Architecture for the harmonization of clinical cohort data in the IMI EMIF* project

Rudi Verbeeck, Luiza Gabriel and Michel van Speybroeck

Janssen Pharmaceutical Companies of Johnson & Johnson, Beerse, Belgium

Abstract. The European Medical Information Framework (EMIF) project under the Innovative Medicines Initiative (IMI) developed an architecture and harmonization framework for assessment and analysis of real world data. The framework was applied to construct a pooled data set of retrospective cohorts of Alzheimer's disease (AD) patients. Information on local variables and harmonization targets and the mapping between them is encapsulated into a structure called a knowledge object. Analysis applications use a distributed knowledge object library of pooled data.

Keywords: IMI EMIF, data pooling, biomedical ontologies, data privacy

1 Introduction

Secondary use for research purposes of medical data that is collected in real world, longitudinal studies requires measures to ensure privacy protection and correct interpretation of local data. Disease specific research cohorts typically contain a deep phenotypic characterization of the cohort and informed consent forms often allow access to anonymized patient level data. However, obstacles to access to the data can be large and include differences in technical implementation, local protocols and codes.

2 Data harmonization

Analysis tools in EMIF allow a direct comparison of data source variables. Source measurements need to be transformed into harmonized variables to ensure semantic compatibility to a level that allows meaningful conclusions from these comparisons.

EMIF tries to increase efficiency by distributing the effort to people with the appropriate knowledge and by making the results re-usable for new research questions. A harmonization framework was developed that allows data source custodians (providers) to capture local protocol information and researchers (consumers) to define

* The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under EMIF grant agreement n° 115372, re-sources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in-kind contribution (http://www.imi.europa.eu).

harmonization targets. The framework enables subject matter experts to take ownership of data harmonization and allows custom, may-to-many mappings between variables resulting in a tree structure of dependent variables (a dependency tree). The variable (protocol) description and mapping code are encapsulated in the tree nodes, called knowledge objects. As more research questions make use of the framework, a library is constructed that increases the likelihood of re-use of a knowledge object. Access to the data is controlled by the data source custodians and security and provenance information is propagated down the dependency tree to the analysis variables.

3 Architecture

To ensure privacy protection, data is kept at the source; exploratory applications that only return aggregated results, federate their queries. For the analysis of patient level data, anonymized data is transferred to a secure environment after approval by data custodians. To ensure outflow tools present measurements consistently across data sources, target knowledge objects are managed centrally and are imported by each source. The EMIF scientific community is responsible for the maintenance of the content of the central library. The library contains target knowledge object definitions and mappings only, not actual data. Data extraction and mappings to source variables are defined and maintained at the source.

Semantic web technology is used to implement the full data flow. An ontology was developed to represent clinical data. Annotation properties and (public) vocabularies are used to define knowledge objects semantically. Rules (code snippets) implement the mappings and access restrictions. The range of required statistical calculations in the mappings outweigh the capabilities of reasoners (although in practice most mappings can be implemented in SPARQL rules). To support languages such as R for complex statistical modelling, a script recursively follows dependencies and executes code snippets using the appropriate engine. The script is implemented using a graphical ETL tool (Pentaho) that was adapted to work with a triple store (Stardog). Harmonized local data is stored in the local triple store that presents a query endpoint to the outside world.

4 Results

The harmonization method described above was applied to the EMIF Multimodality Biomarker Discovery (MBD) Study. This study investigated AD progression in two balanced subgroups (amyloid positive vs. negative) using historical subject level data. 11 cohorts agreed to participate and contributed data for a total of 1221 subjects (676 amyloid negative, 545 amyloid positive). A combined data set was constructed of 324 predefined variables, containing data for 3399 patient encounters. 4 cohorts of the MBD Study and 3 supplementary cohorts provided clinical data for additional subjects. A single, virtual cohort of 3358 subjects from 14 source cohorts was composed. The size of the library of harmonized variables increased to 380. Data cover up to 17 visits, the total number of patient encounters is 10086.