Semantic Integration of Patient Data and Quality Indicators based on *open*EHR Archetypes

Kathrin Dentler*[a,b], Annette ten Teije[a], Ronald Cornet[b], and Nicolette de Keizer[b]

a) Dept. of Computer Science, VU University Amsterdam, The Netherlands
b) Dept. of Medical Informatics, Academic Medical Center, University of Amsterdam, The Netherlands

Abstract. Electronic Health Record (EHRs) contain a wealth of information, but accessing and (re)using it is often difficult. Archetypes have been shown to facilitate the (re)use of EHR data, and may be useful with regard to clinical quality indicators. These indicators are often released centrally, but computed locally in several hospitals. They are typically expressed in natural language, which due to its inherent ambiguity does not guarantee comparable results. Thus, their information requirements should be formalised and expressed via standard terminologies such as SNOMED CT to represent concepts, and information models such as archetypes to represent their agreed-upon structure, and the relations between the concepts. The two-level methodology of the archetype paradigm allows domain experts to intuitively define indicators at the knowledge level, and the resulting queries are computable across institutions that employ the required archetypes. We tested whether *open*EHR archetypes can represent both elements of patient data required by indicators and EHR data for automated indicator computation. The relevant elements of the indicators and our hospital's database schema were mapped to (elements of) publicly available archetypes. The coverage of the public repository was high, and editing an archetype to fit our requirements was straightforward. Based on this mapping, a set of three indicators from the domain of gastrointestinal cancer surgery was formalised into archetyped SPARQL queries and run against archetyped patient data in OWL from our hospital's data warehouse to compute the indicators. The computed indicator results were comparable to centrally computed and publicly reported results, with differences likely to be due to differing indicator definitions and interpretations, insufficient data quality and insufficient and imprecise encoding. This paper shows that *open*EHR archetypes facilitate the semantic integration of quality indicators and routine patient data to automatically compute indicators.

Keywords: Semantic Integration, EHRs, Secondary Use of Clinical Data, Quality Indicators, openEHR Archetypes, OWL, SPARQL

1 Introduction

Today, increasing volumes of clinical data are being routinely recorded, and there is tremendous potential to benefit from reusing the resulting data sources both for individual patients and society in general. In fact, according to a recent report by PricewaterhouseCoopers, "using data for secondary purposes is one of the most promising ways

[*] corresponding author: k.dentler@vu.nl

to improve health outcomes and costs." [16]. Secondary purposes include research, the recruitment of eligible patients for clinical trials, the early detection of epidemics, reimbursement, clinical audit, the generation or testing of medical hypotheses and quality monitoring or reporting. Since patient data often resides in various heterogeneous systems, it needs to be integrated to be (re)usable. In addition, this patient data needs to be meaningful for applications that reuse it.

*Open*EHR archetypes [1] have been proposed to standardise clinical data to achieve semantic interoperability. They have been shown to facilitate the integration of data from several sources [15], to empower multi-centre clinical research [7] and to be a solid basis for ubiquitous computing [6]. Also, archetypes have been shown to facilitate the reuse of patient data for clinical trials [9] and guideline systems [3], [12]. In this paper, we focus on the reuse of patient data for the automated computation of quality indicators, which are measurable elements of practice performance for which there is evidence or consensus that they can assess the quality of provided care, and thus also change in quality [10]. Our main objective was to represent both patient data from our hospital's data warehouse and national quality indicators in terms of *open*EHR archetypes to automatically compute quality indicators.

To apply formal representation to ensure semantic interoperability and to be able to perform automated reasoning with the archetypes and the patient data, we employ an OWL 2[1] representation of archetypes, representing the patient data as its instances. To the best of our knowledge, this is the first time that real patient data is being represented based on *open*EHR archetypes in OWL and used to compute clinical quality indicators.

The structure of this paper is as follows: Section 2 introduces quality indicators and archetypes, and Section 3 our methods and materials. We report on our case study in Section 4. Finally, lessons learned and future challenges are discussed in Section 5. Section 6 concludes this paper.

2 Background: Quality Indicators and Archetypes

This section provides background information on quality indicators and archetypes.

Quality Indicators are employed internally by hospitals to measure and improve the quality of care and externally for accountability and hospital comparison. For the latter, it is essential that the same measurements are performed in each hospital. Quality indicators are often expressed as a fraction, where the denominator defines the criteria of patients to whom the indicator applies, and the numerator those criteria indicating whether the patients received high-quality care. Exclusion criteria can apply. These indicators can be computed automatically by running two queries against the required patient data: one for the denominator and another for the numerator. A sample indicator is the evidence-based process indicator "Number of examined lymph nodes after colon resection" as defined by the Dutch Healthcare Inspectorate[2] for the reporting year 2010:

[1] `http://www.w3.org/TR/owl2-overview/`

[2] `http://www.zichtbarezorg.nl/page/Ziekenhuizen-en-ZBC-s/Kwaliteitsindicatoren`

Numerator: Number of patients who had 10 or more lymph nodes examined after resection of a primary colon carcinoma.
Denominator: Number of patients who had lymph nodes examined after resection of a primary colon carcinoma.
Exclusion criteria: Previous radiotherapy and recurrent colon carcinomas

Archetypes are knowledge-level models that represent clinical concepts and define the structure to record, exchange and integrate clinical data. *Open*EHR archetypes are created based on the consensus of domain experts, and are available via the public archetype repository *Clinical Knowledge Manager*[3]. They define occurrence and cardinality constraints, as well as constraints on the values to be entered. The main categories are Action (e.g. Procedure undertaken), Evaluation (e.g. Diagnosis), Observation (e.g. Blood Pressure) and Instruction (e.g. Medication order). Figure 1 depicts the publicly available archetype “Tumour - lymph node metastases”[4]. The optional archetype node “Number of nodes examined” constrains the number of examined lymph nodes to be greater than or equal to 0.

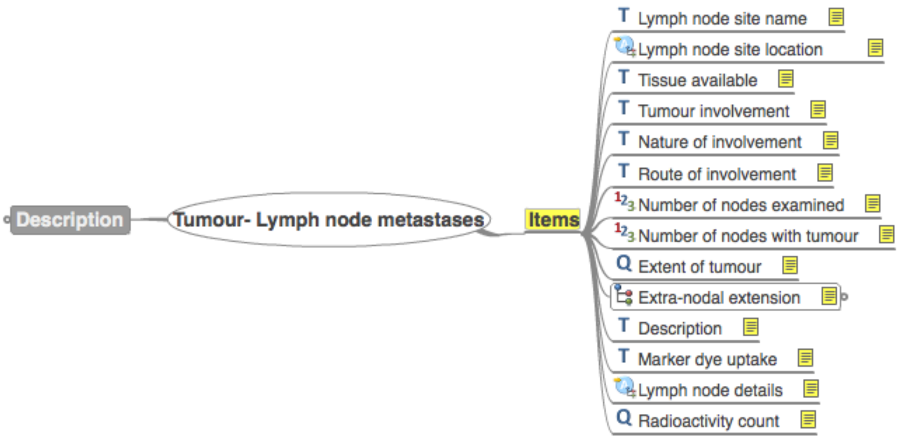

Fig. 1: Archetype “Tumour - lymph node metastases”. The icons depict the datatypes that are to be used. The *T* stands for free or coded text, *Q* for a quantity, *123* for a count, the globe depicts a slot (cluster) that can include other archetypes, such as “Precise anatomical location” for the node “Lymph node site location”, and the tree icon depicts a cluster.

According to the Semantic Health Report [18], semantically interoperable EHR systems rely upon three layers to represent meaning: standard generic reference models such as the *open*EHR Reference Model or the Health Level 7 Clinical Document Architecture (HL7 CDA), agreed clinical structure definitions such as *open*EHR archetypes or HL7 templates, and clinical terminology systems such as LOINC[5] or SNOMED CT

[3] http://www.openehr.org/knowledge
[4] http://openehr.org/knowledge/OKM.html#showarchetype_1013.1.396_5
[5] http://loinc.org/

[4]. Archetype-enabled EHR architectures are based on the two-level methodology [1], which separates the knowledge level from the information level. Archetypes on the knowledge-level constrain the standardised stable and generic reference model on the information level that consists of few abstract classes. The reference model can either be implemented directly by EHR systems or mapped to a local data structure. Unlike the reference model, archetypes evolve together with medical knowledge. Here, we use the term *information model* to refer to both archetypes and their underlying reference model.

The two-level methodology allows queries against patient data to be constructed at the knowledge level, enabling clinical domain experts to contribute to the formalisation of quality indicators without having to know the underlying structure of the patient data. It also makes the resulting queries computable across systems that employ the required archetypes. If data is stored in proprietary systems or represented in competing standards, the required elements have to be mapped from the locally employed information model to the (elements of the) archetypes used to identify required elements to compute the quality indicator.

3 Methods and Materials

This section describes the sample set of employed quality indicators, their formalisation, our patient data and how we related it to SNOMED CT codes, the translation of archetypes to OWL and how patient data was dealt with in OWL.

Quality Indicators and their Formalisation

Besides the sample indicator "Number of examined lymph nodes after resection" described above, the other two evidence-based indicators of the sample set are the process indicator "Patients with rectum carcinoma who have been discussed in a preoperative multidisciplinary meeting", and the outcome indicator "Unplanned re-interventions after resection of a primary colorectal carcinoma". We previously formalised the same sample set with our quality indicator formalisation method CLIF [5], employing a self-defined information model and self-generated patient data. CLIF consists of eight steps. In step 1), relevant concepts have to be identified in the indicator text and encoded in a terminology such as SNOMED CT. In step 2), the elements of the information model are defined and related to each other. In step 3) to 5), temporal, numeric and Boolean constraints are defined. In step 6, constraints can be grouped by Boolean connectors. Finally, exclusion criteria are defined in step 7), and the difference between denominator and numerator is made explicit in step 8). Of all steps, the second is the most relevant in the context of this paper, because we improve the formalisation by using public *open*EHR archetypes as information model. Besides, real patient data from our hospital's data warehouse is being used.

Patient Data and SNOMED CT Codes

We worked on a subset of our hospital's data warehouse, beginning from 2009.[6] The central patient table (1,672,104 entries) contains demographic information and patient IDs. Other relevant tables contain diagnoses (2,925,156), operations (144,860), admissions (259,005), encounters (3,244,586) and pathology reports (92,870). The diagnosis table from the data warehouse contains ICD-9-CM codes for ca. half of the diagnoses, which we mapped to the latest SNOMED CT release (January 2012) via the SNOMED CT to ICD-9-CM crossmap included in the release. The procedures in the operation table contain codes from the Dutch procedure classification of nearly 40,000 codes that are not mapped to any other terminology. Therefore, we manually mapped a relevant subset that refers to "colorectal" procedures to SNOMED CT.

The sample set of quality indicators is computed centrally by the Dutch Surgical Colorectal Audit based on data submitted by Dutch hospitals for all operations on patients with a primary colorectal carcinoma. To extract a manageable but relevant set of patient data, we matched the data submitted by our hospital to the DSCA from 2009 to 2011 with the data stored in our data warehouse. In absence of a mapping between the patients in both systems, we searched DSCA patients in our data warehouse based on sex, year of birth, operation and discharge date as well as the procedures that they underwent. This strategy allowed us to match 192 of the 229 patients for whom data has been submitted to the DSCA.

In total, for the 192 patients 2,656 diagnoses have been recorded, of which 1,515 have (271 distinct) ICD-9 codes, the others are not encoded in ICD-9. 1,325 (239 distinct) of these codes are present in the SNOMED CT to ICD-9 crossmap, and related to 17,611 (3,878 distinct) SNOMED CT codes. 724 (201 distinct) procedures have been recorded, of which 287 (32 distinct) are present in our manually created mapping table, and related to 949 (50 distinct) SNOMED CT procedure codes. This results in 191 of the 192 patients being related to SNOMED CT procedure codes, and 190 patients to diagnosis codes. Data required that is recorded in our hospital but not contained in the data warehouse is information on radiotherapy and multi-disciplinary meetings. However, it is present in the DSCA dataset and thus we retrieve it from there. We also retrieved the number of examined lymph nodes from the DSCA dataset, which is present in our data warehouse, but only in Dutch free text.

Archetypes in OWL

We reused the Archetype Ontologizer[7] [11] to create the OWL 2 ontologies for the archetypes required to represent the patient data and the quality indicators. The translated ontologies are based on the "*open*EHR Specific Data Structures and Data Types" ontology[8] [13] that represents the *open*EHR reference model, containing its data structures and data types with all their properties. We made minor adaptations to the translator so that the default namespace includes the ID of the respective archetype, and

[6] In the Netherlands, there is no need for patient consent when, as in our study, individual patients are not directly involved. The use of the data is officially registered according to the Dutch Personal Data Protection Act.

[7] `http://oe.dynalias.net:8080/JSPWebArchetypeOntologizer/`

[8] `http://klt.inf.um.es/~cati/ontologies/OpenEHR-SP-v2.0.owl`

added the internal node IDs to class names for nodes. Furthermore, we made use of the OWL 2 reasoners HermiT and Pellet to check the consistency of the ontologies and the satisfiability of all classes. We used Pellet's explanation feature to identify the causes for unsatisfiable classes and improved the translator until all classes were satisfiable (for example, a datatype used in combination with a property from the "*open*EHR Specific Data Structures and Data Types" ontology had to be changed from integer to float to conform to the properties range). With this adapted translator, we translated the 5 archetypes needed to represent the patient data and the quality indicators from ADL, the Archetype Definition Language, to OWL. We then merged the resulting OWL 2 ontologies with the "*open*EHR Specific Data Structures and Data Types" ontology. The final ontology consists of 2,001 logical axioms, and has the expressivity $\mathcal{ALCHIQ}(\mathcal{D})$.

Patient Data in OWL
The patient data was originally stored in a MySQL database, and transformed into OWL using the OWL API. To run the queries against the patient data, we loaded the full closure of SNOMED CT (January 2012), the merged archetype ontology and the transformed patient data into OWLIM-SE 5.0 [2], and ran it in combination with Sesame 2.6.5[9], because it supports SPARQL 1.1[10].

4 Case Study

To establish whether open*EHR archetypes are suitable to semantically integrate routine clinical data and quality indicators*, we first transformed patient data from our data warehouse into archetyped patient data (Section 4.1) and modelled the concepts of our sample set of quality indicators in terms of *open*EHR archetypes (Section 4.2). We then constructed archetyped SPARQL queries (Section 4.3) and ran them against the archetyped patient data to compute the indicators (Section 4.4).

4.1 Transforming Patient Data into Archetyped Patient Data

The first step of the transformation process is to map the data structure of our data warehouse and the DSCA dataset to *open*EHR archetypes. We make use of archetypes from the Clinical Knowledge Manager[11], as it can be assumed that publicly available archetypes are most widely employed.

Table 1 provides an overview of the mapping. Most database tables and their relevant columns can be mapped directly to (elements of) archetypes. The patient table is mapped to the demographic archetype "Patient", and the patient ID to its mandatory node "Name"; SNOMED CT diagnosis codes are mapped to the node "Diagnosis" of the archetype "Diagnosis", and operation codes to the node "Procedure" of the archetype "Procedure undertaken". For radiotherapy, multidisciplinary meeting and pathology, exact codes are neither available nor required, as they are not specified by the

[9] `http://www.openrdf.org/`

[10] `http://www.w3.org/TR/sparql11-query/`

[11] `http://www.openehr.org/knowledge`

Table 1: Mapping between the local data structure and *open*EHR archetypes. The added element is italicised. Database tables have been mapped to archetypes, and database columns to nodes of archetypes. Data warehouse is abbreviated by DWH, and SNOMED CT by SCT.

Table	**Column**	**Archetype**	**Node**
Patient (DWH)		Patient	
	Identifier (DWH)		Name
Admission (DWH)		Patient Admission	
	Admission Date (DWH)		Admission Date
	Discharge Date (DWH)		*Discharge Date (added)*
Diagnosis (DWH)		Diagnosis	
	ICD-9 (DWH) (SCT code via ICD-9 - SCT mapping)		Diagnosis
Operation (DWH)		Procedure undertaken	
	Dutch procedure code (DWH, SCT code via manual mapping)		Procedure
(DSCA)	Radiotherapy	Procedure undertaken	Procedure with fixed SCT code (SCT_108290001)
(DSCA)	Multidisciplinary meeting	Procedure undertaken	Procedure with fixed SCT code (SCT_312384001)
Pathology (DWH, only lymph node examination)		Procedure undertaken	Procedure with fixed SCT code (SCT_284427004)
	Number of examined lymph nodes (DSCA)	Tumour- Lymph node metastases	Number of nodes examined

indicators, so fixed codes were set. To represent the number of examined lymph nodes, we employ the archetype “Tumour - lymph node metastases”, depicted in Figure 1, to record findings of lymph node metastases. While admissions and admission dates can be mapped directly, the admission archetype does not contain the required patient’s discharge date, and at the time of writing, an archetype “Patient discharge” did not exist either. Consequently, we added the node “Discharge date/time” to the archetype “Patient admission”. All procedure dates are represented via *open*EHR’s reference model.

Based on the mapping, the patient data was transformed into OWL individuals of the archetype classes. Our program transforms every patient into an OWL individual of the archetype “Patient”, with an arbitrary patient number represented in the obligatory archetype node “Name”. Then, all SNOMED CT diagnoses and procedures with their corresponding dates, and all admissions are transformed into OWL 2 individuals. The number of examined lymph nodes is added, and the date of the first pathology report after the operation is set as lymph node examination date. Finally, data from the DSCA database table related to radiotherapy and multi-disciplinary meetings is added. The resulting dataset contains 52,495 logical axioms, and its expressivity is $AL(D)$.

Let us consider the data for an example candidate patient for the lymph node indicator as depicted in Figure 2. The patient has an instance of a “Diagnosis” for the diagnosis primary colon carcinoma, an instance of a “Procedure” for a colectomy and one for lymph node examination, and an additional instance of “Tumour - lymph node metastases”. The diagnosis and procedures are related to their respective SNOMED CT codes via the property “value_element”. Relationships between a patient and other individuals are expressed by the “links” property.

4.2 Modelling Quality Indicators in terms of *open*EHR Archetypes

This section discusses the archetype-level modelling of quality indicators in terms of *open*EHR archetypes as employed information model.

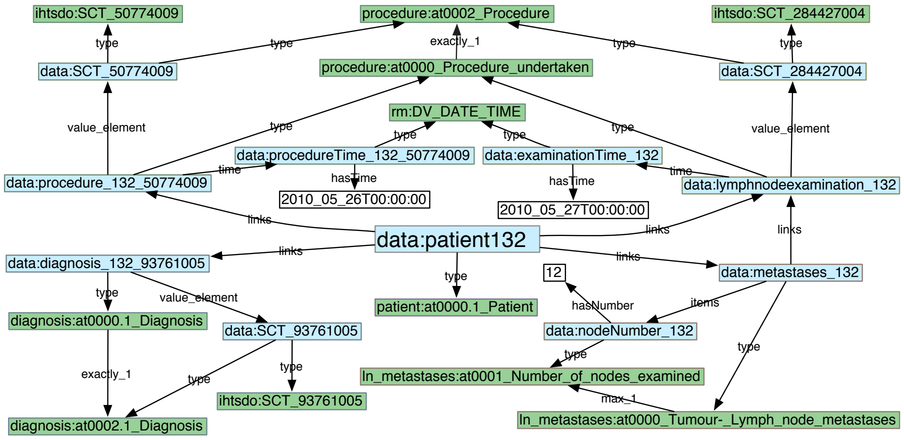

Fig. 2: Example Patient. Green elements are the classes that stem from the OWL archetypes. Blue elements are instances of these classes, and white elements are literals. Note that not all relations defined between nodes of the archetypes are depicted. For example, “diagnosis:at0002.1_Diagnosis” is a node of the archetype “diagnosis:at0000.1_Diagnosis”.

Bind concepts from a terminology to concepts of an information model For each SNOMED CT code that defines a diagnosis or procedure occurring in the indicator texts (as identified in step 1 of CLIF [5]), a corresponding archetype node has to be identified, and the code and the node have to be related to each other. This mapping is straightforward: all diagnosis codes are mapped to the node “Diagnosis” of the archetype “Diagnosis”, and all procedure codes to the node “Procedure” of the archetype “Procedure undertaken”.

Defining relations between assigned concepts of the information model Subsequently, relations between the assigned concepts of the information model have to be defined, i.e. relations that concern the instances to be queried. As intra-archetype relations are part of the archetype definition, only inter-archetype relations need to be defined. According to the problem-oriented patient model paradigm, all procedures should be related to the diagnosis that they are associated with, and this should be feasible via the node “Reason/s for procedure” of the archetype “Procedure undertaken”. Unfortunately, such relations are not present in our data warehouse, so that performing this substep was not possible. We related patients to their diagnoses and procedures, and the “number of examined lymph nodes” to the “lymph node examination” via the property “links” that stems from *open*EHR’s reference model.

4.3 Constructing Archetyped SPARQL Queries

The SPARQL queries were constructed based on the mapping between relevant elements occurring in the quality indicators and their corresponding (elements of) *open*EHR archetypes. Defining quality indicators with the help of archetype elements makes them, in principle, computable across systems that make use of the required archetypes to store clinical data. We defined the graph patterns to be matched based on the translated OWL classes and properties and their inter- and intra-archetype relations. Patients with SNOMED CT classes or subclasses identified in the indicator were retrieved with the help of the SNOMED CT closure. For brevity[12], the following query-extract shows only a query for archetyped patients with diagnoses of the SNOMED CT concept 93761005, i.e. "Primary malignant neoplasm of colon (disorder)":

```
PREFIX patient: <http://few.vu.nl/˜kdr250/archetypes/openEHR-DEMOGRAPHIC-PERSON.person-patient.v1.owl#>
PREFIX diagnosis: <http://few.vu.nl/˜kdr250/archetypes/openEHR-EHR-EVALUATION.problem-diagnosis.v1.owl#>
PREFIX schemarm: <http://klt.inf.um.es/˜cati/ontologies/OpenEHR-SP-v2.0.owl#>
PREFIX sct: <http://www.ihtsdo.org/>
SELECT DISTINCT ?patient WHERE {
    ?patient a patient:at0000.1_Patient .
    ?patient schemarm:links ?diagnosis .
    ?diagnosis a diagnosis:at0000.1_Diagnosis .
    ?diagnosis schemarm:value_element ?diagnosiscode .
    ?diagnosiscode a diagnosis:at0002.1_Diagnosis .
    ?diagnosiscode a sct:SCT_93761005 .
} ORDER BY ?patient
```

4.4 Calculating the Indicators by Running the Queries

Table 2 compares our computed indicator results to the results contained in the report generated for our hospital by the DSCA, and the results publicly reported[13].

Table 2: Comparison of our results to those reported by the DSCA and publicly reported results. Note that some of the indicator definitions and interpretations differ: For example, the reoperation indicator publicly reported includes all colorectal operations, and not only those due to a colorectal carcinoma. Also, it defines re-operations as having taken place within 30 days after the operation, while our indicator - as specified in the indicator description - in addition includes re-operations during the same admission.

Indicator / Results	**Our Result**	**DSCA**	**Publicly Reported**
Lymph nodes	85,71% (42/49)	80,00% (43/54)	-
Meeting	91,66% (22/24)	100% (21/21)	-
Re-operation	1,66% (1/60)	9% (7/75)	8,33% (20/240)

The results reported here are a first approximation, and a thorough analysis is required to determine their reliability, validity and all causes for differing results. As a first evaluation, we analysed the results for the denominator of the indicator "Patients with rectum carcinoma who have been discussed in a preoperative multidisciplinary meeting", which retrieves patients with rectum carincoma who have been operated in the reporting year. The query on the DSCA dataset retrieves 21 patients, whereas the query on the data warehouse retrieves 24. Three out of the 21 patients retrieved on the DSCA dataset were not mapped to patients of our data warehouse. Thus, the query on

[12] Translated archetypes, extract of synthetic patient data and constructed queries: http://www.few.vu.nl/~kdr250/archetypes/

[13] http://www.ziekenhuizentransparant.nl/

the data warehouse retrieved 6 patients who were not retrieved by the other query. All of these patients are registered with a carcinoma located in the *Colon sigmoideum* in the DSCA dataset. In the data warehouse, this is represented with the ICD-9 code 154.00 (Malignant neoplasm of rectosigmoid junction) in all cases. Via the ICD-9 to SNOMED crossmap, this code is mapped to 4 different SNOMED CT concepts, some of which are subconcepts of "Primary malignant neoplasm of rectum (disorder)", which is employed in the indicator query. Thus, these patients are retrieved as rectum carcinoma patients, whereas they have been classified as colon carcinoma patients by the surgeon who entered the data.

5 Discussion

This section discusses the most notable lessons learned during our case study.

Differing Indicator Results and Encoded Data. Besides from differing indicator definitions and interpretations, differing indicator results are likely to be caused by missing patients, who could be not be mapped from the DSCA dataset to our data warehouse based on their properties. The fact that not all patients could be mapped indicates insufficient data quality. Another cause might be insufficient encoding: only a little more than half of the diagnoses in our data warehouse are encoded in ICD-9. Also, no mapping from the Dutch procedure classification to SNOMED CT exists. We detected that patients who have been classified to have a carcinoma in the colon sigmoideum by our surgeons are retrieved as rectum carcinoma patients. This might be due to an incorrect ICD-9 code in the data warehouse. If those patients are indeed sigmoid colon carcinoma patients, the ICD-9 code 153.3 (Malignant neoplasm of sigmoid colon) would have been preferable. The general question remains whether ICD-9 is suitable for our use case, as it is not intended to support the secondary use of clinical data. Routine data must be of sufficient quality, structured, complete, and encoded in detailed, correct concepts from standard terminologies to be (re)usable. Clinical quality indicators must be well-formalised so that comparable results can be obtained. In future, we will investigate the effect of data quality on the reliability and validity of obtained indicator results.

Coverage of the *open*EHR archetype repository. Modelling both the patient data and the quality indicators at the archetype level was intuitive. With regard to the coverage of the *open*EHR archetype repository, we were able to map nearly all required elements to (elements of) archetypes. A missing element was "discharge date", which we expected to be present as a node "discharge date" in the "admission" archetype, or as a separate archetype. Editing the "admission" archetype to fit our requirements was easy, and it would have been possible to contribute our addition to the public repository. In total, we made use of 6 nodes from 5 archetypes.

Archetypes in OWL and Properties. The Archetype Ontologizer proved to be useful after some minor adaptions, and working with the OWL representation of archetypes was practical due to the wide range of Semantic Web tools available.

Regarding the archetyped patient data, all employed properties stem from the reference model, except from *hasTime*, *hasNumber* and *hasBoolean*. In OWL, XML Schema datatypes are used in typed literal values, while the reference model defines datatypes such as DV_DATE_TIME. As literals can not be instances of classes by definition, the relationship between the literals and the classes can not be expressed directly. Defining properties between OWL classes of the archetypes and their instances was complex, as it was unclear which properties would be the correct ones to use for inter-archetype relationships. We chose the property "links" from the reference model to relate patients to their diagnoses and procedures. The use of more meaningful alternatives will be explored in future work. In our data warehouse, procedures are not related to the diagnoses due to which they have been carried out. This forces us to employ heuristics (e.g., a procedure is typically being carried out after the corresponding diagnosis has been recorded), which might negatively impact the validity of indicator results.

Automated Reasoning with Patient Data and Information Models in OWL: Past and Future Many researchers have demonstrated the added value of patient information models represented in OWL. Lezcano et al. [11] integrated archetypes in OWL 2 with SWRL (Semantic Web Rule Language) rules, which are then to be applied to instances of clinical data. Rector et al. [17] represented a set of information models and bindings to a coding system (i.e. allowed codes) in OWL and validated it with a reasoner. They also validated whether individual data structures conform to the information model with the help of added closure axioms. In a comparable study, the *open*EHR library of archetypes was translated into OWL classes and subsequently validated with OWL reasoners [14]. Heymans et al. [8] formalised a subset of the constraints in the implementation guide on *Using SNOMED CT in HL7 Version 3* as OWL Integrity Constraints and automatically validated CDA documents using the OWL 2 reasoner Pellet.

The OWL representation of archetypes and patient data opens new opportunities for automated reasoning: First, reasoning may be useful at a patient data level. The massive data sources currently locked up in EHRs contain a wealth of implicit knowledge that could be made explicit by formal reasoning. In addition, the OWL representation of archetypes could be used to validate whether the patient data fulfils the constraints defined in the corresponding archetypes. For example, it could be checked whether the number of examined lymph nodes is indeed greater than or equal to 0. Finally, it may be possible to infer archetype class memberships for patient data. Reasoning is also required at the archetype-level: It is unrealistic to expect publicly available archetypes to be expressive enough to cover all possible clinical concepts required for all kinds of use cases. Thus, users of the two-level methodology define their own archetypes, and it is important to be able to infer subsumption and equivalence relationships between self-defined and publicly available archetypes. Finally, as information models and terminologies are developed independently from each other, they may overlap, and different systems and users will make different modelling choices. It must be possible to detect semantically equivalent constructs.

6 Conclusion

Our research question for this paper was whether open*EHR archetypes are suitable to semantically integrate patient data and quality indicators*, with the goal to reuse routine patient data for secondary purposes such as the computation of indicators. Mapping both our local database schema and elements of patient data occurring in indicators to (elements of) archetypes was intuitive. This can be attributed both to the two-level methodology, which also makes the resulting queries computable across institutions employing the required archetypes, and the high coverage of *open*EHR's public Clinical Knowledge Manager. We edited an existing archetype to fit our requirements. Based on our mappings, we archetyped the patient data and formalised our sample set of indicators as SPARQL queries with our indicator formalisation method CLIF. We ran the resulting queries against the archetyped patient data to prove the concept. Since *open*EHR archetypes are applicable to represent both patient data and elements of patient data required to compute clinical quality indicators, we conclude that they are suitable for semantic integration of patient data and quality indicators. Further research is required into the potential benefit of automated reasoning based on the OWL representation of archetyped patient data.

References

1. T. Beale. Archetypes: Constraint-based domain models for future-proof information systems. In *OOPSLA 2002 workshop on behavioural semantics*, pages 1–18, 2002.
2. B. Bishop, A. Kiryakov, D. Ognyanoff, I. Peikov, Z. Tashev, and R. Velkov. OWLIM: A family of scalable semantic repositories. *Semantic Web*, 2(1):33–42, 2011.
3. R. Chen and P. Georgii-Hemming. Representing a chemotherapy guideline using openEHR and rules. *Stud Health Technol Inform*, pages 653–657, 2009.
4. R. Cornet and N. de Keizer. Forty years of SNOMED: a literature review. *BMC medical informatics and decision making*, 8 Suppl 1:S2, Jan. 2008.
5. K. Dentler, A. ten Teije, R. Cornet, N. de Keizer, and N. De. Towards the Automated Calculation of Clinical Quality Indicators. In *Knowledge Representation for Health-Care*, volume LNCS 6924, pages 51–64, 2012.
6. S. Garde, E. Hovenga, J. Buck, and P. Knaup. Expressing clinical data sets with openEHR archetypes: a solid basis for ubiquitous computing. *International Journal of Medical Informatics*, 76 Suppl 3:S334–41, Dec. 2007.
7. S. Garde, P. Knaup, T. Schuler, and E. Hovenga. Can openEHR Archetypes Empower Multi-Centre Clinical Research? *Studies in health technology and informatics*, 116:971–6, Jan. 2005.
8. S. Heymans, M. McKennirey, and J. Phillips. Semantic validation of the use of SNOMED CT in HL7 clinical documents. *Journal of biomedical semantics*, 2(1):2, July 2011.
9. C. Kohl, S. Garde, and P. Knaup. Facilitating secondary use of medical data by using openEHR archetypes. *Studies in health technology and informatics*, 160(Pt 2):1117, 2010.
10. M. Lawrence and F. Olesen. Indicators of Quality in Health Care. *European Journal of General Practice*, 3(3):103–108, Jan. 1997.
11. L. Lezcano, M.-A. Sicilia, and C. Rodríguez-Solano. Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules. *Journal of Biomedical Informatics*, 44(2):1–11, Nov. 2010.
12. M. Marcos, J. 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. *Artificial Intelligence in Medicine*, pages 276–285, 2011.

13. C. Martínez-Costa, M. Menárguez-Tortosa, J. T. Fernández-Breis, and J. A. Maldonado. A model-driven approach for representing clinical archetypes for Semantic Web environments. *Journal of biomedical informatics*, 42(1):150–64, Feb. 2009.
14. M. Menárguez-Tortosa and J. Fernández-Breis. Validation of the openEHR Archetype Library by using OWL Reasoning. *Studies in health technology and informatics*, 169:789, Jan. 2011.
15. D. Moner, J. Maldonado, D. Bosca, J. T. Fernandez, C. Angulo, P. Crespo, P. J. Vivancos, and M. Robles. Archetype-based semantic integration and standardization of clinical data. *Proceedings of the 28th IEEE EMBS Annual International Conference*, 1:5141–4, Jan. 2006.
16. PricewaterhouseCoopers. Transforming healthcare through secondary use of health data. 2009.
17. A. L. Rector, R. Qamar, and T. Marley. Binding ontologies and coding systems to electronic health records and messages. 4:51–69, 2009.
18. V. Stroetmann, D. Kalra, P. Lewalle, A. Rector, J. Rodrigues, K. Stroetmann, G. Surjan, B. Ustun, M. Virtanen, and P. Zanstra. Semantic Interoperability for Better Health and Safer Health Care. *European Commission*, (January), 2009.