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]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001923)

[1]Departamento de Informática y Sistemas, Universidad de Murcia, Murcia, Spain
[2]Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Valencia, Spain
[3]Department of Computer Engineering and Science, Universitat Jaume I, Castellón, Spain
[4]Fundación para la Formación e Investigación Sanitaria, Murcia, Spain

Correspondence to
Dr Jesualdo Tomás Fernández Breis, Departamento de Informática y Sistemas, Universidad de Murcia, Murcia 30100, Spain; jfernand@um.es

JTFB, JAM, and MM contributed equally.

Received 15 April 2013
Revised 12 July 2013
Accepted 26 July 2013
Published Online First 9 August 2013

To cite: Fernández-Breis JT, Maldonado JA, Marcos M, *et al*. *J Am Med Inform Assoc* 2013;**20**:e288–e296.

ABSTRACT

Background 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.

Objective 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 (ie, data level) and the rest using ontologies (ie, 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 have been used and the patients have been successfully classified by the risk of developing colorectal cancer.

Conclusions 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, for example 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.[4–6] 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] clinical document architecture 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 web ontology language (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.[13–15] 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.

MATERIALS 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 (ie, data level) and the rest using ontologies (ie, 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 these 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

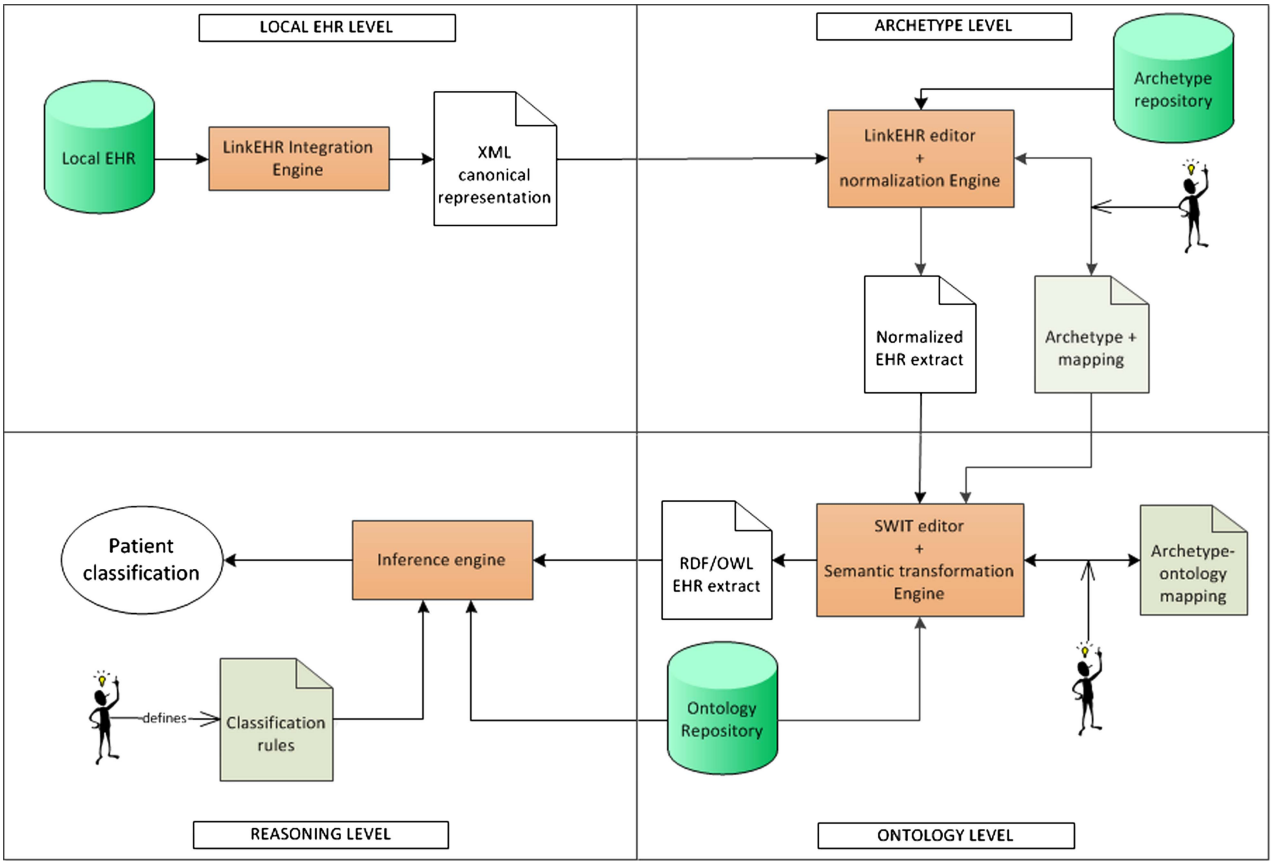

Figure 1 Overview of the methodological approach.

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 reused 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 online supplementary annex I.

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 (eg, medical encyclopedias, clinical guidelines) ▶ Terminological resources (eg, 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]

EHR, electronic healthcare records; SNOMED, systematized nomenclature of medicine; UMLS, unified medical language system.

RESULTS

Case study description

Colorectal cancer is one of the most important causes of mortality in many developed countries according to the 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[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. First, 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

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]

EHR, electronic healthcare records; SWIT, semantic web integration tool.

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, that is, 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.

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 semantic web integration tool (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

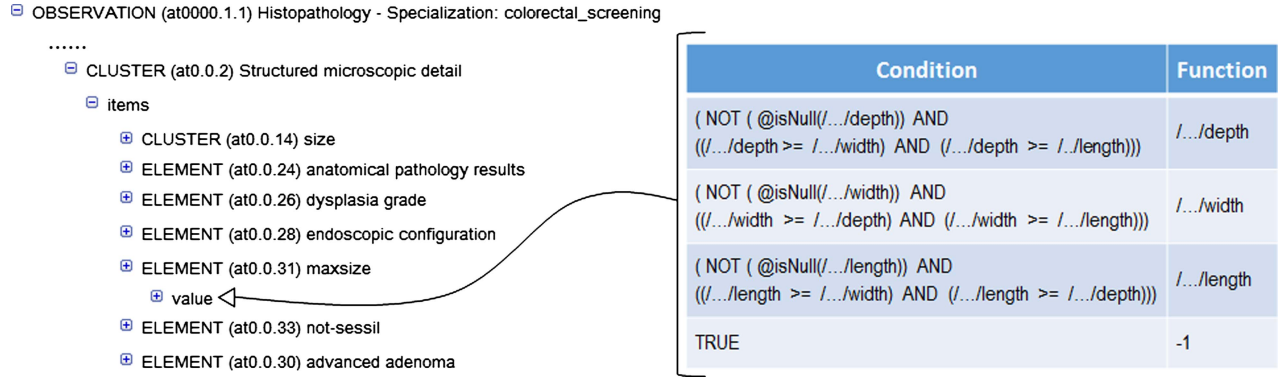

Figure 2 Value correspondence for calculating the maximum size (depth, width, or length) of an adenoma finding. As it can be observed, if none of the dimensions has been recorded the value −1 is assigned.

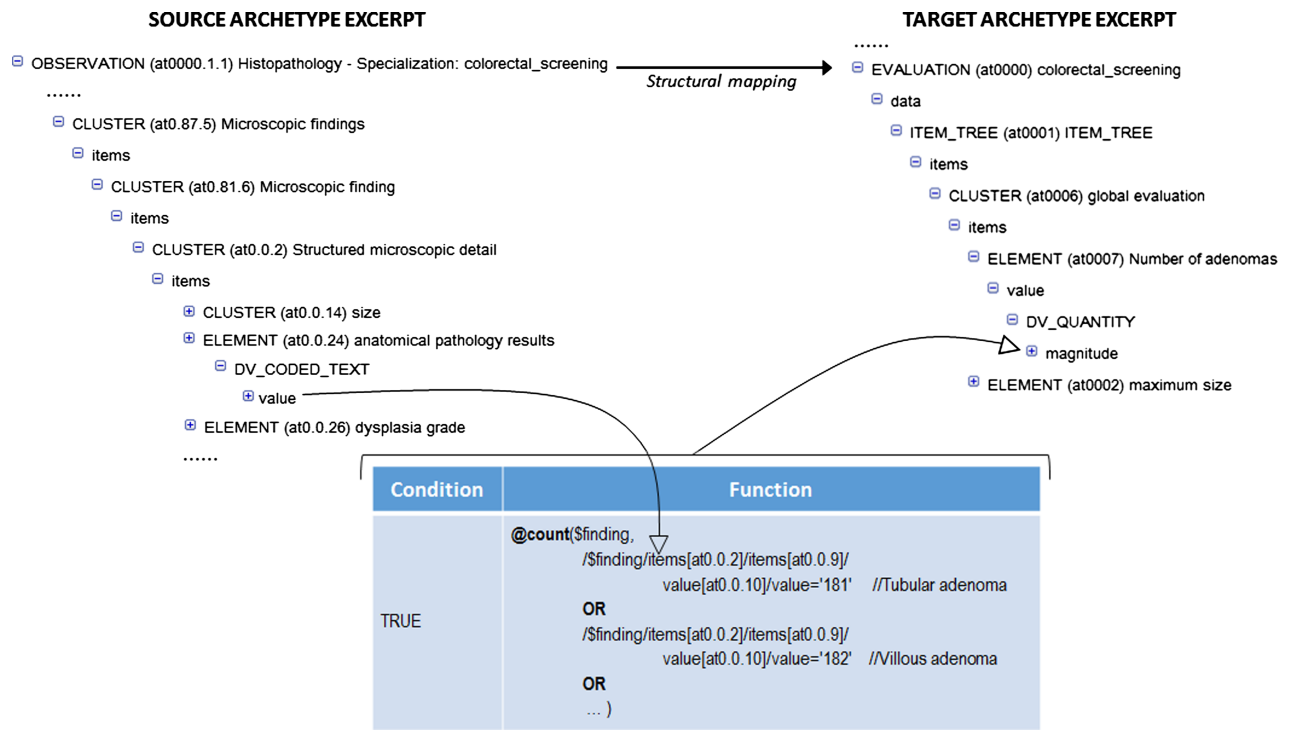

Figure 3 Calculation of the number of adenomas in a histopathology study. The variable $finding represents the path of microscopic findings (CLUSTER at 0.81.6) in the source archetype. The complete source instance provides the context for grouping and counting in the structural mapping. The values 181 and 182 identify the types of adenoma in the database.

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).

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-description logics (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]

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% of 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, for example 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 ms per report, and about 150 ms 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 s.

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 (eg, 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 64.4% of the cases. Among the discrepancies (35.6% of the cases), 58% correspond 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 non-compliance, as physicians were not supposed to follow it. The list of discrepant cases and the corresponding data files are available at 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 two classifications. The physician determined that the OWL classification was the correct one in 100% of 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

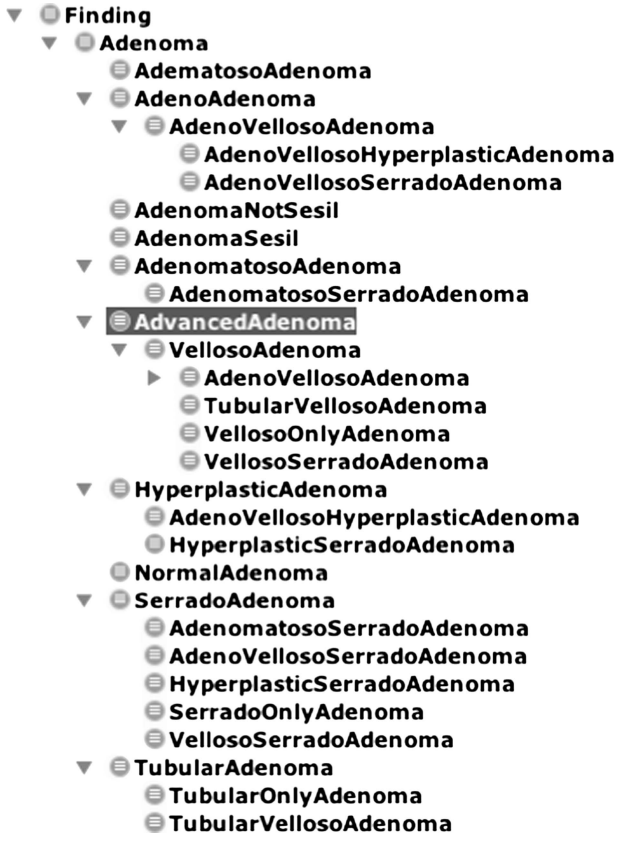

Figure 4 Excerpt of the domain ontology.

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. For example, 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 (eg, 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 for example, 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, for example, 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 simple protocol and resource description framework query language (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.

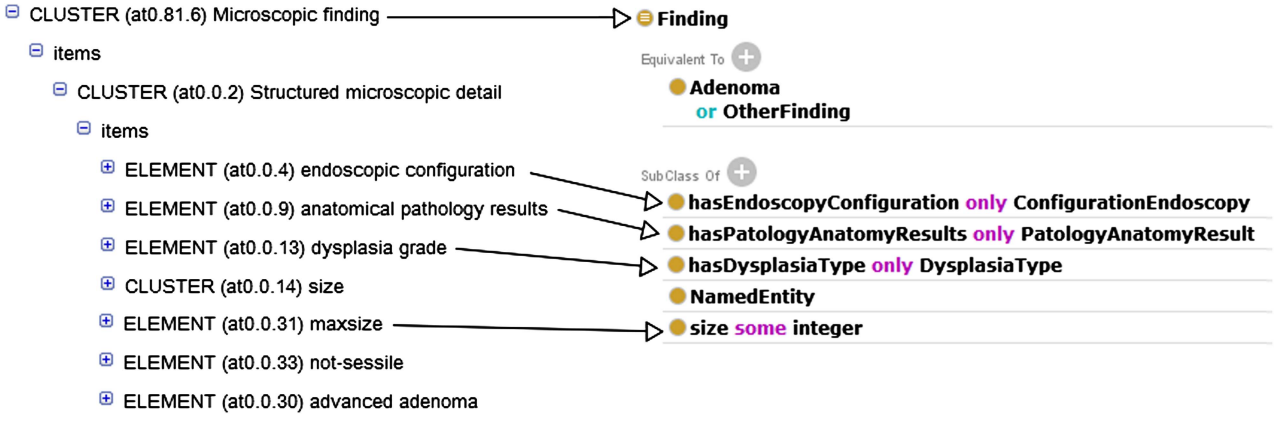

Figure 5 Partial view of the mapping for Finding.

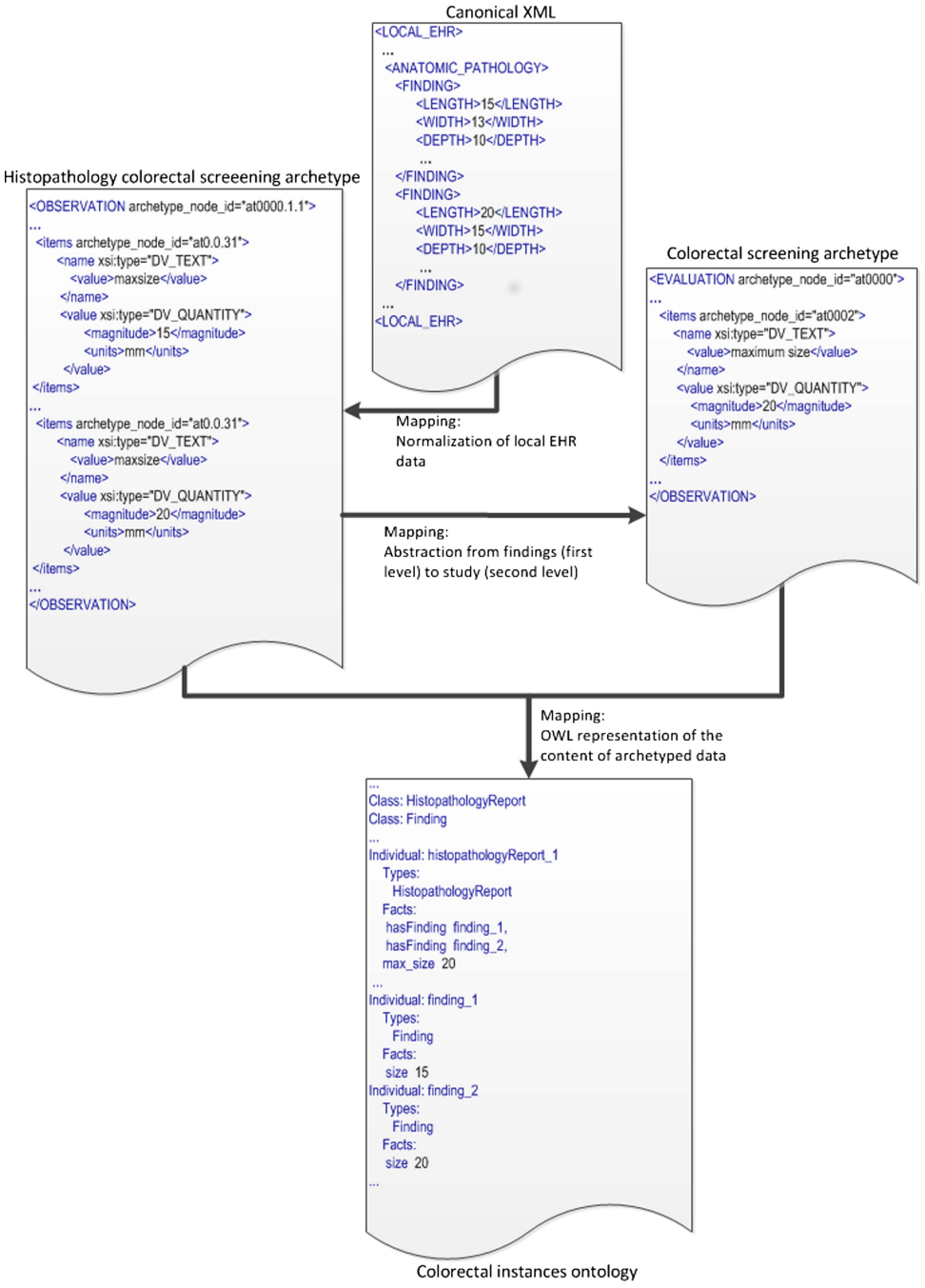

Figure 6 Example of a complete data transformation process.

Besides, the SPARQL inference possibilities based on properties are limited. For example, 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, for example, 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, also taking 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 work we have managed systematized nomenclature of medicine (SNOMED)-CT content in both EHR-archetype[38 39] and archetype-ontology[17] transformation layers. In our view the exploitation of the terminological

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 three adenomas, or a histopathology report whose largest adenoma has at least 20 mm 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 20 mm or at least five 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 20 mm 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 two normal adenomas, but does not contain any advanced one. This rule is the same for both protocols.

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 into 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 work 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.

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

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.

Acknowledgements 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.

Contributors 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, and 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 program through grants TIN2010-21388-C01 and TIN2010-21388-C02. MCLG was supported by the Fundación Séneca through grant 15555/FPI/2010.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The website http://miuras.inf.um.es/colorectal contains all the additional data, which is available to everybody.

REFERENCES

1 Cuggia M, Besana P, Glasspool P. Comparing semi-automatic systems for recruitment of patients to clinical trials. *Int J Med Inform* 2011;80:371–88.
2 Sujansky W. Heterogeneous database integration in biomedicine. *J Biomed Inform* 2001;34:285–98.
3 Schadow G, Russler DC, Mead CN, *et al*. Integrating medical information and knowledge in the HL7 RIM. *Proceedings of the AMIA Symposium*, 2000:764–8.
4 Johnson PD, Tu SW, Musen MA, *et al*. A virtual medical record for guideline-based decision support. *Proceedings of the AMIA 2001 Annual Symposium*, 294–8.
5 German E, Leibowitz A, Shahar Y. An architecture for linking medical decision-support applications to clinical databases and its evaluation. *J Biomed Inform* 2009;42:203–18.
6 Peleg M, Keren S, Denekamp Y. Mapping computerized clinical guidelines to electronic medical records: knowledge-data ontological mapper (KDOM). *J Biomed Inform* 2008;41:180–201.
7 Maldonado JA, Martínez-Costa C, Moner D, *et al*. Using the ResearchEHR platform to facilitate the practical application of the EHR standards. *J Biomed Inform* 2012;45:746–62.
8 Parker CG, Rocha RA, Campbell JR, *et al*. Detailed clinical models for sharable, executable guidelines. *Stud Health Technol Inform* 2004;107:145–8.
9 Clinical Information Modeling Initiative. http://informatics.mayo.edu/CIMI/index.php/Main_Page (accessed Jun 2013).
10 W3C, OWL2 Web Ontology Language. http://www.w3.org/TR/owl2-overview/ (accessed Jun 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 Jun 2013).
13 Martínez-Costa C, Menárguez-Tortosa M, Fernández-Breis JT, *et al*. A model-driven approach for representing clinical archetypes for Semantic Web environments. *J Biomed Inform* 2009;42:150–64.
14 Iqbal AM. An OWL-DL ontology for the HL7 reference information model. Toward useful services for elderly and people with disabilities Berlin: Springer, 2011:168–75.
15 Tao C, Jiang G, Oniki TA, *et al*. A semantic-web oriented representation of the clinical element model for secondary use of electronic. *J Am Med Inform Assoc* 2013;20:554–62.
16 Heymans S, McKennirey M, Phillips J. Semantic validation of the use of SNOMED CT in HL7 clinical documents. *J Biomed Semantics* 2011;2:2.
17 Menárguez-Tortosa M, Fernández Breis JT. OWL-based reasoning methods for validating archetypes. *J Biomed Inform* 2013;46:304–17.
18 Lezcano L, Sicilia MA, Rodríguez-Solano C. Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules. *J Biomed Inform* 2011;44:343–53.
19 Tao C, Wongsuphasawat K, Clark K, *et al*. Towards event sequence representation, reasoning and visualization for EHR data. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI'12)*. New York, NY, USA: ACM:801–6.
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 Jun 2013).
23 UMLS Terminology Services. https://uts.nlm.nih.gov/home.html (accessed Jun 2013).
24 The openEHR Foundation, openEHR Clinical Knowledge Manager. http://www.openehr.org/knowledge/ (accessed Jun 2013).
25 Maldonado JA, Moner D, Boscá D, *et al*. LinkEHR-Ed: a multi-reference model archetype editor based on formal semantics. *Int J Med Inform* 2009;78:559–70.
26 SAXON XSLT and XQuery processor. http://saxon.sourceforge.net/ (accessed Jun 2013).
27 NCBO Bioportal. http://bioportal.bioontology.org/ (accessed Jun 2013).
28 The Protégé Ontology Editor and Knowledge Acquisition System. http://protege.stanford.edu/ (accessed Jun 2013).
29 Semantic Web Integration Tool. http://sele.inf.um.es/swit (accessed Jun 2013).
30 Hermit Reasoner. http://www.hermit-reasoner.com/ (accessed Jun 2013).
31 The OWLAPI. http://owlapi.sourceforge.net/ (accessed Jun 2013).
32 Institute for Health Metrics and Evaluation. Global Burden of Disease. http://www.healthmetricsandevaluation.org/gbd (accessed Jun 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 Jun 2013).
35 DL Query. http://protegewiki.stanford.edu/wiki/DL_Query (accessed Jun 2013).
36 SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/ (accessed Jun 2013).
37 Semantic Web Rule Language. http://www.w3.org/Submission/SWRL/ (accessed Jun 2013).
38 Marcos M, Maldonado JA, Martínez-Salvador B, *et al*. Interoperability of clinical decision-support systems and electronic health records using archetypes: a case study in clinical trial eligibility. *J Biomed Inform* 2013;46:676–89.
39 Marcos M, Maldonado JA, Martínez-Salvador B, *et al*. An archetype-based solution for the interoperability of computerised guidelines and electronic health records. *Lect Notes Comput Sci* 2011;6747:276–85.
40 MobiGuide: Guiding patients anytime everywhere. http://www.mobiguide-project.eu/ (accessed Jun 2013).
41 EURECA: Enabling information re-Use by linking clinical RE search and Care. http://eurecaproject.eu/ (accessed Jun 2013).
42 Rea S, Pathak J, Savova G, *et al*. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. *J Biomed Inform* 2012;45:763–71.
43 Clinical Element Models. http://informatics.mayo.edu/sharp/index.php/CEMS (accessed Jun 2013).