Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200425

Knowledge Driven Phenotyping

Honghan WU [a,1], Minhong WANG [a], Qianyi ZENG [a], Wenjun CHEN [a], Thomas NIND [a], Emily JEFFERSON [a], Marion BENNIE [a], Corri BLACK [a], Jeff Z. PAN [a], Cathie SUDLOW [a] and Dave ROBERTSON [a]

[a] *Working Group of Graph-Based Data Federation for Healthcare Data Science (Sprint Exemplar Project funded by Health Data Research, United Kingdom)*

Abstract. Extracting patient phenotypes from routinely collected health data (such as Electronic Health Records) requires translating clinically-sound phenotype definitions into queries/computations executable on the underlying data sources by clinical researchers. This requires significant knowledge and skills to deal with heterogeneous and often *imperfect* data. Translations are time-consuming, error-prone and, most importantly, hard to share and reproduce across different settings. This paper proposes a knowledge driven framework that (1) decouples the specification of phenotype semantics from underlying data sources; (2) can automatically populate and conduct phenotype computations on heterogeneous data spaces. We report preliminary results of deploying this framework on five Scottish health datasets.

Keywords. health data, phenotype computation, data integration, ontology

1. Introduction

Health data in the UK is stored in different local communities, meaning they are maintained locally and stored in inconsistent formats and languages. A key technical challenge haunting almost all clinical studies is to extract or compute accurate patients' phenotypes (traits of symptoms, diseases, medications or biochemistry test results) from such a fragmented data space. Specifying the computations of a phenotype requires (each time) significant human effort to understand database details, good data science skills to do the querying and data manipulating, and caution & patience to deal with data incompleteness/inconsistencies. Such phenotype specifications can hardly be reused as the consequence of the heterogeneous data models across jurisdictions. This significantly impedes the reusability and reproducibility of clinical researches.

2. Method

The main aim of this study is to realise a clinical data science framework [1] that makes the underlying data sources *transparent* to phenotype computations. The key is to decouple the formalisation of phenotype semantics and the technical details of underlying data sources. We propose an architecture to implement such a decoupling, which is composed of the following two aspects.

[1] Corresponding Author: 9 Little France Road, Edinburgh EH16 4UX, UK; E-mail: honghan.wu@ed.ac.uk.

Phenotype Formalisation Framework This has three components: (1) A database independent phenotype formalisation using Semantic Web knowledge representation technologies to define phenotype semantics; (2) A core phenotype ontology serving as the base vocabulary linking to standard clinical terminologies available at BioPortal (e.g., SNOMED CT, ICD10); (3) A query formatter that generates ontology queries from an user interface. The formatter can automatically translate phenotype definitions between standard terminologies (e.g., READ to SNOMED CT).

Ontology Based & Rule Driven Data Access To *automatically* compute the above formalised phenotypes on actual data, we adopt ontology based data access (OBDA) techniques [2] but with a novel extension to support rules. Such an extension is necessary because the semantics of most phenotypes are not fixed. They either change with research focuses or different researchers might have different opinions about certain rules related to a phenotype. For this reason, in our framework, we minimise the ontology. Instead, a **rule engine** is implemented to allow flexible definitions of phenotypes. The engine can automatically convert user specified rules into new data mappings, which will be translated on the fly to do phenotype computations on databases.

3. Deployment and Evaluation

This study was supported by Health Data Research UK (`https://www.hdruk.ac.uk/projects/graph-based-data-federation-for-healthcare-data-science/`) and the Medical Research Council [grant number MC_PC_18029] as an exemplar to create a federation of distributed health data in Scotland. The above described framework has been deployed on 5 synthetic data sets generated using BadMedicine [3], which represents data/schema characteristics learnt from real data. Due to space limitations, we put the full benchmark and evaluation details on a Github page: `https://github.com/Honghan/KGPhenotyping/tree/master/evaluation`.

4. Conclusion

To overcome *obscure* phenotype computation, which makes experiments difficult to understand and reproduce, we developed a new framework to allow clinical researchers to formalise phenotype semantics independently to the data and, more importantly, in a computer understandable way so that its computation can be automated on the underlying data sources. We implemented a knowledge-driven (based on ontologies and rule languages) approach to define an interlingua in which practitioners can represent the phenotype semantics they want to use and automatically translates this to computations as database queries on participating data sources.

References

[1] Knowledge Driven Phenotyping - an open source Github repository for computing phenotypes across heterogeneous datasets: https://github.com/Honghan/KGPhenotyping.

[2] Xiao, G., Calvanese, D., Kontchakov, R., Lembo, D., Poggi, A., Rosati, R. and Zakharyaschev, M., 2018, July. Ontology-based data access: A survey. IJCAI.

[3] BadMedicine - a synthetic Scottish health data generator: https://github.com/HicServices/BadMedicine