EU FP7 TRANSFoRm Project: Query Workbench for Participant Identification and Data Extraction

Lei Zhao, MSc[1], Sarah N. Lim Choi Keung, PhD[1], Christopher Golby, MSc[1], Jean-François Ethier, MD[2], Vasa Curcin, PhD[3], Hilde Bastiaens, MD, PhD[4], Anita Burgun, MD, PhD[2], Brendan C. Delaney, MD[3], Theodoros N. Arvanitis, DPhil[1]
[1]Institute of Digital Healthcare, WMG, University of Warwick, UK; [2]INSERM UMR 1138, Université Paris Descartes - Sorbonne Paris Cité, France; [3]Department of Primary Care and Public Health Sciences, King's College London, UK; [4]Department of Primary and Interdisciplinary Care, University of Antwerp, Belgium

Abstract

The TRANSFoRm query workbench facilitates building eligibility criteria and querying heterogeneous phenotype/genotype datasets. With archetypes for clinical data specifications and support of logical and temporal constraints, complex queries can be built to expedite research processes.

Introduction and Background

The EU FP7 TRANSFoRm project (www.transformproject.eu) aims to develop a common digital infrastructure to support the learning healthcare system. Query Workbench (QW) is one of TRANSFoRm software tools to support epidemiological research using primary care electronic health records, genomic and other datasets.

Methods

While most existing solutions, e.g. i2b2, employ Extract, Transform, Load functions on data from different sources into a central data warehouse, TRANSFoRm uses a semantic mediation approach to connect distributed heterogeneous phenotype/genotype data repositories without needing them to be in a common schema. Instead, clinical data models in TRANSFoRm are described via openEHR archetype definition language. Archetypes are then mapped to local database schema through TRANSFoRm Clinical Data Integration Model (CDIM) and Data Source Models [1,2]. CDIM is an ontology of primary care domain elements that captures the structural and semantic variability, with archetypes and CDIM bindings on the implementation side.

Results and Discussion

Based on the archetype models, researchers can formulate queries with complex Boolean and temporal constraints to flag eligible subjects in target data sources and define data items to retrieve for the flagged subjects. Once appropriate authorizations are granted by individual data sources, the data are extracted, linked and anonymized in a trusted third party for researchers to access. An example is to extract HbA1c results for Type 2 Diabetes patients which should be within 6 months before patients' Sulfonylurea treatment started and within 3-12 months after their commencing Sulfonylurea. This tool is being tested and evaluated by a pilot study which investigates associations between well selected single nucleotide polymorphisms (SNPs) in Type 2 diabetic patients with variations in drug response to Sulfonylurea. This study aims to include a total of 500 patients from 5 pilot sites across UK, the Netherlands and Belgium.

References

1. Lim Choi Keung SN et al. Detailed Clinical Modelling Approach to Data Extraction from Heterogeneous Data Sources for Clinical Research. AMIA 2014 Joint Summits on Translational Science, 2014.
2. Ethier JF et al. A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm. J Am Med Inform Assoc 2013;20:5 986-994