A Loosely Coupled Framework for Terminology Controlled Distributed EHR Search for Patient Cohort Identification in Clinical Research

Lei ZHAO [a,1], Sarah N. LIM CHOI KEUNG [a], Adel TAWEEL [b], Edward TYLER [a], Ire OGUNSINA [a], James ROSSITER [a], Brendan C. DELANEY [b], Kevin A. PETERSON [c], F.D. Richard HOBBS [d] and Theodoros N. ARVANITIS [a]

[a] *University of Birmingham, Birmingham, United Kingdom*
[b] *King's College London, London, United Kingdom*
[c] *University of Minnesota, Minneapolis, United States*
[d] *University of Oxford, Oxford, United Kingdom*

Abstract. Heterogeneous data models and coding schemes for electronic health records present challenges for automated search across distributed data sources. This paper describes a loosely coupled software framework based on the terminology controlled approach to enable the interoperation between the search interface and heterogeneous data sources. Software components interoperate via common terminology service and abstract criteria model so as to promote component reuse and incremental system evolution.

Keywords. EHR search, terminology controlled, loosely coupled framework, patient identification

Introduction

Patient cohort identification and subsequent recruitment is often a time-consuming and costly process in the whole lifecycle of clinical studies. With the increased adoption of electronic health records (EHR), clinical research staff are now able to search eligible patients on individual EHR repositories. The heterogeneity of EHR systems, however, has presented a major bottleneck to search on multiple EHR data sources, particularly for large-scale multi-center clinical studies. Not only are these EHR systems often implemented in different data structures with diverse access interfaces, the data itself are also encoded in different coding schemes. A number of standards for representation of clinical data in EHRs have been proposed [1], but none of them has achieved universal acceptance. In this paper, we present an extensible software framework for automated distributed EHR search, which accommodates the inevitable heterogeneity in both the EHR data models and coding schemes.

[1] Corresponding Author: Lei Zhao, University of Birmingham, Birmingham B15 2TT, United Kingdom; E-mail: l.zhao@bham.ac.uk.

1. Related Work

In recent years many efforts have been made to connect distributed EHR data sources for clinical research. Particularly, ePCRN (electronic Primary Care Research Network [2]) and I2B2 (Informatics for Integrating Biology & the Bedside [3, 4]) have developed software solutions to support queries across distributed EHR databases to identify potentially suitable subjects for research. The ePCRN developed the pilot infrastructure for the US Federation of Practice-based Research Networks and the implementation was based on open source grid middleware OGSA-DAI [5] and Globus Toolkit [6]. The I2B2 concentrated mainly on hospitals and the entire architecture was built on web services. Although ePCRN and I2B2 have a different community focus and are different in their implementation details, they share many common approaches. In order to address the EHR data model heterogeneity, both have chosen a single cross-site standard data model and have mapped source data to the standard model. ePCRN adopted the ASTM standard CCR (Continuity of Care Record [7]), while I2B2 chose to develop their own "star schema" where observations about patients were stored in a central observation fact table. In order to address the coding scheme heterogeneity across data sources, both have chosen a "terminology controlled" approach where a standard terminology is mapped to local original codes. ePCRN used the public web service of NCI Metathesaurus [8] to provide a mapping of concepts to terms among source coding schemes. I2B2 developed their own set of standard concepts and also added support for coding schemes frequently used in the US hospital systems such as International Classification of Diseases (ICD), National Drug Code (NDC), and Logical Observation Identifiers Names and Codes (LOINC). Both projects also developed a graphical user interface to help clinical researchers specify query criteria, using provided standard terminology. However, the constructed queries are bound to the database schemas and are restricted to a certain logical structure.

The initial version of ePCRN implementation essentially followed a bottom-up approach, where the initial search system was designed around the underlying data model, i.e. CCR. This has limited its flexibility and interoperability with other systems. Expanding on the ePCRN approach, we propose a more extensible, loosely coupled search framework in the new version of the technology to alleviate these limitations. The current extension and implementation follows a more top-down, model-driven approach.

2. A Loosely Coupled Search Framework

2.1. Conceptual Architecture

Instead of enforcing a single standard EHR model across the whole system, the new architecture focuses on the design of interoperable interfaces between the search component and local EHR data sources (Figure 1). A specially designed terminology web service allows end users to browse and select standard concepts from the common terminology service (CTS), which provides mappings from standard concepts to local coding schemes, builds search criteria using selected concepts, and submits to the search coordinator. The search coordinator distributes the search request to local data sources and coordinates their execution. In the current ePCRN implementation, administrators can configure a limit on how many concurrent searches can be running

at the same time, in order to avoid overloading the system. In the future, advanced scheduling algorithms can be developed to dynamically optimize system throughput or give certain users higher priority. Similar to the initial ePCRN version, the search criteria are captured in an abstract representation, which is neutral to local data source implementations. However, in this approach local search brokers translate the abstract criteria into executable local statements suitable to run the search on the local EHR system and return the result to the coordinator. Finally the coordinator aggregates individual results and presents back to the users. This architecture enables loose-coupling between data sources and the search interface, and thus allows flexible implementation options for individual EHR source.

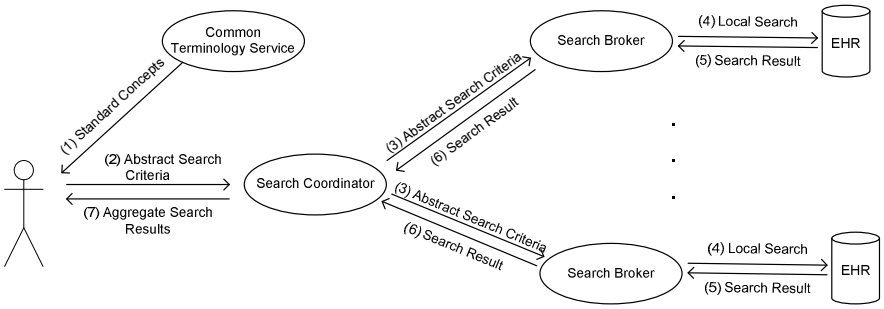

Figure 1. Conceptual architecture.

2.2. Common Terminology Service

Instead of integrating and storing terminology mappings at each data source, which is the I2B2 approach, the common terminology service provides a unified platform to deliver terminology mappings and is shared by all entities. A centralized terminology service is easy to reuse, maintain and evolve. The Unified Medical Language System (UMLS) Metathesaurus is a comprehensive multi-lingual biomedical terminology database which covers many terminologies used for clinical care, translational and basic research and provides cross-mappings between these source terminologies [9]. The NCI Metathesaurus [10], version 4.3, is based on the UMLS Metathesaurus, but only supports the English language, which has restricted its usage outside the US. With the aim of supporting European languages and including more European terminologies, we have developed an integrated terminology service [11] (as part of the TRANSFoRm project [12]), based on a UMLS Metathesaurus subset. The terminology service covers the most commonly used coding schemes in European primary care systems, including SNOMED CT, ICD-10, ICPC, Read Codes, etc. The service was implemented using LexEVS 5.1 and is currently being upgraded to version 6. LexEVS is an open source general purpose terminology service solution which supports HL7 CTS 2 Draft Standard for Trial Use [13].

2.3. Abstract Search Criteria Model

The abstract search criteria model is key in achieving interoperability between individual data sources (Figure 2). This model is based on the initial version of ePCRN,

however has been extended to allow a higher level of abstraction, by generalising search criteria concepts. A single criterion follows the pattern of {concept, value, time range} which provide basic building blocks to construct arbitrarily complex logical statements. The model is easy to extend by introducing more complex value type and new time range specifications.

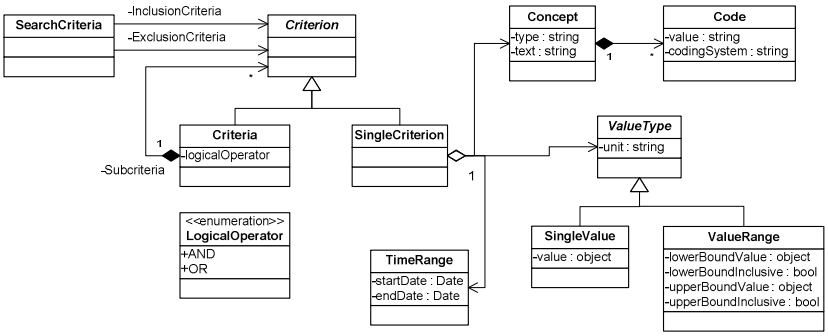

Figure 2. Abstract search criteria model.

2.4. Local Search Translation

```
<SearchCriteria>
  <InclusionCriteria>
    <Concept>
      <Type>Disease or Syndrome</Type>
      <Text>Diabetes Mellitus</Text>
      <Code>
        <Value>C0011849</Value>
        <CodingSystem>UMLS</CodingSystem>
      </Code>
    </Concept>
  </InclusionCriteria>
</SearchCriteria>
```

```
SELECT COUNT(DISTINCT C.ANON_ID)
FROM CCR_XML C
WHERE
/* --- Inclusion Criteria --- */
/* --- Problem Diabetes Mellitus --- */
EXTRACTVALUE(CCR_XML_String,'/
ContinuityOfCareRecord/Body/Problems/
Problem/Description/Code/Value') REGEXP
'Cyu2.|^Cyu2.| C10..|^C10..|
C1000|^C1000|'
```

Figure 3. Translated search statement snippets.

We have developed the search broker for the pilot ePCRN data repositories that we are setting up in the UK. The data are coded in Read Codes version 2. The actual implementation uses MySQL to store CCR XML strings. We do not have space to present the complete algorithm here, so instead we use code snippets to demonstrate the process (Figure 3). The example criteria encode a single criterion to search on the concept Diabetes Mellitus whose concept identifier is C0011849 in UMLS. The automatically generated query statement searches the CCR problem sections, based on the concept's semantic type Disease or Syndrome, for the mapped Read Codes. As demonstrated it is relatively straightforward to translate the abstract search criteria into CCR queries. We are also investigating connection with I2B2 data repositories.

3. Discussion

This paper concentrates on the distributed EHR search framework. A system interacting with EHR data also needs to address data privacy concerns, maintain

institutional autonomy, fulfill regulatory obligations and data sharing agreements. It is therefore essential for the search framework to interoperate and integrate with de-identification, authorisation, and auditing frameworks, which require considerable further research.

Many challenges exist in reusing data from EHR systems for clinical research [14]. The main focus of EHR is to support healthcare transactions. It is still not clearly understood what data elements should be supported and what search capability should be provided by EHR systems from the clinical research perspective. Therefore, we advocate a loosely coupled software framework to promote component reuse and incremental system evolution, and encourage more collaboration and research on identifying common search requirements from various clinical studies.

Acknowledgement

This work was supported in part by the National Institutes of Health, under contract No. HHS268N200425212C, "Re-engineering the Clinical Research Enterprise"; the European Commission – DG INFSO (FP7 247787) for the TRANSFoRm project; and the National Institute for Health Research Birmingham and Black Country Comprehensive Local Research Network (NIHR BBC CLRN).

References

[1] Ohmann C, Kuchinke W. Future developments of medical informatics from the viewpoint of networked clinical research. Interoperability and integration. Methods Inf Med 2009;48:45 – 54

[2] Delaney BC, Peterson KA, Speedie S, Taweel A, Arvanitis TN, Hobbs FDR. Envisioning a Learning Health Care System: The Electronic Primary Care Research Network, A Case Study. Ann Fam Med. 2012; 10 (1):54 – 59

[3] Murphy SN, Weber G, Mendis M, Chueh HC, Churchill S, Glaser JP, Kohane IS. Serving theEnterprise and beyond with Informatics for Integrating Biology and the Bedside (i2b2). J Am Med Inform Assoc. 2010; 17 (2):124 – 130

[4] Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): A prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009; 16 (5): 624 – 630

[5] Antonioletti M, Atkinson MP, Baxter R, Borley A, Chue Hong NP, Collins B, Hardman N, Hume A, Knox A, Jackson M, Krause A, Laws S, Magowan J, Paton NW, Pearson D, Sugden T, Watson P, Westhead M. The Design and Implementation of Grid Database Services in OGSA-DAI, Concurrency and Computation: Practice and Experience. 2005; 17 (2-4): 357 – 376

[6] Globus Toolkit, www.globus.org (2012 Jan 25)

[7] ASTM International Standard Specification for Continuity of Care Record (CCR). ASTM Designation E2369-05. 2005 Dec. www.astm.org/Standards/E2369.htm (2012 Jan 25)

[8] NCI Metathesaurus (NCIm), ncimeta.nci.nih.gov (2012 Jan 25)

[9] Unified Medical Language System (UMLS), www.nlm.nih.gov/research/umls/ (2012 Jan 25)

[10] Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thesaurus and Ontology. Journal of Web Semantics 2003; 1(1).

[11] Lim Choi Keung SN, Zhao L, Tyler E, Arvanitis TN. Integrated Vocabulary Service for Health Data Interoperability. Fourth International Conference on eHealth, Telemedicine, and Social Medicine (eTELEMED), Spain. IARIA. 2012: 124 - 127

[12] TRANSFoRm, www.transformproject.eu (2012 Jan 25)

[13] LexEVS, https://cabig.nci.nih.gov/tools/LexEVS_Server (2012 Jan 25)

[14] Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med. 2009; 48:38 – 44