Storing and Querying Longitudinal Data Sets in an Open Source EHR

John CHELSOM[a,1] and Naveed DOGAR[a, b]
[a] *Seven Informatics Ltd, Oxford, UK.*
[b] *University of Oxford, UK.*

Abstract. The cityEHR is an example of an open source EHR system which stores clinical data as collections of XML documents. The records gathered in routine clinical care are a rich source of longitudinal data for use in clinical studies. We describe how the standard language XQuery can be used to identify cohorts of patients, matching specified criteria. We discuss methods for ensuring good data quality and the issues in implementing XML queries on longitudinal data sets.

Keywords. EHR, Longitudinal Data, HL7 CDA, XQuery, Cohort Studies

1. Introduction

The Electronic Health Record (EHR), stored as a set of clinical documents, can provide a rich source of longitudinal data that can be used in secondary clinical studies. Longitudinal data are measurements or observations that are repeated for each individual patient over a period of time.

This paper describes the storage of clinical documents in the cityEHR open source health records system [1] and the implementation of queries to retrieve cohorts of patients matching specified constraints; its aim is to explore the main advantages and limitations of the pure XML architecture and implementation of cityEHR.

The cityEHR is an example of the generic data modeling approach to EHR described by Johnson [2]. A base ontology, expressed in the Web Ontology Language, OWL [3], represents concepts from the ISO 13606 [4] and HL7 CDA [5] standards. The assertions in this ontology are used to build templates for CDA documents which are then instantiated with data for individual patients.

HL7 CDA is an open, XML-based standard for representing clinical data in the form of documents and is widely used as a means for storing and exchanging clinical documents. The structure of an HL7 CDA document is itself an instantiation of the HL7 Reference Information Model and is also closely aligned to the overall structure of an electronic health record defined in ISO 13606.

The modeling approach in cityEHR is similar to the use of archetypes in ISO 13606 and openEHR [6], but whereas these use a specialised Archetype Definition Language (ADL), cityEHR uses the open standards XML, XPath and OWL. An advantage of the XML-based approach in cityEHR is that the information architecture, models and tooling are compatible with any other XML-based systems. Other studies

[1] Corresponding Author.

have also explored the use of archetypes to generate HL7 CDA templates, for example Moner, et al [8].

The majority of openEHR systems are implemented using relational database technology [10, 11] but cityEHR is implemented using a native XML database, storing standard HL7 CDA clinical documents directly as XML. With this pure XML approach, cohorts of patients are identified by running queries on the XML document store using the standard XQuery language [7]. We will show the implementation of this approach in cityEHR and explore the limits to which longitudinal data sets can be interrogated, beyond which data must be extracted from the EHR and analysed using more sophisticated statistical techniques (see for example, Liang and Zeger [9]).

2. EHR as a Source of Longitudinal Data

2.1. Record Storage

The cityEHR is an Electronic Health Record implemented as a records management system. In contrast to a relational database implementation, where the data are normalised before persistent storage and retrieval, the store of HL7 CDA documents may include redundant data. The smallest unit of meaningful clinical information in HL7 CDA is the Clinical Statement, consisting of an Entry containing one or more Values (Elements in the ISO 13606 model). Entries are recorded in Clinical Documents (Compositions in ISO 13606) – typically forms, letters or messages in the EHR.

Longitudinal data sets can be built up in several ways. A simple example is that of laboratory test data, which are received as HL7 messages from laboratory information systems and may be repeated many times over the lifetime of a patient. Figure 1 shows another example – the smoking habits of a patient, recorded in a lifestyle questionnaire during consultation with a general practitioner. After the first such consultation, the questionnaire is pre-filled with the values recorded previously, so that the observation of smoking habits is reaffirmed on each consultation, or recorded as having changed.

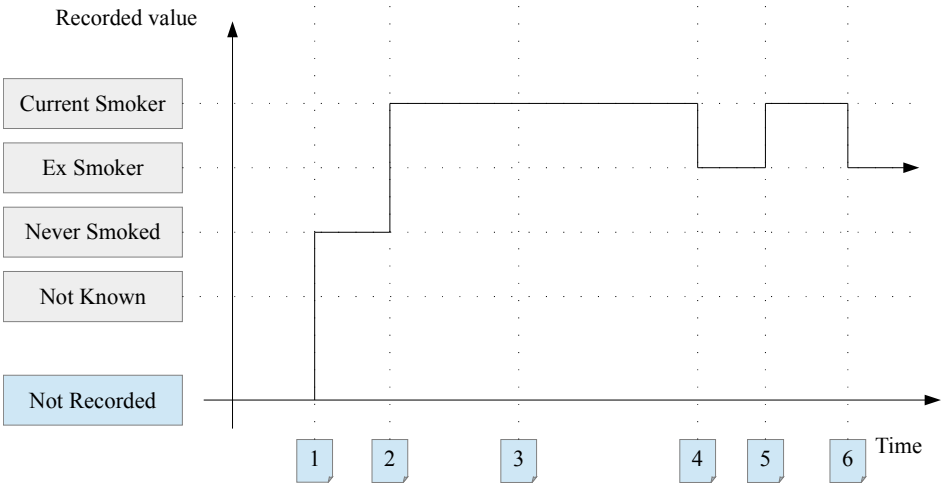

Figure 1. Smoking Habits Recorded at Six Points in Time

2.2. Accurate Data for Cohort Studies

The records management in an EHR for routine clinical care must implement a number of essential features if it is to be used as an accurate source of longitudinal data for clinical studies.

Whilst for routine care it is important that the EHR is designed to be immutable (i.e. data, once recorded, cannot be changed) for cohort studies it is necessary to exclude data that were found to be recorded in error. This may apply both to data within an individual document or for the entire record for a patient.

Where patient data are recorded in error, the entire clinical document must be superseded by a new version, with the erroneous data corrected. The original document is not deleted, since it forms an important part of the historic record for the patient; clinical decisions may have been made on the basis of the erroneous data and these may need to be justified using the historic record. Moreover, errors may have been identified based on an individual clinical judgment which itself is later deemed to have been an error, in which case the original document is restored. When querying the data for secondary uses, it is important that the superseded documents have been moved to an archive which is not included in the query; for routine clinical use, the links between the archived (original) documents and the superseding documents must be maintained so that a complete historic record is maintained.

Similarly, there may be cases where a patient has duplicate records in the EHR. It must be possible to merge any such records, so that the number of patients in cohorts formed by searches is not misreported. There may also be cases where entire records are stored in the EHR in error (for example, fictitious records used for testing) and these too must be moved to an archive, outside the scope of the searched record set.

3. Retrieval of Data For Cohort Studies

3.1. Identifying Cohorts and Assembling Data Sets

Since cityEHR stores patient data as XML documents, queries to retrieve data can be expressed in the standard XQuery language [7]. There are two stages to data retrieval for cohort studies: identification of patients in the cohort, followed by export of longitudinal data sets for each patient in the cohort. In this paper, space allows us only to focus on the first of these: the formulation of queries to define cohorts of patients with data matching the query constraints.

The web-based user interface in cityEHR provides a user-friendly environment for formulating an XQuery to find patients matching constraints on a Clinical Statement. Figure 2 shows the XQuery which finds a cohort of patients assessed within a particular date range. The XQuery returns a set of <patient> elements, with the patient id, the effective time of the matched observation and the value that satisfied the constraints.

The clinical context for a query is either at the level of the Clinical Statement (i.e. matching constraints on a set of Elements within a specified Entry) or the Clinical Document (i.e. matching constraints on a set of Entries recorded in the same Composition). The cohorts returned from each individual query can be combined using logical operators (AND, OR, NOT) to build up complex constraints on the final cohort.

```
for $value in /descendant::cda:observation[cda:id/@extension eq '#ISO-13606:Entry:assessdate']
                /descendant::cda:value[@extension eq '#ISO-13606:Element:Date']
                [@value ge '2016-01-01'][@value le '2016-07-31']
let $document := $value/ancestor::cda:ClinicalDocument
let $effectiveTime := $document/cda:effectiveTime/@value
return
<patient id="{$document/descendant::cda:patientRole/cda:id/@extension}"
          effectiveTime="{$effectiveTime}" value="{$value/@value}"/>
```

```
<exist:result exist:hits="3" exist:start="1" exist:count="3"  xmlns:exist="http://exist.org">
   <patient id="47474774" effectiveTime="2016-01-15T12:01:18" value="2016-01-13"/>
   <patient id="12341234" effectiveTime="2016-04-15T09:26:10" value="2016-04-03"/>
   <patient id="47474774" effectiveTime="2016-05-22T11:53:36" value="2016-05-21"/>
</exist:result>
```

Figure 2. Simple XQuery and Results Document.

3.2. Effective Time of Matched Observations

Returning to the example of Smoking Habits shown in Figure 1, a search for patients recorded as Current Smoker would return multiple hits for the patient shown; the query is for patients who have *ever* been recorded as a Current Smoker, and for this patient that has happened on more than one occasion (at times 2, 3 and 5). However, the most recent observation is that the patient is an Ex Smoker. To support constraints on the effective time, a second XQuery can be used to return the set of all recorded instances of Smoking Habits. Post processing of the two results sets can then compare the effective times to match the first recorded value, the last recorded (i.e. current) value or even the n^{th} recorded value.

3.3. Expanding the Clinical Context

Suppose patients are prescribed a drug during an initial assessment and are then followed up at 1 month, 4 months and 12 months to check compliance. The monitoring form filled out at each stage is the same, resulting in three Compositions stored in the patient record, each with the same set of Entries.

A query to return the cohort of patients observed as non-compliant is easily formulated using the method shown above, but the query to return the cohort of patients observed as non-compliant on the 12 month follow-up requires the context of the query to be extended beyond the simple Clinical Statement. Using the Clinical Document (Composition) as context allows a single XQuery to combine constraints on clinical statements recorded together – in this example, the type of follow-up (1 month, 4 months or 12 months) and the level of compliance.

Entries recorded in the same Composition have the same effective time and can be considered concurrent. However, Entries recorded on different Compositions will have different effective times and the definition of 'concurrent' may depend on the clinical context (e.g. defined in the query as being within 1 hour, 1 week, 3 months, etc.).

3.4. Comparing Values within the Record

It is often necessary to compose queries that compare values of different observations for a patient, for example to find the cohorts of patients who have:

1. sustained a hip fracture whist resident in a care home
2. fractured their left hip more times than their right hip
3. given up smoking for a period of less than 12 months

In query (1) it is not sufficient to combine the cohort of patients who have ever had a hip fracture with the cohort that has been resident in a care home, since the two observations may not have been simultaneous. In (2) the criteria may be satisfied at multiple time intervals, if a patient fractures one left hip, then the other and subsequently re-fractures. To form the cohort in (3) queries must find patients recorded as Current Smoker, then as Ex Smoker and then again as Current Smoker, comparing times between observations.

In general, these types of query are not easy (or may not be possible) to formulate in a single XQuery – they require several separate queries and post-processing of the results sets. In some cases this is feasible with in-memory XML processing, but more complex cases may require export and processing in external statistics packages.

4. Conclusions

An Electronic Health Record system, implemented as a store of HL7 CDA XML documents, provides a rich source of longitudinal data sets for use in clinical studies.

Accurate identification of patient cohorts requires a number of essential features in the EHR – patient merge/unmerge, archival of patients and superseding of erroneous data entries. XQuery can be used to retrieve patient cohorts matching specified constraints and post-processing of results can apply further constraints. However, some more complex constraints can only be applied by exporting results sets and processing in external statistics packages.

References

[1] Chelsom, J. J, Pande I, Summers R, Gaywood I. (2011) Ontology-driven development of a clinical research information system. 24th International Symposium on Computer-Based Medical Systems, Bristol. June 27-June 30 ISBN: 978-1-4577-1189-3

[2] Johnson, S.B., 1996. Generic data modeling for clinical repositories. Journal of the American Medical Informatics Association, 3(5), pp.328-339.

[3] Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F. and Rudolph, S., 2009. OWL 2 web ontology language primer. W3C recommendation, 27(1), p.123.

[4] ISO 13606-1:2008 Health informatics - Electronic health record communication - Part 1: Reference model. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=40784

[5] Dolin RH, Alschuler L, Boyer S et al. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc. 2006;13(1):30-9.

[6] Beale, T., 2002, November. Archetypes: Constraint-based domain models for future-proof information systems. In OOPSLA 2002 workshop on behavioural semantics (Vol. 105).

[7] Boag, S., Chamberlin, D., Fernández, M.F., Florescu, D., Robie, J., Siméon, J. and Stefanescu, M., 2002. XQuery 1.0: An XML query language.

[8] Moner, D., Moreno, A., Maldonado, J.A., Robles, M. and Parra, C., 2012, August. Using archetypes for defining CDA templates. In *MIE* (pp. 53-57).

[9] Liang, K.Y. and Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika, 73(1), pp.13-22.

[10] Chen, R. and Klein, G., 2007. The openEHR Java reference implementation project. *Medinfo*, *129*, pp.58-62.

[11] Frade, S., Freire, S.M., Sundvall, E., Patriarca-Almeida, J.H. and Cruz-Correia, R., 2013, June. Survey of openEHR storage implementations. In Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (pp. 303-307). IEEE.