical terms that we will need to reuse some concepts from different ontologies and that the resulting ontology infrastructure will probably be a networked ontology. Thus, repositories of existing ontologies are another input for this step.

Finding the appropriate concepts for reuse is not an easy task. Repositories like Bioportal [11] provide search options but other tools like Watson [12] might help to find relevant concepts and ontologies on the web. In case of different candidate ontologies for re-use, decisions will

have to be made based on the quality of the ontologies and their appropriateness for the task. It should be noted that the ontologies available do not usually meet all our requirements. On the one hand, existing ontologies might not include all the concepts we need. On the other hand, many currently available biomedical ontologies have been designed for annotation purposes and therefore are not suitable for automated reasoning. For example, Bioportal returns 23 matches for the concept adenoma in ontologies like SNOMED-CT, NCI Thesaurus, MedDRA, MeSH or Experimental Factor Ontology. However, to the best of our knowledge, none of them includes the properties we need to classify adenomas as normal or advanced, as clinicians need for supporting their decisions.

Consequently, extensions and re-engineering of ontologies are likely to be required. This is usually a manual task done with the support of tools like Protégé [13], which permits the addition of new concepts and axioms to existing ontologies, as well as creating new ones. In case new ontologies are developed, principles like the ones proposed by the OBO Foundry [14] could contribute to produce quality, interoperable ontologies. For this purpose we implement the ontologies in OWL-DL [15], which is the OWL subset based on Description Logics. By proceeding in this way, the domain knowledge is made explicit in a set of OWL ontologies and therefore ready to be exploited by means of automated reasoning.

From archetyped data to OWL

This step aims to transform the normalized EHR extracts into OWL. This step has three main inputs: (1) the normalized EHR extracts, which are the source data to be transformed; (2) the archetypes used for representing these extracts, since they provide the data schema; and (3) the domain ontologies, which provide the knowledge schema. This transformation of archetyped data into OWL does not pursue representing in OWL the archetype structure as we did in previous work with a different purpose [16], but to represent the data according to an appropriate domain conceptualization, that is, the transformation of the clinical content.

Technically speaking, the transformation method has two main tasks: (1) definition of the mapping rules between the archetype and the ontology; and (2) application of mapping rules to the normalized EHR extracts. The first task permits to map to source and target schemas, which drives the transformation of the EHR data and provides a flexible way to determine which pieces of data have to be transformed into OWL. Each mapping rule associates archetype entities with OWL classes and properties. The execution of a particular mapping rule will permit the transformation of a data unit into its semantic representation. A mapping rule would associate, for instance, the representation of the adenoma size in the archetype with its corresponding class and property in the ontology. The execution of this mapping would take the value of the size (e.g. 10) from the normalized EHR extract and create an individual of the adenoma class in the ontology, with a value of 10 for the property. The mapping rules are manually defined and require a good understanding of both the archetypes and ontologies involved, whereas the transformation of the data can be automatically performed once the mapping rules have been defined. It should be noted that once the mapping rules are defined, they can be stored and reused in further transformation processes. Given that the EHR data is transformed in a one-by-one basis, the method must avoid the duplicity of OWL individuals. This does not only refer to the problems related to the existence of different values for a functional property (like size) but also to the issues derived from combining the information about a particular patient coming from different extracts. Identity conditions permit to deal with this issue.

The tasks performed in this step have been supported by our semantic transformation engine [17,18], called SWIT, which is capable of generating RDF and OWL content from both relational and XML data. The result of this step is the collection of OWL individuals representing the content of the normalized EHR extracts in a semantic format.

OWL reasoning

The final step is to obtain a partition of the patients according to the phenotyping algorithm. A major motivation for using OWL in this step is its capability to perform sound and complete automated reasoning. The inputs of this step are: (1) the collection of OWL individuals representing the clinical data; (2) the domain knowledge specified in OWL ontologies; and (3) the phenotyping algorithm. First, the phenotyping algorithm has to be implemented in OWL. For this purpose an OWL ontology will be created. This ontology will reuse the domain ontologies previously developed, which will permit the joint exploitation of classification rules, domain knowledge and EHR data by means of automated reasoning.

The ontology will contain one class per group of interest. OWL-DL classes have sets of axioms associated and two types of axioms are relevant in this work: (1) *subClassOf*; and (2) *equivalentClass*. The latter are the most relevant ones for reasoning, because they permit to define the sufficient conditions for an OWL individual to be classified as a member of the OWL class. Defining the inclusion/exclusion criteria as *EquivalentClass* axioms will permit the reasoner to automatically partition the clinical data into the groups of clinical interest.

These axioms have to be carefully designed and optimized taking into account: (1) OWL's open world assumption; (2) how reasoners work; and (3) distribution of activities between the archetype and ontology layers. Our approach tries to perform the activities in the most appropriate layer. For example, operations like counts, negations or obtaining the maximum value would drop the time performance if done using OWL reasoning. Consequently, we propose performing these operations at the archetype level, whereas conceptual classifications like detecting whether an adenoma is normal or advanced and obtaining the category to which a patient belongs are in principle performed at the ontology level.

Once this ontology is ready, an OWL-DL reasoner like Hermit [19] can be applied over the complete semantic dataset in order to infer all the possible information given the EHR data. In particular, we are interested in obtaining the classification of the EHR data according to the phenotyping algorithm. Hence, the final activity in this step is to retrieve the resulting classifications. This can be done by using different query languages available for OWL content such as DL-query [20] or SPARQL [21], or through a programmatic API like OWLAPI [22]. This has been the choice made in this project given that we needed to run simple queries, asking about the membership of the extracts in the phenotyping algorithm classes.

REFERENCES

1. UMLS Terminology Services. https://uts.nlm.nih.gov/home.html (accessed: June 2013)
2. The openEHR Foundation, openEHR Clinical Knowledge Manager. http://www.openehr.org/knowledge/ (accessed: June 2013)
3. Maldonado JA, Moner D, Boscá D, Fernández-Breis JT, Angulo C, Robles M. LinkEHR-Ed: A multi-reference model archetype editor based on formal semantics. International Journal of Medical Informatics 2009;**78**:559–570.
4. Marcos M, Maldonado JA, Martínez-Salvador B, Moner D, Boscá D, Robles M. An Archetype-Based Solution for the Interoperability of Computerised Guidelines and Electronic Health Records. Lecture Notes in Computer Science 2011; **6747**:276-285
5. Marcos M, Maldonado JA, Martínez-Salvador B, Boscá D, Robles M. Interoperability of clinical decision-support systems and electronic health records using archetypes: a case study in clinical trial eligibility. Journal of Biomedical Informatics 2013, http://dx.doi.org/10.1016/j.jbi.2013.05.004
6. Maldonado JA, Martínez-Costa C, Moner D, Menárguez-Tortosa M, Boscá D, Miñarro-Giménez JA, Fernández-Breis JT, Robles M. Using the ResearchEHR platform to facilitate the practical application of the EHR standards. Journal of Biomedical Informatics 2012;**45**:746-762.
7. W3C. XQuery 1.0: An XML Query Language. http://www.w3.org/TR/xquery/ (accessed: June 2013)
8. SAXON XSLT and XQuery processor. http://saxon.sourceforge.net/ (accessed: June 2013).
9. SNOMED-CT http://www.ihtsdo.org/snomed-ct/ (accessed: June 2013)
10. Rector AL, Brandt S, Drummond N, Horridge M, Puleston C, Stevens R. Engineering use cases for modular development of ontologies in OWL. Applied Ontology 2012, **7(2)**:113–132.
11. NCBO Bioportal. http://bioportal.bioontology.org/ (accessed: June 2013)

12. Watson, exploring the Semantic Web. http://kmi-web05.open.ac.uk/WatsonWUI/ (accessed: June 2013)
13. The Protégé Ontology Editor and Knowledge Acquisition System. http://protege.stanford.edu/ (accessed: June 2013)
14. The Open Biomedical Ontologies Foundry. http://www.obofoundry.org/ (accessed: June 2013)
15. W3C, OWL2 Web Ontology Language. http://www.w3.org/TR/owl2-overview/ (accessed: June 2013)
16. Martínez-Costa C, Menárguez Tortosa M, Fernández-Breis JT: Clinical data interoperability based on archetype transformation. Journal of Biomedical Informatics (2011)**44(5):** 869-880
17. Miñarro-Giménez JA, Fernández-Breis JT, Ontology-driven Method for Integrating Biomedical Repositories, Lecture Notes in Computer Science 2011;**7023**:473-482
18. Semantic Web Integration Tool. http://sele.inf.um.es/swit (accessed: June 2013)
19. Hermit Reasoner. http://www.hermit-reasoner.com/ (accessed: Junes 2013)
20. DL Query. http://protegewiki.stanford.edu/wiki/DL_Query (accessed: June 2013)
21. SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/ (accessed: June 2013)
22. The OWLAPI. http://owlapi.sourceforge.net/ (accessed: June 2013)