Interoperability of clinical decision-support systems and electronic health records using archetypes: a case study in clinical trial eligibility

Mar Marcos[a,*], Jose A. Maldonado[b], Begoña Martínez-Salvador[a], Diego Boscá[b], Montserrat Robles[b]

[a]*Dept. of Computer Engineering and Science, Universitat Jaume I*
Av. de Vicent Sos Baynat s/n, 12071 Castellón, Spain
[b]*Biomedical Informatics Group, ITACA Institute, Universidad Politécnica de Valencia*
Camino de Vera s/n, 46022 Valencia, Spain

Abstract

Clinical decision-support systems (CDSSs) comprise systems as diverse as sophisticated platforms to store and manage clinical data, tools to alert clinicians of problematic situations, or decision-making tools to assist clinicians. Irrespective of the kind of decision-support task CDSSs should be smoothly integrated within the clinical information system, interacting with other components, in particular with the *electronic health record* (EHR). However, despite decades of developments, most CDSSs lack interoperability features.

We deal with the interoperability problem of CDSSs and EHRs by exploiting the *dual-model methodology*. This methodology distinguishes a reference model and archetypes. A reference model is represented by a stable and small object-oriented model that describes the generic properties of health record information.

*Corresponding author. *Telephone:* +34 964 72 82 88. *Fax:* +34 964 72 84 86.
Email address: Mar.Marcos@uji.es (Mar Marcos)

For their part, archetypes are reusable and domain-specific definitions of clinical concepts in the form of structured and constrained combinations of the entities of the reference model. We rely on archetypes to make the CDSS compatible with EHRs from different institutions. Concretely, we use archetypes for modelling the clinical concepts that the CDSS requires, in conjunction with a series of knowledge-intensive mappings relating the archetypes to the data sources (EHR and/or other archetypes) they depend on.

We introduce a comprehensive approach, including a set of tools as well as methodological guidelines, to deal with the interoperability of CDSSs and EHRs based on archetypes. Archetypes are used to build a conceptual layer of the kind of a *virtual health record* (VHR) over the EHR whose contents need to be integrated and used in the CDSS, associating them with structural and terminology-based semantics. Subsequently, the archetypes are mapped to the EHR by means of an expressive mapping language and specific-purpose tools. We also describe a case study where the tools and methodology have been employed in a CDSS to support patient recruitment in the framework of a clinical trial for colorectal cancer screening.

The utilisation of archetypes not only has proved satisfactory to achieve interoperability between CDSSs and EHRs but also offers various advantages, in particular from a data model perspective. First, the VHR/data models we work with are of a high level of abstraction and can incorporate semantic descriptions. Second, archetypes can potentially deal with different EHR architectures, due to their deliberate independence of the reference model. Third, the archetype

instances we obtain are valid instances of the underlying reference model, which would enable e.g. feeding back the EHR with data derived by abstraction mechanisms. Lastly, the medical and technical validity of archetype models would be assured, since in principle clinicians should be the main actors in their development.

Keywords: Clinical Decision Support Systems, Electronic Health Records, Systems Integration, Clinical Trials, Terminology, SNOMED CT, Artificial Intelligence.

---

1. Introduction

A *clinical decision-support system* (CDSS) can be defined as "any computer program designed to help health professionals make clinical decision" [1]. This definition encompasses systems as diverse as sophisticated platforms to store and manage clinical data, tools to alert clinicians of problematic situations (e.g. drug-drug interactions), or decision-making tools to assist clinicians by providing patient-specific recommendations. In a broader sense, other systems which use clinical data to support decisions not directly related to patient care can also be considered to be CDSSs. Systems to support patient recruitment for clinical research trials are a representative example of such CDSSs.

Irrespective of the kind of decision-support task, ideally CDSSs should be smoothly integrated into the computer tools that are routinely used by clinicians, and more importantly they should be able to operate without the manual entry of

data already entered using some other system [1]. This implies some interaction with other components of the clinical information system, in particular with the *electronic health record* (EHR) to access all the clinical data required. However, after more than 3 decades of developments most of CDSSs have been either stand-alone systems or small components embedded within EHR or physician order entry systems [1], [2].

An important problem is the heterogeneity of clinical data sources, which may differ in the data models, schemas, naming conventions, and degree of detail used to represent similar data [3]. On the other hand, CDSSs very often require data at a level of abstraction higher than raw clinical data, a problem which has been referred to as the "impedance mismatch" between the CDSS and the EHR [4], [5]. There have been several initiatives, involving standardisation bodies, to define generic EHR architectures for the communication of health data, such as CEN/ISO EN13606 [6], openEHR [7], HL7 CDA [8], or CDISC ODM [9]. However, their use is not widespread in current CDSSs.

One of the main contributions of recent EHR architectures is the *dual-model methodology* [10] for the description of the structure and semantics of health data. The dual model methodology distinguishes a reference model and archetypes. A *reference model* is represented by a stable and small object-oriented model that describes the generic properties of health record information (such as folder, document, section, and audit). The generality of the reference model (RM) is complemented by the particularity of archetypes. An *archetype* is a detailed, reusable and domain-specific definition of a clinical concept (such as Apgar score, discharge

report, and primary care EHR) in the form of a structured and constrained combination of the entities of the RM. The principal purpose of archetypes is to provide a powerful way of managing the description, creation, validation and querying of EHRs. From a data point of view, archetypes are a means for providing structural and terminology-based semantics to data instances that conform to some RM.

We deal with the interoperability problem of CDSSs and EHRs by exploiting dual-model EHR architectures. In previous articles we propose a solution that exploits openEHR archetypes for the interoperability of CDSSs based on clinical guidelines [11], [12]. In this article we take a further step and describe the implementation of a prototype that demonstrates the feasibility of our proposal. The prototype is based on a case study dealing with the determination of patient eligibility in a clinical trial (CT) for colorectal cancer screening. Typically, both clinical guideline recommendations and CT eligibility criteria are intended to be shared across different institutions, at national or even at international level, and thus the standardised access to the EHR becomes a pressing need in CDSSs for these purposes.

2. Background

The advantages of integration with the EHR were already acknowledged in early CDSSs. Thus, different authors have sought such integration while pursuing the shared use of CDSSs, in particular in guideline-based CDSSs. One of the early approaches was to separate the site-specific data references from the logic rules. The best example of this approach is the Medical Logic Modules (MLM)

of Arden syntax [13], [14], currently a HL7 standard for representing clinical logic. In Arden Syntax MLMs, the site-specific mappings (queries) to EHR data are defined in a separated section, known as the data section. In this section, the specific details for retrieving a required data element from a data source, such as an EHR, are enclosed in a pair of curly braces. The problem of combining site-specificity with a standard syntax has been known as the "curly braces problem".

The problem of combining data residing at different sources and providing a unified view of these data, known as data integration [15], is not exclusive of the health-care domain. Among the different approaches to data integration, federated information systems are the most widely used. These systems leave data at the sources and provide querying access to the set of data sources through a virtual federated view (schema). The federation relies on schema mapping for the integration of data sources. The mediator/wrapper architecture [16] is one of the most commonly used approaches to achieve data federation. A mediator is a read-only virtual database which is introduced between the data sources and the client applications and is capable of answering queries about the underlying data [17].

Starting from the federated approach, other initiatives rely on the definition of a global virtual schema, known as Virtual Medical Record or Virtual Health Record (VHR), over a set of local EHR systems, and on a set of mappings from the VHR to the local EHR systems. The VHR includes an information model that defines generic concepts (such as Observation, Instruction, etc.) for representing patient data, domains for attributes in the information model (e.g. terminologies), and a query language [18]. Queries for patient data in the CDSS are posed against

the VHR schema. In order to answer them they are translated into an equivalent set of local subqueries that are executed against the local data sources, whose results are then combined. This approach alleviates the curly braces problem since it is only necessary to define the mappings between the VHR and the CDSS once. When a CDSS is to be bound to a new EHR system, only the mappings between the EHR system and the virtual view are needed, thus the CDSS remains unaltered and its portability is facilitated.

3. Approach

We are concerned with the use of archetypes within CDSSs as a standardised mechanism for the interaction with the EHR, in order to obtain CDSSs that can be shared across institutions without the need for modifications in the implementation. This problem is mentioned by Sujansky as one of the heterogeneous database integration challenges in Medical Informatics [3], and is usually solved by means of abstractions that make the CDSS compatible with clinical databases from different institutions. We propose to use archetypes to build a semantically-rich VHR for this purpose. More precisely, our proposal is to develop a series of archetypes for the data/concepts that the CDSS requires, and to include references to these archetypes in the parts of the CDSS knowledge base (KB) where interactions with the EHR should occur. It is important to note that our interest in shared use (and reuse) is not limited to the KB as a whole but also covers the archetypes modelling the necessary clinical data/concepts.

We are also concerned with technical solutions to implement our approach.

Technical implementation requires on one hand a platform for the access to the EHR data via archetypes, in the likely case that the EHR does not support archetypes natively. On the other hand, an inference engine supporting the use of archetypes is required. For the former, we have used the data integration engine of the LinkEHR Normalization Platform [17] (see section 4 for more details). With respect to the inference engine, in the absence of engines that support data access via archetypes, we have chosen to use an existing guideline execution engine in combination with a specific mediator module which allows taking input data from a variety of external data sources. Concretely, we have used the Tallis Engine, which is a non-commercial execution tool for the PROforma guideline representation language [19]. PROforma is particularly powerful with regard to decision models [20], [5] which makes it very well-suited for describing the eligibility criteria of our case study. Lastly, as archetype framework we have chosen the proposal by the openEHR Foundation [7], [21], which stands out for its web-based repository of reusable archetypes. Figure 1 depicts the overall architecture, showing the PROforma and LinkEHR modules involved. Notice that despite our particular choices of PROforma and openEHR, our approach is rather generic.

The architecture is particularly well adapted to CDSSs, which usually require performing a series of operations or abstractions on the EHR data and subsequently using these elaborated data to provide the decision support itself. The previous operations/abstractions can be resolved by the LinkEHR modules, based on a series of predefined mappings, while the decision support tasks can be performed by the inference engine. An important feature is that the LinkEHR trans-

formations are applied to data from clinical databases working under the closed world assumption, which allows us to conclude e.g. that the patient does not suffer from a condition unless it has been documented in the EHR. Handling negation at this level can be advantageous e.g. if the inference engine of choice works under the open world assumption. Finally, note that although we employ medical terminologies/ontologies for mapping purposes, advanced reasoning over these terminologies is not considered a priority and therefore is beyond the scope of this work.

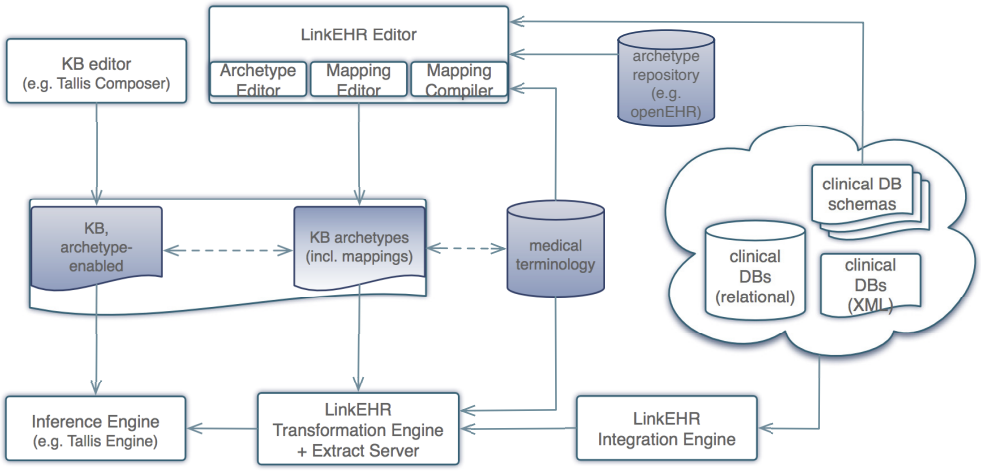

Figure 1: An architecture to deal with the interoperability of CDSSs and EHRs using archetypes.

3.1. Linking the CDSS to heterogeneous EHRs

We rely on archetypes to make the CDSS compatible with EHRs from different institutions. Concretely, we use archetypes to provide a uniform (and abstract) interface to clinical data, which can be subsequently used to connect with hetero-

geneous EHR implementations. This view has several implications in regard to knowledge modelling:

1. it is necessary to design a collection of archetypes suitable for the decision-support tasks carried out in the CDSS

2. it is necessary to ensure that the CDSS KB is compliant with these archetypes

3. it must be ensured that the connection with the desired source EHR (or clinical databases) via the designed archetypes is feasible

With respect to (1.), to increase the chances of reuse it is important that the CDSS archetypes are designed considering the available archetypes and standards. This is consistent with the philosophy of the openEHR archetypes we propose to use, since by design they are intended for wide reuse. Requirement (2.) is also crucial since KBs are often modelled without regard to the interaction with the EHR, which hinders interoperability. Here again, KBs should be modelled taking into account EHR standards and available archetypes all along. Finally, requirement (3.) involves the definition of a series of mappings relating the archetype elements from (1.) to the corresponding data items in the source EHR. Because CDSSs often operate on data abstracted from lower-level EHR data, these mappings may relate one archetype to several data items (or even other archetypes), e.g. by means of logical abstractions. Much of the work presented in this article relates to the design and mapping of archetypes (requirements (1.) and (3.)).

3.2. The openEHR architecture

The openEHR architecture uses the dual-model methodology for the description of EHR data [22]. A key feature is the separation of clinical knowledge, described using archetypes, from the information or recording model, referred to as the RM. Thanks to this two-level modelling, openEHR-based systems should be able to easily accommodate changes in clinical concepts, which will only require modifications to the archetypes.

An openEHR archetype is a model for the capture of clinical knowledge [23]. It is a machine processable specification of a domain/clinical concept, in the form of structured constraints and based on the openEHR RM. An archetype extensively describes the structure and content of clinical concepts such as "diagnosis" or "blood pressure" [24]. In principle archetypes have been defined for wide reuse, however they can be specialised for adaptation to local singularities. To promote reuse, archetypes include all the relevant attributes about a specific clinical concept, according to clinicians' criteria. In this sense, they can be considered as maximal data sets.

With respect to the openEHR archetype formalisms, we can mention the Archetype Object Model (AOM), which is an object-oriented model for the representation of archetypes in memory, and the Archetype Definition Language (ADL), which is the normative abstract serialisation of archetypes based on F-logic queries with terminology [25]. Other archetype serialisations, e.g. XML-based serialisations, are often used for practical purposes.

3.3. Reusing openEHR archetypes

When it comes to archetype reuse, it is essential to rely on sources of a certain quality. We have used the openEHR Clinical Knowledge Manager [26] (CKM), which is a web-based repository allowing for archetype search, browse and download. Archetypes in the CKM have been created by independent domain experts, mainly clinicians and computer scientists, and then they have been released to the community as freely available content. Before publication, archetypes undergo an iterative review process to ensure that they cover as many use-cases as possible and thus constitute a reasonable maximal data set (with a high reuse potential).

According to openEHR, the main categories for the description of clinical concepts are observation, evaluation, instruction and action. This categorisation is related to the way in which information is created during the care process: an *observation* is created by an act of observation, measurement, or testing; an *evaluation* is obtained by inference from observations, using personal experience and/or published knowledge; an *instruction* is an evaluation-based instruction to be performed by healthcare agents; and an *action* is a record of the interventions that have occurred, instruction-related or not. The number and specificity of archetypes in the CKM differ significantly among and within categories, possibly because they have been developed according to individual interests.

For the purposes of our case study we have mainly used the CKM archetype `openEHR-EHR-EVALUATION.problem.v1`. It is intended to be used for recording any information about general health-related problems, understood as issues that negatively affect the physical, mental and/or social wellbeing of an

individual. The archetype contains slots for the identifier of the problem, its clinical description, and its severity, as well as for a number of relevant dates (e.g. date of initial onset, and date clinically recognised), amongst other things. As we explain later, the archetype `openEHR-EHR-EVALUATION.problem.v1` has been specialised to meet the needs of data sources used in the CDSS of our case study.

4. Material and methods

4.1. The LinkEHR platform

The LinkEHR Normalization Platform [27] is a set of modules that allow: i) the creation of an archetype-based customisable view over a set of heterogeneous and distributed EHR data sources [17]; ii) the editing of archetypes based on different RMs (standards), as long as an XML Schema is available [28] (several RMs have been tested successfully: CEN/ISO EN13606, openEHR, HL7 CDA, CDISC CDM and CCR); and iii) the specification of declarative mappings between archetypes and data sources, and from these mappings the automatic generation of XQuery scripts which translate source XML data into XML documents that are archetype compliant.

LinkEHR employs archetypes for both the semantic description of legacy EHRs and the publication of existing clinical information in the form of standardised EHR extracts. Since health data reside in the underlying EHR systems, it is necessary to define some kind of mapping information that links entities in the archetype to data elements in data repositories (e.g. elements and attributes in the

case of XML sources). Basically, these mappings specify how to create archetype instances from the content of the data sources.

Different LinkEHR modules are involved in our CDSS interoperability framework (see Figure 1). A crucial tool is the LinkEHR archetype and mapping editor. During archetype editing, the tool provides support to ensure that the archetype being edited is valid with respect to the RM (and parent archetype, if any), e.g. showing the elements allowed [29]. Subsequently, the tool supports both the editing of archetype-source declarative mappings (through a wide array of transformation functions, see section 4.2 for more details), and the automatic generation of adequate XQuery transformation scripts, based on the mapping specifications and the source schemas. Also important, the LinkEHR integration engine works as a data integration module that provides a virtual, integrated and global XML view over distributed clinical data sources, XML or not [17]. Finally, the LinkEHR transformation engine and the extract server jointly provide an archetype and RM-compliant extract of the clinical data, from the data supplied by the integration engine and the above mentioned XQuery scripts.

4.2. Mapping methods in LinkEHR

At the schema level, the above mentioned mappings require an explicit representation of how the source schema (either an EHR schema or a set of archetypes) and target schema (archetype) are related to each other. The effort required to create and manage such mappings is considerable. The common case is to write intricate and non-reusable programs to perform the required transformations. This is

even more complex in the case of archetypes, since they are used to model generic concepts without regard to the internal architecture of the EHR. LinkEHR allows the specification of high-level declarative mappings, by defining a set of correspondences between the entities of archetypes and source schemas. Two types of correspondences are supported, namely value and structural correspondences. The former specify how to calculate atomic values, whereas the latter may be used to control the generation and grouping of elements in the target.

In our case study we have primarily used value correspondences. They are defined by a set of pairs, each consisting of a mapping function that specifies how to calculate a value in the target from a set of source values, and a condition that source data must satisfy so that the transformation is applied. References to source data/schemas are frequent in both mapping functions and conditions. These may take the form of an XPath location path in the case of XML sources (i.e. `/step/step/...`), or an ADL path [25] in case other archetypes are used as source schemas. With respect to the mapping functions, the simplest kind is the identity function, which copies a source value into a target value. However, quite often it is necessary to specify arbitrarily complex functions. For this purpose the LinkEHR tool comes with a wide range of functions such as type conversion, as well as mathematical, logical, string, date and time, and metadata functions (which allow access to archetype metadata such as descriptions or type names). Additionally, mapping functions can be easily extended. For example, a number of terminology functions have been added for the interoperability of CDSSs and EHRs. These functions allow terminology abstraction by reasoning over the

acyclic taxonomic (*is-a*) hierarchy of SNOMED CT, among other things (for examples refer to section 5.5). Aggregation functions have been also added for interoperability purposes. As illustration we can cite the counting and adding functions, both operating on a given context (see also section 5.5).

The example in Table I illustrates a simple value correspondence for transforming gender codes. It transforms the local gender code in the path `/patient/gender` of an XML EHR fragment (source data) into a normalised code to be stored somewhere within an archetype (target data). Note that the order of mapping specification pairs is relevant, and that only the first applicable one is used.

Table I: A simple mapping transforming the gender codes from an XML source.

Condition	**Mapping 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`

In mapping scenarios with complex nested structures as those induced by EHR information models, a key aspect is the grouping semantics, i.e. how we have to group and nest data to build a target instance. LinkEHR comes with a default grouping semantics based on Partition Normal Form, which has resulted adequate in many mapping scenarios since it tends to group together data with the same clinical context (date, author, etc). In those scenarios where this default semantics is not suitable, structural mappings should be used. In short, structural mappings define how to generate and group data in the target on the basis of source data [30].

From the set of high-level declarative mapping specifications (value and struc-

tural correspondences) and archetype constraints, an XQuery script is generated. The resulting script transforms a source schema instance (EHR data or archetype instance) into an XML document compliant with both the underlying RM and the target archetype.

5. Results

5.1. Case study: a CT for colorectal cancer screening

As case study we have used a CT from the ClinicalTrials.gov repository [31], which is a registry of clinical trials conducted in the US and worldwide. We have chosen a CT for colorectal cancer screening which has been designed to compare the efficacy of 2 different screening procedures. Concretely, the goal is to compare the efficacy of biennial immunochemical fecal occult blood test versus colonoscopy every 10 years for the reduction of colorectal cancer-related mortality at 10 years in average-risk population [32]. It is an ongoing trial coordinated by Hospital Clinic of Barcelona (Spain) and conducted in collaboration with several Spanish hospitals. This particular CT has been chosen on the basis of the complexity of the clinical concepts it requires, e.g. involving definitions in terms of arithmetic, aggregation and/or logic operations based on other concepts, possibly complex ones (see section 5.2).

In general, CT inclusion and exclusion criteria can be readily used to implement a CDSS for patient eligibility determination. Starting exactly from these criteria, we have implemented a (single decision) CDSS in the PROforma representation language. On the other hand, the clinical concepts to which the CT

criteria refer constitute the minimum data needed for the operation of the CDSS, and hence requiring archetypes. Furthermore, to make the CDSS work the concepts used in the decision mechanism must be tuned in to the terms and concepts used in the EHR as far as possible. For instance, one issue is recognising semantic equivalence in the face of the multiplicity of terms used to describe a disease. This and other issues justify the need for a shared concept representation [33]. In the rest of the section we review important aspects related to the PROforma CDSS for CT patient eligibility and, particularly, to the design and mapping of CT archetypes. Before that, we give an overview of CT concepts and provide working definitions thereof.

5.2. Specification and representation of CT concepts

The inclusion and exclusion criteria of the colorectal cancer screening CT mainly refer to demographic data like *sex* and *age* (inclusion of "men and women aged 50-69 years"), or to data on health problems like *colorectal cancer* and *colorectal adenoma* (e.g. exclusion of patients with "personal history of colorectal cancer, colorectal adenoma,..."), which in principle are all expected to be found in the EHR. An important issue to consider regarding health problems is that CTs often refer to rather generic conditions describing a wealth of more specific problems –*terminology abstractions* according to Peleg *et al.* [34]. To unravel such terminology abstractions we have resorted to additional information sources as well as to UMLS Terminology Services [35], in particular to the SNOMED CT® terminology [36].

Another difficulty is the utilisation of terms that can be defined by means of more or less complex expressions referring to other lower-level terms –*definitions of abstract terms* according to Peleg *et al.*. To clarify these terms, we have turned to an Oncology Specialist. A good example is the problem *severe comorbidity* (exclusion of patients with "severe comorbidity"), which turned out to be a rather high-level/abstract concept. Following the definition of the widely used Charlson index [37], a comorbidity score (and thus a severity grade) can be calculated as the sum score of the morbidities affecting the patient. A total of 19 morbidities are considered, ranging from less severe to more severe diseases such as *AIDS* and *metastatic solid tumor*. Among these, several terms corresponded in turn to complex terms (terminology abstractions or abstract definitions) that required further analysis.

One of our concerns is the reuse of the archetypes designed for the clinical data/concepts required by the CDSS (see section 3). To increase reuse chances, we have gathered the information obtained using the procedure outlined above in descriptions to document the corresponding archetypes. In order to produce comprehensive and unambiguous descriptions, we have employed OWL language [38] expressions based on SNOMED CT terms whenever it was possible (e.g. in the case of terminology abstractions). Tables II, III, and IV list the concepts involved in the CT with their respective informal and formal (OWL) descriptions, both using SNOMED CT terms. The list of concepts is exhaustive, except for the lower-level terms on which *severe comorbidity* is based (in Tables III and IV).

It is noteworthy that we have identified different types of concept descriptions,

being the most frequent one expressions combining several SNOMED CT terms using set theory operations such as union (`or`) and set difference (`and not`, i.e. an intersection followed by a complement). Thus, union expressions capture the specific problems included in two (or more) categories of problems, and set difference expressions represent the problems that fall in one first category but not in a second one (see e.g. *colorectal cancer* and *colorectal polyposis* in Table II). Additionally, we have used expressions referring to some SNOMED CT term with a particular qualifier, typically *severity*. The previous expressions can be properly described in OWL. However, complex terms that are calculated based on the existence of two or more independent problems are not expressible in OWL. An example is the concept *metastatic solid tumor*, which depends on the existence of both a (primary) solid tumor and a secondary tumor or metastasis (see Tables III and IV).

5.3. Design of a PROforma plan for the CT

A CDSS for patient eligibility determination can be implemented in the PROforma language in a straightforward way. PROforma is among the leading languages for guideline representation according to the literature [20]. In a wider sense, it can be considered as a language for modelling clinical processes. In PROforma these are modelled in terms of hierarchically organised tasks [19]. PROforma tasks fall into four main categories, namely: *actions*, *enquiries*, *decisions* and *plans*. *Actions* and *enquiries* represent the basic interactions with the environment: actions are used to initiate some external procedure, human or auto-

Table II: Description of concepts involved in the case study.

Concept	**Informal definition using SNOMED CT terms**	**OWL definition using SNOMED CT terms**
colorectal adenoma	*any* Adenoma of large intestine[a]	`'Adenoma of large intestine'`
colorectal cancer	*any* Malignant tumor of colon[a] *or* *any* Malignant tumor of rectum[a]	`'Malignant tumor of colon' or 'Malignant tumor of rectum'`
colorectal polyposis	*any* Intestinal polyposis syndrome[a] *except* *any* Polyp of small intestine[a]	`'Intestinal polyposis syndrome' and not 'Polyp of small intestine'`
familial colorectal cancer	Family history of cancer of colon[b], [c]	N/A
family history of colorectal polyposis	Family history of polyp of colon[b], [c]	N/A
inflammatory bowel disease	*any* Inflammatory bowel disease[a]	`'Inflammatory bowel disease'`
Lynch syndrome	Hereditary nonpolyposis colon cancer[a]	`'Hereditary nonpolyposis colon cancer'`
previous total colectomy	*any* Total colectomy[d]	`'Total colectomy'`
severe comorbidity	Charlson index *greater or equal than* 6, calculated as the sum score of the morbidities of the patient (see lower-level terms in Table III)	N/A

[a]Disorder.
[b]Situation.
[c]Approximate definition, due to missing SNOMED CT terms.
[d]Procedure.

mated, and enquiries serve to obtain information about the environment, be it from the user or from a database. *Decisions* are points at which some choice has to be made based on reasons for and/or against the different alternatives or candidates. Finally, *plans* can be used to group together other tasks and are thus a key element for the hierarchical organisation of tasks. Furthermore, tasks may have a number of properties that determine the way they will be executed, such as preconditions and scheduling constraints.

Patient eligibility determination can be implemented in PROforma using merely one decision task followed by two action tasks corresponding to the

Table III: Description of concepts required for the definition of *severe comorbidity* concept.

Concept	**Informal definition using SNOMED CT terms**	**OWL definition using SNOMED CT terms**
AIDS	*any* AIDS[a]	`'AIDS'`
any tumor	*any* Malignant neoplastic disease[a]	`'Malignant neoplastic disease'`
diabetes with end organ damage	*any* Diabetic neuropathy[a] *or* *any* Diabetic oculopathy[a] *or* *any* Diabetic renal disease[a] *or* *any* Peripheral circulatory disorder associated with diabetes mellitus[a], [b]	N/A
hemiplegia	*any* Hemiplegia[a]	`'Hemiplegia'`
leukemia	*any* Leukemia[a]	`'Leukemia'`
lymphoma	*any* Malignant lymphoma[a]	`'Malignant lymphoma (clinical)'`
metastatic solid tumor	*exists any* Primary solid tumor *and* *exists any* Metastatic tumor (see lower-level terms in Table IV)	N/A
moderate or severe liver disease	*any* Disease of liver[a] *with* Severity[c] Moderate, Moderate to severe *or* Severe[d]	`'Disease of liver' and ('Severity' some ('Moderate' or 'Moderate to severe' or 'Severe'))`
moderate or severe renal disease	*any* Kidney disease[a] *with* Severity[c] Moderate, Moderate to severe *or* Severe[d]	`'Kidney disease' and ('Severity' some ('Moderate' or 'Moderate to severe' or 'Severe'))`

[a]Disorder.
[b]Approximate definition, due to missing SNOMED CT terms.
[c]Attribute.
[d]Qualifier values.

exclusion and inclusion outcomes. To illustrate this, Figure 2 shows the key elements of the PROforma plan for our case study in the PROforma textual notation. A decision requires the specification of both the candidates of the decision (in this case, `exclude_patient` and `include_patient`) and the different arguments for and/or against them. In turn, an argument consists of a logical expression (e.g. `colorectal_cancer_present = ''true''`) and a support mode (e.g. `for`) specifying the conditions under which a candidate (e.g. `exclude_patient`) must be selected or discarded. Additionally, a series of recommendation rules for the candidates plus

Table IV: Description of concepts required for the definition of *metastatic solid tumor* concept.

Concept	**Informal definition using SNOMED CT terms**	**OWL definition using SNOMED CT terms**
metastatic tumor	*any* Secondary malignant neoplastic disease[a]	`'Secondary malignant neoplastic disease'`
primary solid tumor	*any* Malignant neoplastic disease[a] *except* *any* Secondary malignant neoplastic disease[a] *except* *any* Malignant tumor of lymphoid hemopoietic and related tissue[a]	`'Malignant neoplastic disease' and not ('Secondary malignant neoplastic disease' or 'Malignant tumor of lymphoid hemopoietic and related tissue')`

[a]Disorder.

a choice mode (single vs. multiple selection) must be specified in the decision. Finally, to couple the decision with the subsequent actions, the latter must include a precondition to ensure that they will only be executed in case the corresponding candidate has been selected in the decision (e.g. `result_of(inclusion_decision) = exclude_patient`).

In the above implementation the CT exclusion criteria have been all encoded as arguments for the candidate `exclude_patient`. These arguments contain a logical expression that refers to data elements coming from (or derivable from) the EHR, such as `colorectal_cancer_present`. The LinkEHR transformation engine is the module responsible for providing a value for these data elements, in accordance with the specified mappings (see section 5.5). The final decision is calculated by counting all the arguments for the candidate `exclude_patient`, resulting in the exclusion of the patient if the number of arguments for this candidate is equal or greater than one (`exclude_patient` recommendation is `netsupport(inclusion_decision, exclude_patient)` $\geq$ `1`). Conversely, the inclusion decision is made if the number of arguments for the

```
/** PROforma (plain text) version 1.7.0 **/

plan  :: 'NCT00906997_plan' ;
     caption :: "NCT00906997_plan" ;
     description :: "" ;
     component :: 'NHC_enquiry' ;
          caption :: "NHC_enquiry" ;
          task_definition :: 'NHC_enquiry' ;
     component :: 'main_enquiry' ;
          caption :: "main_enquiry" ;
          task_definition :: 'main_enquiry' ;
          schedule_constraint :: completed('
               NHC_enquiry') ;
     component :: 'inclusion_decision' ;
          caption :: "inclusion_decision" ;
          task_definition :: 'inclusion_decision' ;
          schedule_constraint :: completed('
               main_enquiry') ;
     component :: 'include_patient_action' ;
          caption :: "include_patient_action" ;
          task_definition :: 'include_patient_action
               ' ;
          schedule_constraint :: completed('
               inclusion_decision') ;
     component :: 'exclude_patient_action' ;
          caption :: "exclude_patient_action" ;
          task_definition :: 'exclude_patient_action
               ' ;
          schedule_constraint :: completed('
               inclusion_decision') ;
end plan.

action  :: 'exclude_patient_action' ;
     caption :: "exclude_patient_action" ;
     precondition ::result_of(inclusion_decision) =
          exclude_patient;
end action.

action  :: 'include_patient_action' ;
     caption :: "include_patient_action" ;
     precondition ::result_of(inclusion_decision) =
          include_patient;
end action.
```

```
decision  :: 'inclusion_decision' ;
     caption :: "inclusion_decision" ;
     candidate :: 'include_patient' ;
          recommendation ::netsupport(
               inclusion_decision, exclude_patient)
               < 1;
     candidate :: 'exclude_patient' ;
          argument :: for,colorectal_adenoma_present
                = "true"  attributes
               argument_name :: '
                    exclude_patient_Arg_04' ;
          end attributes ;
          argument :: for,colorectal_cancer_present
                = "true"  attributes
               argument_name :: '
                    exclude_patient_Arg_05' ;
          end attributes ;
          ...
          argument :: for,severe_comorbidity_present
                = "true"  attributes
               argument_name :: '
                    exclude_patient_Arg_09' ;
          end attributes ;
          recommendation ::netsupport(
               inclusion_decision, exclude_patient)
               >= 1;
end decision.

enquiry  :: 'main_enquiry' ;
     caption :: "main_enquiry" ;
     source :: 'colorectal_adenoma_present' ;
          caption :: "colorectal_adenoma_present?" ;
          data_definition :: 'textType' ;
     source :: 'colorectal_cancer_present' ;
          caption :: "colorectal_cancer_present?" ;
          data_definition :: 'textType' ;
     ...
     source :: 'severe_comorbidity_present' ;
                caption :: "
                     severe_comorbidity_present?"
                     ;
                data_definition :: 'textType' ;
end enquiry.
...
```

Figure 2: Details of the PROforma plan for CT patient eligibility determination.

same candidate is less than one (`include_patient` recommendation is `netsupport(inclusion_decision, exclude_patient) < 1`).

5.4. Design of CT archetypes

The case study requires the design of a series of archetypes suitable for the decision-support tasks carried out in the above PROforma plan. In this stage we have used as a tool the LinkEHR archetype editor. As already mentioned, we have started from the CKM archetype `openEHR-EHR-EVALUATION.problem.v1`, which fits well with the kind of clinical concepts identified in the CT

(see section 5.2). The archetype has been specialised to meet the data needs of the PROforma plan. This specialisation, named `openEHR-EHR-EVALUATION.problem-DS.v1`, incorporates a boolean element to store the presence/absence of the problem plus a numeric element to record the associated comorbidity score, if required. The latter archetype has been in turn specialised in a number of specific archetypes, one for each of the identified concepts (i.e. `openEHR-EHR-EVALUATION.problem-DS-colorectal_cancer.v1`, and so on). In this way we seek to achieve a "separation of concerns" in the design of archetypes, and ultimately to facilitate the subsequent mapping process. Note that although from a definitional point of view some of the concepts are based (and hence depend) on another, the corresponding archetypes have been designed as independent objects.

5.5. Mapping of CT archetypes to a summary health record

The archetypes from the previous step are conceived to connect the PROforma CDSS with alternative EHRs. To accomplish this, we have used the LinkEHR archetype mapping tool to define a set of mappings relating the (target) archetype elements to the (source) data items of the EHR under consideration. The EHR schema that we have chosen in our case study is part of a normalisation project carried out at Hospital de Fuenlabrada (Spain). The schema has been designed as a summary health record and integrates the list of problems and medications of the patient, as reflected in different health information systems (primary care systems, hospital systems, and medication databases) [39]. The only difference

with respect to the schema in use at Hospital de Fuenlabrada (HF) is that we assume a SNOMED CT encoding of patient problems.

For mapping purposes we can make a distinction between the archetypes corresponding to terminology abstractions and those corresponding to definitions of abstract terms. The former have been termed *first-level/base archetypes*, since their value can be obtained directly from EHR data by means of rather simple expressions. The latter, which require data that can be derived from the EHR but are not available as such, have been named depending on the level of the archetypes they use as source. E.g. *second-level archetypes* use first-level archetypes (and possibly EHR data items), and so forth. Figure 3 depicts the dependences among the different archetypes and/or the EHR. Although not strictly necessary, the mapping process was carried out starting with first-level archetypes and continuing with second-level ones, and so forth. Thereby we were able to validate at each step the XQuery transformation script generated by the LinkEHR mapping tool.

As an illustration, Table V shows part of the mapping functions for the first-level archetype `openEHR-EHR-EVALUATION.problem-DS-metastatic_tumor.v1`, which uses HF summary health record as source schema. Concretely, this mapping corresponds to the boolean element storing the presence/absence of a metastatic tumor. According to the definition listed in Table IV, this problem is present if the patient record stores any problem within the category `'Secondary malignant neoplastic disease'`. The mapping condition checks the number of occurrences of such problems,

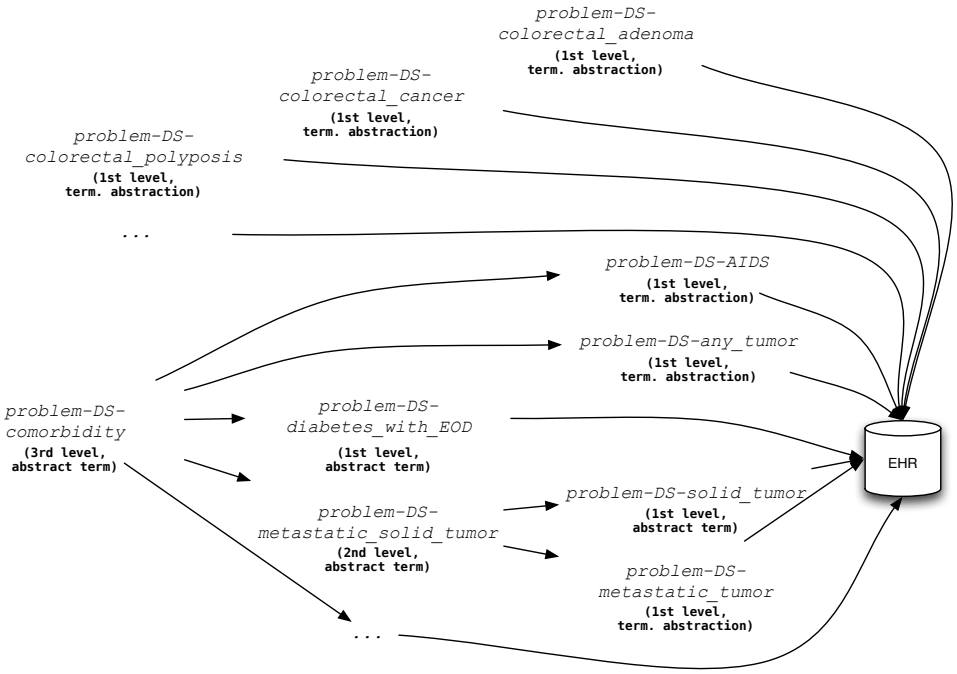

Figure 3: Graphical representation of archetype dependences in terms of the sources they use. Note that the full archetype names include the prefix `openEHR-EHR-EVALUATION.` as well as the suffix `v1.adl`.

if greater than zero, in both primary care problems (in the archetype path `$context/resumida/problemas_OMI/problema_OMI`) and hospital ones (in `$context/resumida/problemas_SELENE/problema_SELENE`). The path expressions use the variable `$context`, which refers to a path in the source and sets the context of the data for a particular patient. Note also that two parameters are used in `@count` expressions: the context to be used for counting and a condition specifying the elements to be counted. Additionally, the expression `@descendents("128462008")` has been used to obtain the SNOMED CT codes for the category itself and for all of its subcategories.

Table V: A mapping transforming the problem codes from an XML summary health record to a boolean value indicating the presence/absence of a metastatic tumor. This mapping corresponds to the first-level archetype openEHR-EHR-EVALUATION.problem-DS-metastatic_tumor.v1.

Condition	Mapping Function
(@count($context/resumida/problemas_OMI/problema_OMI, @in($context/resumida/problemas_OMI/problema_OMI/codigo, @descendents("128462008"))) + @count($context/resumida/problemas_SELENE/problema_SELENE, @in($context/resumida/problemas_SELENE/problema_SELENE/codigo, @descendents("128462008")))) > 0	TRUE
TRUE	FALSE

Table VI shows the mappings for the presence/absence element of a second-level archetype, namely openEHR-EHR-EVALUATION.problem-DS-metastatic_solid_tumor.v1. In this case, the element value depends on the presence of both a primary solid tumor and a metastatic one (see definition in Table III). The latter information is derived from the HF summary health record by means of the mappings defined for the first-level archetypes openEHR-EHR-EVALUATION.problem-DS-solid_tumor.v1 and openEHR-EHR-EVALUATION.problem-DS-metastatic_tumor.v1. Consequently, these archetypes have been used as data sources in the mapping of openEHR-EHR-EVALUATION.problem-DS-metastatic_solid_tumor.v1. Notice that in this case the mapping conditions can be hard to read, due to the fact that archetypes are used as sources. However, in practice the user relies on an editing tool that allows entering archetype paths by simply browsing the archetype nodes and clicking on the appropriate one (e.g. in all openEHR-EHR-EVALUATION.problem-DS.v1 specialisations, the path ending with [at0000.1.1]/data[at0001]/items[at0.12]-

/value[at0.13]/value points to the boolean value storing the presence/absence of the problem).

Table VI: A mapping transforming the presence/absence values of two source archetypes to a boolean value indicating the presence/absence of a metastatic solid tumor. This mapping corresponds to the second-level archetype openEHR-EHR-EVALUATION.problem-DS-metastatic_solid_tumor.v1.

Condition	Mapping Function
(/entity_data_root[ENTITYDATAROOTat]/ ... problem_ds_metastatic_tumor___v1[at0000.1.1]/data[at0001]/ items[at0.12]/value[at0.13]/value = "true") AND (/entity_data_root[ENTITYDATAROOTat]/ ... problem_ds_primary_solid_tumor___v1[at0000.1.1]/data[at0001]/ items[at0.12]/value[at0.13]/value = "true")	TRUE
TRUE	FALSE

6. Discussion

We have developed all the necessary components to implement a prototype for the determination of patient eligibility in the framework of a CT for colorectal cancer screening, which is designed to operate by taking as input patient data as stored in a real-life EHR system. Firstly, we have implemented a CDSS for patient eligibility determination in the PROforma language, using basically one decision element. Second, we have developed a set of openEHR archetypes tailored to the clinical concepts in the CT and at the same time suitable for the decision-support tasks of the CDSS. Third, starting with an EHR schema in use in a Spanish hospital, we have defined the necessary mappings to generate archetype-compliant instances for use in the PROforma CDSS. Finally we have performed a series of tests of the prototype, checking mainly the proper functioning of the instance generation scripts created by LinkEHR but also the smooth access of the PROforma

execution engine to the generated archetype instances. For the latter we have employed a mediator module which allows connecting the engine with any kind of external data source, including (XML) archetype instances.

Representation of clinical concepts. As representation language for the description of clinical concepts, we have used jointly the OWL language and the SNOMED CT terminology. In most cases we have obtained adequate OWL+SNOMED CT expressions. Exceptions are a few terms that are not included in SNOMED CT (e.g. *family history of cancer of rectum*), surely for well-founded reasons at the discretion of the developers, and some definitions of abstract terms using arithmetic or logical expressions beyond the expressive power of OWL (e.g. *metastatic solid tumor* would require existential quantification over two variables). Concerning the lack of a SNOMED CT concept for specific terms, the preferred solution in our approach is to use post-coordination at the SNOMED level. An additional option to consider is using post-coordination at the archetype level. As regards to the OWL language, it is well-suited to describe terminology abstractions as well as concept definitions based on set theory operations. However, it is clear that an alternative language has to be used to meet the needs of other kind of definitions. On the other hand, in our view the choice of SNOMED CT is beyond question, as it has recently emerged as global standardised terminology.

Design of archetypes. With regard to the archetypes developed, we have chosen to keep the entire structure of the archetype used as starting point

(openEHR-EHR-EVALUATION.problem.v1) in our specialisations. In this way we intend to leave open the possibility of reuse of our archetypes in other different contexts, e.g. by applications in health-care settings. Moreover, the archetypes correspond to actual clinical concepts, which should also increase the chances of reuse. To support reuse, the definitions we have obtained in the stage of concept analysis can be of great help if included as documentation within the archetypes. Although we have not yet provided a solution for this, the option under consideration is to incorporate an additional element to our specialisations.

Mapping of archetypes. Likewise, concept definitions are crucial in the mapping stage. As shown in section 5.5, the expressions in mapping conditions bear some resemblance with the definitions of the corresponding clinical concepts. For instance, an informal concept definition like '*any* <disease>' must be translated into an expression to determine whether a patient suffers from any problem within the category <disease>, including subcategories. This results in the following expression pattern, which checks the number of occurrences of these problems among the values of a particular element and for a given context:

```
@count(<context>, @in(<ehr-element>,
                  @descendents(<disease-SNOMED-code>))) > 0
```

Such mapping expressions, which in our case seem to be characteristic of first-level archetypes, can be reused to a large extent when mapping the archetypes to other EHRs, by replacing the EHR-specific parts (i.e. <context> and <ehr-element>). The situation is more favourable in the case of second

and further level archetypes, which do not depend on the EHR but on other archetypes. Consequently, the mappings we have developed can be reused as they are. An important issue to resolve in this case concerns visualisation, since resulting expressions when (possibly multiple) archetypes are used as source can be difficult to read.

To conclude discussion on mapping issues, it is worth noting that although archetypes are defined as maximal data sets, with slots for any data item that can be possibly required for a concept, the mapping definitions do not have to be necessarily exhaustive. This is because mapping is only required for those archetype nodes which have been defined as mandatory in either the RM, the parent archetype, or the archetype itself.

Prototype tests. We have carried out a series of tests of the archetype instances' generation scripts created by LinkEHR. These tests have been limited to simulated data, although using an EHR schema currently in production in a Spanish hospital. Our tests have shown that the response time of queries is affordable. However, more comprehensive tests in a realistic setting are still pending. On the other hand, the tests have served to debug and validate the add-ons implemented specifically for our case study, and ultimately as a proof-of-concept of our approach. The add-ons are all new LinkEHR functionalities, falling into two main categories: new functions for the description of mappings, and integrated support for using archetypes as data sources.

Some of the new LinkEHR mapping features have proven to be crucial for our purposes. One is the implementation of a basic SNOMED CT query module

(limited to the *is-a* relationships), which allows to refer to e.g. the descendants of a given concept/category in mapping expressions (the previously mentioned `@descendents` function). Even more important is the possibility of using multiple archetypes as source schema. This feature has been extensively used in our case study, and is expected to continue to be so in other CDSS interoperability projects.

Scope and limitations of this work. The approach presented in this article has been put into practice in a case study for CT patient recruitment, resulting in a prototype with the desired interoperability characteristics. Despite having used a single case study, the approach is rather generic and thus applicable to other CDSSs, possibly in other clinical domains. To take one example, the approach is currently being successfully applied to a CDSS for colorectal cancer risk assessment, which is based on the results of colonoscopy (and other) tests and therefore requires completely different clinical parameters. In this case the definitions of complex concepts use aggregation functions such as counting and maximum, which are within the functionalities of LinkEHR. It should be noted that no new functionalities were needed for this CDSS. LinkEHR has also proved satisfactory in applications that handle numerical data, e.g. dealing with the numerical values and units of medication information [30]. In our experience, the concept definitions involved in our case study (some beyond the expressive power of OWL) are among the hardest to deal with for the interoperability of CDSSs and EHRs.

In principle the proposed approach can fit in service-oriented environments where a CDSS is offered as a set of services [40]. In such environments one of the

most important services is the decision support one, which receives patient data as input and produces a set of patient-specific conclusions as output. Other useful services are the common terminology service or the entity identification one. The mapping and abstraction capabilities presented in this article could be encapsulated as a service providing standardisation and abstraction functionalities. This service would offer operations at the concept model level (and also at the information model one), thus making decision support independent of the particular details of clinical data sources.

One limitation of our work is the basic SNOMED CT reasoning we use, restricted to reasoning over the is-a hierarchy of concepts. This has been sufficient for terminology abstractions and for concept definitions using set operations, but would not be enough if more advanced support is required. An example could be to determine if two SNOMED CT expressions, e.g. pre- and post-coordinated ones, are equivalent. If this is a requirement, we envisage the integration of an external reasoning service providing an adequate support within our architecture.

7. Related work

Standardisation of the VHR is regarded as an important issue [18], particularly in the definition of the VHR global schema. Several initiatives have based their VHR on standard EHR architectures. The use of a simplified version of HL7 RIM is the prevailing option. The KDOM framework [34], the MEIDA architecture [41], and the works by Lonsdale *et al.* [42] and by Cho *et al.* [43], [44] are remarkable examples of this alternative. All these approaches use a VHR based on

a small subset of HL7 RIM classes, which are also simplified, to make the CDSS compatible with different EHRs. Although EHR architectures impose a tree-like structure to health data, these approaches use a flat relational view over the source EHR to build the VHR. This hides the original semantics of the HL7 RIM, apart from generating data redundancy in the virtual schema. Another approach, that stands apart from the previous ones, is the EGADSS system [45]. It uses HL7 CDA for building the VHR, resulting in a more structured VHR. However, it uses a single type of document (a patient summary) that contains a fixed set of patient data as VHR. The source EHR systems generate this document in XML in response to a clinical event, such as the beginning of a patient encounter.

HL7 acknowledges the difficulties of using EHR standards to define VHRs, and consequently is currently working on the specification of a Virtual Medical Record (VMR) [46] that aims at facilitating the reuse of exiting healthcare data in CDSSs. The HL7 VMR is an information model inspired by existing HL7 version 3 standards which has been designed to accommodate non HL7-based sources as well. It defines a simple model for representing patient data, and at the same time comes with powerful context specification capabilities. It does not try to represent every possible entry in the EHR, although it is powerful enough to represent a large percentage of decision support needs.

The use of a VHR based on a standard EHR architecture is necessary but not sufficient for semantic interoperability between CDSSs and EHRs. The main problem is the partitioning of concepts between the information model (EHR architecture) and terminology [47], [48]. Every different partition decision yields

as a result a different valid representation of data with respect to the information model. In order to solve this problem it is necessary to make explicit all the assumptions about the representation of data. Thus, high-level domain concepts' descriptions are needed rather than generic concepts as those provided by EHR architectures. The SAGE project [48], [49] was the first work to consider this problem. They proposed the use of Detailed Clinical Models (DCMs) as a way of defining explicitly the unique data representation expected by the guidelines. In SAGE the VHR is composed of a set of DCMs, all of them derived from the information model classes by constraining the value of properties. Each DCM describes a particular domain concept, such as diagnosis or blood pressure measurement, needed by the CDSS. In SAGE, as in most of the previous initiatives, the information model is composed of a subset of HL7 RIM classes which are also simplified. SAGE DCMs are also employed by Lonsdale *et al.* for evaluating CT eligibility criteria against EHR systems. DCMs are similar in purpose to archetypes or CDA templates in many ways [50].

Our approach is similar to the idea of these latter platforms which both use clinical models for specifying the VHR. Two distinctive features of our work are i) the utilisation of the full-fledged archetype framework (including inheritance, reutilisation and composition of archetypes, semantic validation, and terminology bindings) for specifying the VHR instead of simple unrelated concept definitions, and ii) the support for any information model, as long as an XML Schema is available, instead of a fixed simplified one as in the previous platforms. Figure 4 compares the main approaches, specifically: those requiring substantial CDSS

modifications (see (a) in Figure 4), like the Arden Syntax, those exploiting a VHR based on some generic EHR architecture (see (b)), like KDOM or MEIDA, and finally those that make use of domain concepts for defining the VHR (see (c)), as our approach.

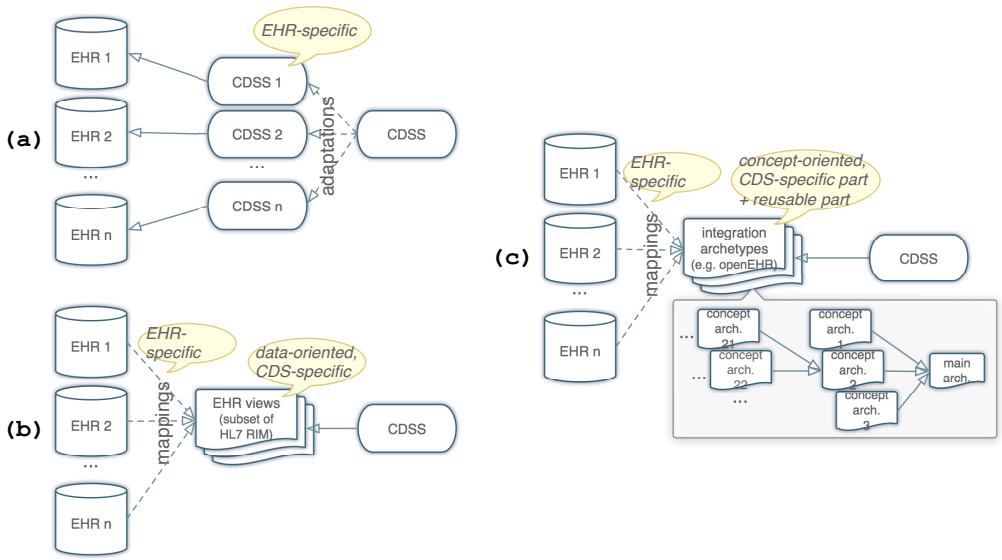

Figure 4: Linking a CDSS to different EHRs (adapted from Fig. 1 by Peleg *et al.* [34]): (a) adapting the CDSS to each EHR; (b) through data views based on (a subset of) HL7 RIM; and (c) through reusable concept views based on archetypes (e.g. openEHR).

Mapping source EHR data to the CDSS is a very complex task, due to the differences and mismatches between heterogeneous formats, models, abstractions levels, and semantics. In the case of a generic VHR (as in KDOM or MEIDA), two levels of mappings have to be considered: from the EHR to the VHR, and from the VHR to the CDSS. The construction of the VHR from the source EHR is a manual process in all the projects that use a generic VHR. In our approach the VHR is composed of a set of domain concepts (archetypes) tailored to a particular

CDSS, although with a high potential of reuse between different CDSSs, and at the same time based on a generic information model. Since our VHR is composed of a set of concepts, the same tools and mapping language are used throughout the whole mapping process, from raw EHR data to the potentially highly abstract CDSS concepts. The mapping capabilities differ also a lot among the analysed projects. EGADSS only supports basic one-to-one mappings expressed in XPath, and the approach by Cho *et al.* provides very basic concept-level mappings. The approach by Lonsdale *et al.* only supports the definition of abstract terms from basic concepts. MEIDA supports powerful temporal abstractions but basic terminology ones, and provides no support for mapping composition and/or reutilisation. KDOM provides terminology abstractions and definition of abstract terms, as well as mapping composition and reutilisation, but only supports basic temporal abstractions. Thanks to the LinkEHR platform, our approach provides support for all the previous mapping capabilities. This includes structural mappings, i.e. mappings that rule structural transformations between complex structures such as classes of information models. Note that no other approach supports this feature, which has proved essential for our purposes. Table VII summarises the solutions and capabilities provided by the main approaches we have analysed.

8. Conclusions

In this article we introduce a comprehensive approach, including a set of tools as well as methodological guidelines, to deal with the interoperability of CDSSs and EHRs based on archetypes. Archetypes are used to build a conceptual layer

Table VII: Comparative analysis of the main approaches to interoperability of CDSSs and EHRs using a VHR (in chronological order).

Features		EGADSS	KDOM	Lonsdale *et al.*	MEIDA	Cho *et al.*	our approach
Source EHR	EHR data model	any	Relational	Relational	Relational	Relational	Relational or XML
VHR	VHR reference model	Fixed CDA document (Patient Summary Document)	Subset of HL7-RIM	Subset of HL7-RIM	Subset of HL7-RIM	Subset of HL7-RIM	any
	VHR generation from EHR	manual	manual	manual	manual	Semi-automatic from high-level declarative mappings	Semi-automatic from high-level declarative mappings
	Clinical models in VHR	no	no	yes, Detailed Clinical Models	no	no	yes, archetypes
	VHR instances	XML instances of the Patient Summary Document	Flat relational view of HL7-RIM	Nested name-value pairs	Flat relational view of HL7-RIM	Flat relational view of HL7-RIM	XML instances compliant with reference model and clinical models
Mapping features & mapping types	High-level declarative mapping language	no	yes	no	yes	basic	yes
	Mapping execution language	XPath	SQL	proprietary	SQL	SQL	XQuery
	Extension of mapping functions	no	yes	no	no	no	yes
	Structural mappings	no	no	no	no	no	yes
	Concept-level mappings	no	no (just attribute level)	yes	no (just attribute level)	basic	yes
	Mapping composition	no	yes	no	no	no	yes
	Mapping reutilisation	no	yes	no	no	no	yes
	Terminology abstractions	no	yes	no	yes	no	yes
	Temporal abstractions	no	basic	no	yes	no	basic
	Definition of abstract terms	no	yes	yes	no	no	yes
Query	Automated query generation	no	yes	yes	yes	yes	yes

of the kind of a VHR over the EHR whose contents need to be integrated and used in the CDSS, associating them with structural and terminology-based semantics –what might be termed *knowledge-rich clinical models based on archetypes*. Subsequently, the archetypes are mapped to the EHR by means of an expressive mapping language and specific-purpose tools. In the article we also describe a case study where the tools and methodology have been employed in a CDSS to support patient recruitment in the framework of a CT for colorectal cancer screening.

The utilisation of archetypes not only has proved satisfactory to achieve interoperability between CDSSs and EHRs but also offers benefits of varying nature. From a data model perspective, the utilisation of archetypes brings about several advantages over similar initiatives. First, the VHR/data models we work with are of a higher level of abstraction (clinical concept level instead of RM one) and can incorporate semantic descriptions (through terminology references). Second, archetypes can potentially deal with different EHR architectures (e.g. CEN/ISO EN13606, openEHR or HL7 CDA), due to their deliberate independence of the RM. Third, no matter what RM is used, the archetype instances we obtain are valid instances of the underlying RM, which would enable e.g. feeding back the EHR with data derived by abstraction mechanisms. Lastly, the medical and technical validity of archetype models would be assured, since in principle clinicians should be the main actors in their development.

In the future we intend to work on different kinds of enhancements to our approach. On one hand we plan to integrate methodologies and tools to deal with an explicit domain (or concept) model, as well as with the interactions thereof

with the archetype (or information) model, in line with the conceptual framework proposed by Rector [47]. On the other hand we envisage to deal with efficiency issues, to ensure that the response time when handling realistic clinical databases is affordable. Also related to the functionalities of our tools, we plan to address the bidirectional interaction of the CDSS with the clinical information system, e.g. the CDSS feedback to the EHR which has been mentioned above.

Acknowledgements

This research has been supported by the Spanish Ministry of Education through grant PR2010-0279, and by Universitat Jaume I through project P11B2009-38. Additionally, this research has been supported by the Spanish Ministry of Science and Innovation under grant TIN2010-21388-C02-01, and by the Spanish Ministry of Economy and Competitiveness under grant PTQ-11-04987.

We would like to thank M.D. Carlos Ferrer-Albiach, Oncology Specialist and Director of the Oncological Institute at Hospital Provincial of Castellón (Spain), for his invaluable help in the description of the clinical concepts of our case study. We are also grateful to Prof. John Fox, from the University of Oxford (United Kingdom), for providing us with an academic license of the PROforma tools. Finally, we would like to thank Dr. Juan M. García-Gómez, from the Biomedical Informatics Group at ITACA Institute (Spain), for granting us access to the PROforma mediator module developed by his group.

References

[1] M. A. Musen, Y. Shahar, and E. H. Shortliffe, *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, ch. Clinical Decision-Support Systems, pp. 698–736. Health Informatics, Springer, 3rd ed., 2006.

[2] E. S. Berner and T. J. La Lande, *Clinical Decision Support Systems*, ch. Overview of Clinical Decision Support Systems. Health Informatics, Springer, 2nd ed., 2007.

[3] W. Sujansky, "Heterogeneous Database Integration in Biomedicine," *J Biomed Inform*, vol. 34, pp. 285–298, Aug. 2001.

[4] G. Schadow, D. Russler, C. Mead, and C. McDonald, "Integrating medical information and knowledge in the HL7 RIM," in *Proc. of the AMIA 2000 Annual Symposium*, pp. 764–768, 2000.

[5] F. Sonnenberg and C. Hagerty, "Computer-Interpretable Clinical Practice Guidelines. Where Are We and where Are We Going?," *IMIA Yearbook of Medical Informatics*, pp. 145–158, 2006.

[6] International Organization for Standardization, "Health informatics — Electronic health record communication — Part 1: Reference model." In `http://www.iso.org/iso/catalogue_detail.htm?csnumber=40784`, 2008.

[7] The openEHR Foundation, "openEHR: future proof and flexible EHR specifications." `http://www.openehr.org/`.

[8] Health Level Seven International, "HL7 Clinical Document Architecture Release 2.0 (CDA R2)." In `http://www.hl7.org/implement/standards/cda.cfm`, 2005.

[9] Clinical Data Interchange Standards Consortium, "Specification for the Operational Data Model (ODM)." In `http://www.cdisc.org/odm`, 2010. Last accessed: August 2012.

[10] T. Beale, "Archetypes. Constraint-based Domain Models for Future-proof Information Systems." In `http://www.openehr.org/publications/archetypes/archetypes_beale_web_2000.pdf`, 2001. Date of access: December 2011.

[11] M. Marcos and B. Martínez-Salvador, "Towards the Interoperability of Computerised Guidelines and Electronic Health Records: An Experiment with openEHR Archetypes and a Chronic Heart Failure Guideline," in *Knowledge Representation for Health-Care* (D. Riaño, A. ten Teije, S. Miksch, and M. Peleg, eds.), vol. 6512 of *LNCS*, pp. 101–113, Springer, 2011.

[12] M. Marcos, J. A. Maldonado, B. Martínez-Salvador, D. Moner, D. Boscá, and M. Robles, "An Archetype-Based Solution for the Interoperability of Computerised Guidelines and Electronic Health Records," in *Proc. of the 13th Conference on Artificial Intelligence in Medicine (AIME 2011)* (M. Peleg, N. Lavrač, and C. Combi, eds.), vol. 6747 of *LNCS*, pp. 276–285, Springer, 2011.

[13] T. Pryor and G. Hripcsak, "The Arden Syntax for Medical Logic Modules," *Int J Clin Monit Comput*, vol. 10, pp. 215–224, 1993.

[14] G. Hripcsak, P. Ludemann, T. A. Pryor, O. B. Wigertz, and P. D. Clayton, "Rationale for the Arden Syntax," *Comput Biomed Res*, vol. 27, pp. 291–324, 1994.

[15] M. Lenzerini, "Data integration: a theoretical perspective," in *Proc. of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 233–246, ACM, 2002.

[16] G. Wiederhold, "Mediators in the Architecture of Future Information Systems," *Computer*, vol. 25, pp. 38–49, Mar. 1992.

[17] C. Angulo, P. Crespo, J. Maldonado, D. Moner, D. Pérez, I. Abad, J. Mandingorra, and M. Robles, "Non-invasive lightweight integration engine for building EHR from autonomous distributed systems," *Int J Med Inform*, vol. 76, no. S3, pp. 417–424, 2007.

[18] P. D. Johnson, S. W. Tu, M. A. Musen, and I. Purves, "A virtual medical record for guideline-based decision support," in *Proc. of the AMIA 2001 Annual Symposium*, pp. 294–298, 2001.

[19] D. Sutton and J. Fox, "The syntax and semantics of the PROforma guideline modeling language," *J Am Med Inform Assn*, vol. 10, no. 5, pp. 433–443, 2003.

[20] M. Peleg, S. Miksch, A. Seyfang, J. Bury, P. Ciccarese, J. Fox, R. Greenes, R. Hall, P. Johnson, N. Jones, A. Kumar, S. Quaglini, E. Shortliffe, and M. Stefanelli, "Comparing computer-interpretable guideline models: A case-study approach," *J Am Med Inform Assoc*, vol. 10, pp. 52–68, 2003.

[21] D. Kalra, T. Beale, and S. Heard, *Regional Health Economies and ICT Services: The PICNIC Experience*, vol. 115 of *Studies in Health Technology and Informatics*, ch. The openEHR Foundation, pp. 153–173. IOS Press, 2005.

[22] The openEHR Foundation, "Architecture Overview." In `http://www.openehr.org/releases/1.0.1/architecture/overview.pdf`, Apr. 2007.

[23] The openEHR Foundation, "Archetype Definitions and Principles." In `http://www.openehr.org/releases/1.0.2/architecture/am/archetype_principles.pdf`, Mar. 2007.

[24] H. Leslie and S. Heard, "Archetypes 101," in *Bridging the Digital Divide: Clinicians, Consumers, Computers - Proc. of the 2006 Health Information Conference (HIC 2006)* (Westbrook, J. *et al.*, ed.), 2006.

[25] The openEHR Foundation, "Archetype Definition Language ADL 2." In `http://www.openehr.org/releases/1.0.2/architecture/am/adl2.pdf`, Mar. 2007.

[26] The openEHR Foundation, "openEHR Clinical Knowledge Manager." http://www.openehr.org/knowledge/.

[27] Biomedical Informatics Group, Universidad Politécnica de Valencia, "LinkEHR Normalization Platform." http://www.linkehr.com/. Last accessed: August 2012.

[28] J. A. Maldonado, D. Moner, D. Boscá, J. Fernández, C. Angulo, and M. Robles, "LinkEHR-Ed: A multi-reference model archetype editor based on formal semantics," *Int J Med Inform*, vol. 78, no. 8, pp. 559–570, 2009.

[29] J. A. Maldonado, D. Moner, D. Boscá, C. Angulo, L. Marco, E. Reig, and M. Robles, "Concept-Based Exchange of Healthcare Information: The LinkEHR Approach," in *Proc. of the 1st IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB 2011)*, pp. 150–157, July 2011.

[30] J. A. Maldonado, C. M. Costa, D. Moner, M. Menárguez-Tortosa, D. Boscá, J. A. Miñarro Giménez, J. T. Fernández-Breis, and M. Robles, "Using the ResearchEHR platform to facilitate the practical application of the EHR standards," *J Biomed Inform*, Nov. 2012.

[31] U.S. National Institutes of Health (NIH), "ClinicalTrials.gov." http://clinicaltrials.gov/ct2/home. Last accessed: August 2012.

[32] Hospital Clínic, Barcelona, "Colorectal Cancer Screening in Average-risk Population: Immunochemical Fecal Occult Blood Testing Ver-

sus Colonoscopy." http://clinicaltrials.gov/ct2/show/NCT00906997. Last accessed: August 2012.

[33] C. G. Chute, "Medical Concept Representation," in *Medical Informatics. Knowledge Management and Data Mining in Biomedicine* (R. Sharda, S. Voß, H. Chen, S. S. Fuller, C. Friedman, and W. Hersh, eds.), vol. 8 of *Integrated Series in Information Systems*, ch. 6, pp. 163–182, Boston: Springer US, 2005.

[34] M. Peleg, S. Keren, and Y. Denekamp, "Mapping computerized clinical guidelines to electronic medical records: Knowledge-data ontological mapper (KDOM)," *J Biomed Inform*, vol. 41, no. 1, pp. 180–201, 2008.

[35] U.S. National Library of Medicine, "Unified Medical Language System (UMLS)." http://www.nlm.nih.gov/research/umls/. Last accessed: July 2012.

[36] International Health Terminology Standards Development Organisation (IHTSDO), "SNOMED CT." http://www.ihtsdo.org/snomed-ct/. Last accessed: July 2012.

[37] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation," *J Chronic Dis*, vol. 40, no. 5, pp. 373–383, 1987.

[38] World Wide Web Consortium (W3C), "OWL 2 Web Ontology

Language Document Overview." http://www.w3.org/TR/owl2-overview/, Oct. 2009. Last accessed: July 2012.

[39] F. J. Farfán Sedano, M. Terrón Cuadrado, Y. Castellanos Clemente, P. Serrano Balazote, D. Moner Cano, and M. Robles Viejo, "Patient Summary and medicines reconciliation: application of the ISO/CEN EN 13606 standard in clinical practice," *Stud Health Technol Inform*, vol. 166, pp. 189–196, 2011.

[40] Health Level Seven International, "HL7 Decision Support Service (DSS), Release 1." In http://www.hl7.org/implement/standards/product_brief.cfm?product_id=12, Aug. 2011. Last accessed: April 2013.

[41] E. German, A. Leibowitz, and Y. Shahar, "An architecture for linking medical decision-support applications to clinical databases and its evaluation," *J Biomed Inform*, vol. 42, no. 2, pp. 203–218, 2009.

[42] D. W. Lonsdale, C. Tustison, C. G. Parker, and D. W. Embley, "Assessing clinical trial eligibility with logic expression queries," *Data Knowl Eng*, vol. 66, pp. 3–17, July 2008.

[43] I. Cho, J. Kim, J. H. Kim, H. Y. Kim, and Y. Kim, "Design and implementation of a standards-based interoperable clinical decision support architecture in the context of the korean ehr," *Int J Med Inform*, vol. 79, pp. 611–622, Sept. 2010.

[44] S. Kim, B. Shim, J. A. Kim, and I. Cho, "SW Architecture for Access to Medical Information for Knowledge Execution," in *Security-Enriched Urban Computing and Smart Grid* (T.-h. Kim, A. Stoica, and R.-S. Chang, eds.), vol. 78 of *Communications in Computer and Information Science*, pp. 574–580, Springer Berlin Heidelberg, 2010.

[45] I. Bilykh, J. H. Jahnke, G. McCallum, and M. Price, "Using the Clinical Document Architecture as Open Data Exchange Format for Interfacing EMRs with Clinical Decision Support Systems," in *Proc. of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS 2006)*, (Washington, DC, USA), pp. 855–860, IEEE Computer Society, 2006.

[46] Health Level Seven International, "HL7 Version 3 Domain Analysis Model: Virtual Medical Record for Clinical Decision Support (vMR-CDS), Release 1." In `http://www.hl7.org/implement/standards/product_brief.cfm?product_id=270`, Apr. 2012. Last accessed: April 2013.

[47] A. L. Rector, "The Interface between Information, Terminology, and Inference Models," *Stud Health Technol Inform*, vol. 84, pp. 246–250, 2001.

[48] C. Parker, R. Rocha, J. Campbell, S. Tu, and S. Huff, "Detailed clinical models for sharable, executable guidelines," *Stud Health Technol Inform*, vol. 107, pp. 145–148, 2004.

[49] S. W. Tu, J. R. Campbell, J. Glasgow, M. A. Nyman, R. McClure, J. Mc-

Clay, C. Parker, K. M. Hrabak, D. Berg, T. Weida, J. G. Mansfield, M. A. Musen, and R. M. Abarbanel, "The SAGE Guideline Model: achievements and overview," *J Am Med Inform Assoc*, vol. 14, no. 5, pp. 589–598, 2007.

[50] W. Goossen, A. Goossen-Baremans, and M. van der Zel, "Detailed clinical models: a review," *Healthc Inform Res*, vol. 16, no. 4, pp. 201–214, 2010.