A Framework for Vocabulary Controlled Queries of Distributed Electronic Healthcare Records

Sarah N. Lim Choi Keung1, Lei Zhao1, Ire Ogunsina1, Theodoros N. Arvanitis2 and Edward Tyler1

1 Department of Primary Care Clinical Sciences
2 School of Electronic, Electrical and Computer Engineering
University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom
{s.n.limchoikeung, l.zhao, i.ogunsina, t.arvanitis, e.tyler}@bham.ac.uk

Abstract. Patient health information is usually stored using different conceptual representations and coding conventions in different health record systems and research databases. The linkage capacity of these data sources strongly determines the reuse of clinical care information within clinical research and vice versa. As part of efforts in addressing the issue of system interoperability within primary and secondary care, we are investigating the use of controlled vocabularies across both healthcare sectors. In this paper, we present research work being carried out in our institution, involving the development of a vocabulary service that allows users to access clinical terminology and vocabulary translations through both a web interface and web service Application Programmable Interface (API). Another area of work is the transformation of federated clinical data for enhanced query performance and analysis.

Keywords: Interoperability, controlled vocabularies, electronic health record.

1 Introduction

The clinical research process is generally long and costly. With the increased use of electronic healthcare records (EHRs), the aim is to promote the secondary use of existing clinical and research data for clinical research. The increasing use of health information systems and Web technologies has led to multiple health records being maintained for the same patient [1]. Therefore there is a need to link EHRs, especially due to the heterogeneity of the systems and data repositories in use. The context of our research involves the linkage of EHR data within primary care and within secondary care research, as well as linkage of primary and secondary care for clinical research. One of the applications that we are currently involved in is patient cohort identification and recruitment to clinical research studies and clinical trials. We are looking into solutions that are scalable and flexible enough to allow for efficient
linkage of data repositories and systems whilst handling their individual features and complexities.

In this paper, we present the controlled vocabulary linkage research work at the University of Birmingham. The remaining parts of this paper are structured as follows. Section 2 discusses the issues of heterogeneity in EHR data and the relevant solutions proposed in the literature. In Section 3, we introduce our approach to vocabulary controlled linkage and describe the linkage aspects in our research work. Section 4 describes a vocabulary service developed to allow clinical terminology access and vocabulary translations via a web service. Section 5 presents linkage solutions in the UK primary care and secondary care research domains. The focus of these solutions is clinical research, in particular to favour the secondary use of EHR data in clinical research. Next, Section 6 provides a discussion of the challenges, experiences, and lessons learnt. Finally, we conclude with the achievements so far, the limitations and future work.

2 Heterogeneity of EHR Data

The heterogeneity of EHR data arises from the use of disparate health information systems within primary and secondary care, both for routine patient care and research. In the UK primary care for instance, where 100% of General Practitioner (GP) clinics use EHRs [2], there are several GP systems that are most commonly in use, including EMIS, and Vision [3]. These systems are for supporting routine patient care and were not designed for use in clinical and biomedical research. However, as suggested by Kahn [4], if 30-50% of research data can be found in EHRs, this is potentially a very useful for supporting biomedical research. Towards the achievement of this goal, we aim to link heterogeneous data so that biomedical researchers can have access to a bigger pool of data. This can have a direct benefit to patient care through translational research. The issues of heterogeneous data include the variation in meaning of the terms used to describe clinical activities, as well as different terms used to mean the same thing. In the UK, this issue is particularly evident in secondary care research, where the dual roles of clinician and investigator have resulted in standalone EHR systems that are only suitable to those who developed them.

2.1 Outline of Approach

We view data linkage at two main levels: information model representation and vocabulary-controlled terminology support. The information model allows relevant patient clinical and research data to be represented in an EHR. Common information models enable data repositories to be more easily linkable and there is less overhead than the need to unify the EHR models. The vocabulary-controlled terminology
ensures that the accurate meaning is used and preserved, especially as data from different repositories are linked. We are concerned with the linkage of both primary care and secondary care research data to support the clinical research process. Each of these two domains has its own particular issues and complexity. These need to be addressed in order to achieve seamless interoperability for the application areas we are interested in. We have developed a vocabulary service that supports the translation of different medical terminologies and is a cross-cutting feature of the data linkage.

2.2 Related Work

Our approach is based on and extends the work achieved in the ePCRN project [5], a National Institutes of Health (US-funded) project investigating a sophisticated electronic infrastructure to support the design and implementation of randomised clinical trials, while facilitating translational research in primary care in the United States. ePCRN provides the researcher with access to a set of applications called the Study Design Workbench. The workbench includes applications that assist the researcher to create and define new studies, using a standard template that includes all the required fields, defined by the World Health Organization and the US National Institute of Health ClinicalTrials.gov service. The workbench also helps the researcher define specific eligibility criteria by interfacing directly with National Cancer Institute’s Enterprise Vocabulary Services (NCI EVS)\(^1\) system. This facility provides access to a large variety of standardised data elements and coding systems. Additionally, the workbench supports the translation of eligibility criteria into actual queries and interacts with the ePCRN Practice Gateway to provide counts of potentially eligible subjects. The ePCRN Practice Gateway is a suite of locally installed applications that assist healthcare providers (clinics) in reviewing research opportunities, recruiting patients, reviewing and compiling requested data, logging of all local data transactions, and providing feedback to researchers about the recruitment process. More specifically, the Practice Gateway imports data from the local EHR into its own database and uses the Globus OGSA-DAI framework with PKI certificates to accept queries.

3 Vocabulary Service

A medical terminology is a controlled vocabulary of well defined medical terms with specified relationships between those terms, capable of being interpreted by both humans and computers. Medical terminologies establish a foundation for clinical information interoperability. They provide systematic approaches for organising

\(^1\) https://cabig.nci.nih.gov/concepts/EVS/
clinical information and define the semantics of information using consistent and computable mechanisms, so clinical data can be processed by computers, and shared between data sources using the same terminology. Nowadays, however, many different medical terminologies are used in primary care EHR systems and secondary care research databases, such as SNOMED CT, ICD-10, and Read Codes, which makes it difficult to link data from these systems.

The ability of a vocabulary service to provide consistent representation and access to a broad set of terminologies enables multiple disparate terminology sources to be available to the community and helps to ensure consistency across the domain space. The Health Informatics team based in the Primary Care Clinical Research and Trials Unit (PC-CRTU) at the University of Birmingham has developed a vocabulary service to facilitate EHR data integration from primary care GP systems and secondary care research databases. This vocabulary service uses Unified Medical Language System (UMLS) Metathesaurus as the major terminology source and employs the LexEVS technology to access the backend UMLS vocabulary database. Additional terminologies such as Read Codes version 2 (Read Codes v2), which are relevant to UK Primary Care but are not included within the UMLS Metathesaurus hosted terminologies, are also integrated as well as their mappings to UMLS Metathesaurus.

3.1 UMLS

The Unified Medical Language System (UMLS) is a compendium of many controlled vocabularies in the biomedical sciences [6]. It provides a mapping structure among these vocabularies and thus allows the translation among the various terminology systems. UMLS consists of Knowledge Sources (Metathesaurus, Semantic Network and SPECIALIST Lexicon) and a set of software tools. The UMLS Metathesaurus is a large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them [7]. The Metathesaurus contains over five million terms, organised by concept and is built from the electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms used in various medical domains. Each concept has specific attributes defining its meaning and is linked to the corresponding concept names in the various source vocabularies. Additionally, each concept in the Metathesaurus is assigned one or more semantic types, which are broad subject categories, such as Disease, Syndrome and Clinical Drug, which are linked with one another through semantic relationships, such as “isa” links [8].
3.2 NCI EVS and LexEVs

NCI Enterprise Vocabulary Services (EVS) is a project of the US National Cancer Institute Center for Biomedical Informatics and Information Technology (NCI CBIIT). NCI EVS provides controlled terminology and ontology that is the semantic base for caCORE, caBIG®, and the new CBIIT semantic infrastructure [9]. LexEVS is the open-source software release of NCI EVS, which implements a comprehensive set of software and services to load, publish and access controlled terminologies. LexEVS supports a wide range of data formats, including UMLS RRF, OWL, OBO, HL7 RIM and LexGrid XML.

NCI EVS develops and supports two vocabulary sources: the NCI Thesaurus (NCIt), a cancer-focused terminology, and the NCI Metathesaurus (NCIm). The NCIm is based on UMLS Metathesaurus and thus provides a mapping of concepts to terms among multiple vocabularies. The NCI Metathesaurus is more US-oriented, and thus has less European coding systems and only supports vocabularies in the English language. With a vision to include more European terminologies and support vocabularies in other European languages, we develop and deploy a local vocabulary service (as part of the TRANSFoRm project [10]), using the NCI EVS’s LexEVs.

3.3 TRANSFoRm Vocabulary Service

The TRANSFoRm Vocabulary Service is designed to allow end users to search and retrieve clinical vocabulary concepts and associated content, through both a web interface and a web service API. The service uses the LexEVs (version 5.1) to access a customised UMLS vocabulary database, and uses direct Java Database Connectivity (JDBC) to access other vocabulary databases, such as Read Codes v2, which are not included within the UMLS Metathesaurus hosted terminologies. The Vocabulary Service provides both a web interface for web client access and a web service interface for programmatic access. It also uses the UMLS vocabulary database through LexEVs, with additional vocabulary databases being accessed directly. LexEVs builds an index on the vocabulary databases using Lucene technology², which enables high performance data searching.

We customised a subset of the UMLS Metathesaurus (2010AA release) using MetamorphoSys [11], a free tool distributed with the UMLS, and imported into a local LexEVs database to serve as a vocabulary data source. The subset consists of those vocabularies most relevant to the European primary care research and clinical

² http://lucene.apache.org/
practice, including SNOMED CT, ICD-9, ICD-10, ICPC-2, ICPC2-ICD10 Thesaurus, ICPC, and Read Codes version 3 (also known as Clinical Terms v3).

![Diagram](image)

**Fig. 1.** UMLS-to-Read Codes v2 association through SNOMED CT

Read Codes v2 terminology is still widely used in the UK primary care systems, but it is not available from UMLS Metathesaurus. In order to cater for this need, a “Read Codes v2” corpus of terms and their associated mappings have been added in a customised database. The UK NHS Connecting for Health Terminology Centre [12] has provided mappings from Read Codes v2 to SNOMED CT. The Read Codes v2 database in TRANSFoRm vocabulary service is set up based on this mapping so that Read Codes v2 concepts can be linked to a UMLS search. As shown in Fig. 1, UMLS concepts generally map to SNOMED CT codes, hence the Read Codes which map to the same set of SNOMED CT codes can be linked to the associated UMLS concepts.

The aim for using the vocabulary service is to enable searching by concept, which is a key feature of the eligibility criteria and query formulation tool that will support researchers to identify eligible patients for their studies. Through an example search, we demonstrate how the vocabulary service works in helping to map concepts to a wide range of medical terminologies, hence enabling searches across heterogeneous healthcare data. Fig. 2 shows a screenshot of the results for an example search for a clinical term (“Type II Diabetes”) on the web-based interface of the TRANSFoRm vocabulary service. In this scenario, the user wants to know the Read Code version 2 for Type II diabetes. The term is entered in the search box and the ReadCodes2 option is ticked to include the terminology as it is not provided within UMLS.

The elements of the result are described below (see annotated screenshot Fig. 3).

1. The UMLS concepts matching the search term are displayed as a list of concepts, shown in the top left box. The search term is matched to all the available terminologies; in this case, all those included within UMLS and Read Codes v2, and 25 concepts were returned for the search term “Type II Diabetes”.

2. Each UMLS concept has a unique UMLS code, semantic type, and definition. For instance, in Fig. 2, the selected concept is “Diabetes Mellitus, Non-Insulin-Dependent” (highlighted by the dashed box in Fig. 3). For this concept, further information is displayed on the top right of Fig. 2 and Fig. 3. The UMLS Code
"C0011860" provides the unique UMLS identifier for that concept, while the Semantic Type "Disease or Syndrome" indicates the subject category within the categories of UMLS concepts.

3. For the selected concept, its super and sub concepts are rendered on the display, describing the important relationships with other concepts. For instance, a super concept for “Diabetes Mellitus, Non-Insulin-Dependent” is “Diabetes Mellitus”, while the sub concept “Diabetes mellitus type 2 in obese” is a more specific concept.

4. Codes from hosted vocabularies (including language and textual description) are also provided for the selected concept. In Fig. 2, only the vocabularies in English have been selected. For instance, in Read Code v2, there are three terms that have been matched to “Type II Diabetes”.

Fig. 2. Result screenshot for the example search using the TRANSFoRm Vocabulary Service web interface.
4 Secondary Use of Primary Care Data in Research

Primary care in the UK consists of Primary Care Trusts (PCTs), which are regional organisations overseeing a group of GP practices. Existing local initiatives at some PCTs include the integration of patient data from different GP systems into a unified EHR repository for commissioning purposes, among others. To promote reuse of clinical data for research, we are investigating the usefulness of this federated data for feasibility studies, such as for initial patient cohort identification with the view of future recruitment of patients to clinical studies and trials.

The PCT we are currently working with is responsible for GP practices that serve around 300,000 patients. We have performed an initial analysis of the data within the anonymised repository in order to evaluate the data quality and the consistency of
coding with Read Codes v2. As described in our previous work [13], a significant proportion of the codes used does not match the reference Read Codes v2. Nevertheless, the codes that do match the reference codes may be sufficiently rich to be used for initial patient cohort identification in feasibility studies. We are now in the process of making the unified data searchable, according to specific sets of inclusion and exclusion criteria. This is due to the Read Codes v2 not being sufficiently organised to enable easy and consistent querying of the data. We are investigating the use of the Vocabulary Service to translate Read Codes v2 to other coding schemes which are more richly structured, such as SNOMED CT.

5 Secondary Care Research

5.1 Secondary Care Research Model

Health Level Seven (HL7)\(^3\), is an international organization providing standards of interoperability across the entire spectrum of health care enterprise. One of its objectives is the provision and support of seamless exchange of information between medical applications such as between different patient record systems or between patient record systems and laboratory systems. The current version of the HL7 standard, the HL7 v3 – is fundamentally different from previous versions by being re-engineering from the ground up; adopting a new and complete object oriented paradigm. HL7 focuses on the capacity of an application to generate and interpret messages sent between applications and systems. At the core the HL7 v3 is the Reference Information Model (RIM). This object based model is an attempt to uniquely represent the health care domain sets of data using 6 high level structural classes. The RIM structure complies with the specifications of HL7 v3 by encapsulating or representing the semantic and lexical connections between information contained in messages [14]. In HL7, a vocabulary domain (concept domain) is a set of concepts and HL7 supports a one-to-many relationship between a given concept and coding systems (implementation), although a specific coded attribute in the RIM will be associated with only one vocabulary domain [15,16].

We are developing HL7 compliant dual-purpose secondary care systems for clinical care and research, to be initially implemented across 5 initial medical specialties. Information models are designed for each specialty subsystem in line with HL7 - enabling messaging and interoperability between subsystems. The design harnesses one of HL7’s strengths, its loose coupling with vocabulary mechanisms and terminology sources such as SNOMED CT and ICD-10. Rather than having its own

\(^3\) http://www.hl7.org/
vocabulary dictionary scheme, it integrates well with many alternatives. Apart from the capacity of the system to access multiple vocabulary databases via the vocabulary service in the vocabulary module as shown in Figure 3, the system also allows the use of ISO-based identifiers, namely Object Identifier Definitions (OIDs). This is a way of uniquely identifying coding systems and identifier namespaces, particularly in HL7 v3 and may be used in representing and identifying coding schemes such as Read Codes and SNOMED CT, as well as external identifier schemes such as unique patient number (e.g. UK NHS number) and GP numbers.

![Simplified component diagram of system cross section](image)

**Fig. 4.** Simplified component diagram of system cross section

Within the information model, a vocabulary domain module exists which effectively is a set of reusable libraries typically providing the following:

1. Use the vocabulary service to access multiple vocabulary databases as shown in Fig. 4. A data element/container may be populated at the request of a subsystem; the request is in turn passed on to the vocabulary module which may consume the vocabulary web service, or may return, for example, a list of race data containing the contents of the race vocabulary domain within a specified context identifiable by its OID.

2. Facilitate correct OID setting in instance value. For example, race as a vocabulary or concept domain definition may be based on SNOMED CT coding scheme in a given subsystem. A potential pitfall is prevented during inter-subsystem activity where multiple coding schemes are utilized by providing additional information about the instance attribute value of concept domains. This additional information may be used for further processing such as with entity-mappings. Table 1 shows some of the relevant OIDs and coding systems in the UK.
3. An additional use of OIDs is inherent in its hierarchical structure making it possible to leverage it as an ontology mechanism and traverse different tree nodes providing information on versions, categories and hierarchical relationships that may exist.

<table>
<thead>
<tr>
<th>OID</th>
<th>Coding System</th>
<th>Description</th>
<th>Issuer</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.16.840.1.113883.6.29</td>
<td>The Read Codes Version 2</td>
<td>The Read Codes Version 2</td>
<td>National Health Service (UK)</td>
</tr>
<tr>
<td>2.16.840.1.113883.6.6</td>
<td>NHS Clinical Terms Version 3 (Read Codes)</td>
<td>NHS Clinical Terms Version 3 Codes</td>
<td>National Health Service (UK)</td>
</tr>
<tr>
<td>2.16.840.1.113883.6.5</td>
<td>SNOMED</td>
<td>SNOMED codes</td>
<td>College of American Pathologists</td>
</tr>
</tbody>
</table>

5.2 Controlled Vocabularies for Secondary Care Research

Currently, secondary care research systems are standalone and developed mainly for the purpose of supporting an individual clinician/investigator. Often, no controlled vocabulary is used to describe the clinical or research activity, resulting in the data being difficult to share or reuse due to the lack of meaning and consistency. With this in view, we are currently researching the coding schemes that would be the most appropriate for use by the clinicians/investigators. For instance, ICD-10 is used in the hospital systems in the UK and was originally proposed to the users. However, several issues were identified. Firstly, due to the specialty domain, consultants found it difficult to find existing ICD-10 codes and descriptions that fit their purposes. Additionally, there were sometimes issues with being able to map several codes to one condition.
6 Discussion and Conclusions

6.1 Vocabulary Service

The current vocabulary service implementation is based on LexEVS 5. A newer version of LexEVS, namely LexEVS 6 has been released, which adds comprehensive support of emerging HL7 terminology service standards [18]. We plan to investigate and migrate to LexEVS 6 to align with the emerging standard. With a vision to support clinical research across Europe in the longer term, we plan to investigate and add more European focused vocabularies, for example, ICPC-2 in different European languages. Vocabularies related to medication present a big challenge, which definitely needs significant future work.

6.2 Reuse of Primary Care Data for Research

Clinical data used in the routine healthcare of patients have the potential to contain enough data of research quality that will enable the clinical research process to be shortened. Currently in the UK, research staff have to go into individual GP practices, and with the help of practice staff, query the GP system for patients that fit the eligibility criteria. A semi-automated approach where unified anonymised data from GP systems are queried for patients that fit the criteria can significantly reduce the time taken to do feasibility studies. For instance, we may want to use the unified data from the PCT with the ePCRN Workbench to help a researcher find the count of patients that fit a particular set of inclusion and exclusion criteria to determine if such a study is feasible. Future work in this area will look at the ways to adapt the ePCRN Workbench approach in the UK healthcare context. Issues of different coding schemes and organizational structures need to be taken into account.

6.3 Secondary Care Research

This relatively new area of research has shown its complexities, but the efforts by medical professionals to be actively engaged with the changes to make EHRs for specialty research interoperable, shows much promise. In the UK, secondary care research is far less integrated than the primary care domain. The development of a secondary care research model is a starting point to represent the structure of an EHR within a research specialty. We are currently working on restructuring the first one of the 5 specialty research EHRs based on the model presented. We intend to model the remaining specialties in a similar manner such that they can be interoperable. Other areas of research include the interoperability of systems concerning medications, which is also an ongoing research area within primary care in Europe. In terms of adopting controlled vocabularies in secondary care research EHRs, we are investigating other coding schemes that are more flexible and allow a greater
capability to describe a condition more accurately, such as SNOMED CT. Future work will look into the most appropriate coding schemes to enable interoperability with other systems and data repositories, as well as serving the dual research and clinical care purpose of secondary care research systems.

Acknowledgements

This work was supported in part by the European Commission - DG INFSO (FP7 247787) for the TRANSFoRm project, and the UK National Institute for Health Research Birmingham and Black Country Comprehensive Local Research Network (NIHR BBC CLRN).

References


