Cohort Identification for Clinical Research: Querying Federated Electronic Healthcare Records Using Controlled Vocabularies...

Article · March 2012
Source: PubMed

9 authors, including:

Sarah N. Lim Choi Keung
The University of Warwick

24 PUBLICATIONS  41 CITATIONS

SEE PROFILE

Brendan Delaney
Imperial College London

238 PUBLICATIONS  4,729 CITATIONS

SEE PROFILE

Adel Taweel
Birzeit University

89 PUBLICATIONS  429 CITATIONS

SEE PROFILE

Theodoros Arvanitis
The University of Warwick

212 PUBLICATIONS  1,769 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Identifying Primary Care Practice Components Leading to Optimal Diabetes Care in Patient-Centered Medical Homes View project

Diet4Elders (http://www.diet4elders.eu/) View project
Cohort Identification for Clinical Research: Querying Federated Electronic Healthcare Records Using Controlled Vocabularies and Semantic Types

Sarah N. Lim Choi Keung, PhD¹, Lei Zhao, MSc¹, Edward Tyler, MSc¹, Adel Taweel, PhD², Brendan Delaney, MD², Kevin A. Peterson, MD, MPH³, Stuart M. Speedie, PhD⁴, F.D. Richard Hobbs, MB ChB⁴ and Theodoros N. Arvanitis, DPhil¹
¹University of Birmingham, United Kingdom; ²King’s College London, United Kingdom; ³University of Minnesota, Minneapolis, MN; ⁴University of Oxford, United Kingdom

Abstract
In the United Kingdom (UK), local initiatives have started to federate electronic healthcare records from different primary care clinical systems, mainly for the purposes of ensuring that health care services effectively meet the needs of the population. The use of such information is being investigated for clinical research, notably in patient cohort identification and recruitment. To achieve these aims, it is essential that the information from different systems can be searched from a single interface. While interoperability is a widely researched topic, interoperable methods and data sources in primary care are largely missing. This paper describes our approach to enabling primary care data in England to be searchable on a platform developed for performing large national collaborative primary care research studies throughout the United States.

Introduction and Background
Patient cohort identification and recruitment is a time-consuming and costly part of the clinical research process. Currently, much of this process is carried out by research staff and primary care staff in individual practices to query the clinical systems for patients matching a particular set of criteria. The use of electronic health records (EHRs) in 100% of practices in UK primary care¹ leads to the opportunity of computationally analysing and processing patient data. Local initiatives have started to integrate patient data from heterogeneous systems for commissioning purposes, and tapping into this resource for clinical research is an important by-product of the integration process. With about 60% of the integrated data correctly coded with the Read Codes Version 2 (RCV2), widely used in primary care², our work aims to tackle the technological interoperability issue of RCV2 with other vocabularies. Our work is based on the experience gathered during the US-funded ePCRN project to apply similar technologies to support researchers in the UK primary care.

Methods
Firstly, as the UMLS does not support RCV2, we have developed a local vocabulary, which allows end users to search and retrieve clinical vocabulary concepts and associated content, through both a web interface and a web service API, using LexEVS to access a customised UMLS vocabulary database and direct Java Database Connectivity to access other vocabulary databases, such as RCV2. Secondly, the eligibility criteria used by the researchers need to have the same meaning, irrespective of the source vocabulary. The ePCRN Workbench enables EHR data to be queried according to a set of eligibility criteria specified by a researcher, and the EHR data is mapped to a Continuity of Care Record (CCR) schema. The integrated data schema is different from the CCR and our approach to maintaining semantic interoperability is to use the UMLS semantic types as the main indicator of the types of EHR content and to map semantic types to the five CCR categories used in the Workbench.

Results and Discussion
We have tested the Eligibility Criteria Tool on a randomly selected sample of 25 patients from the extract. The sample contains 19,035 clinical activities and out of the 1831 distinct RCV2 codes, 963 have semantic types (53%). 54 semantic types (individual and in combination) have been identified following the removal of duplicates. These semantic types have been manually mapped to one of the five CCR categories (Demographics, Clinical Problem, Laboratory Test, Vital Sign and Medication) where applicable. The sample set of patient records have then been transformed into a CCR format and has been used to return counts of patients matching a specific eligibility criteria set. The counts have been verified against direct database queries on the original data extract. In future work, we will consider the inclusion of prescription data; currently there is no medication RCV2 to SNOMED CT maps.

References