Cloud services for patient cohort identification using I2b2

Kavishwar B. Wagholikar MBBS PhD\textsuperscript{1,2,4}, Shreekanth V. Joshi MS\textsuperscript{3}, Vishal V. Pai Vernekar BE\textsuperscript{3}, Yuri Ostrovsky PhD\textsuperscript{3}, Somnath D. Desai MCA\textsuperscript{3}, Pooja B. Magdum BE\textsuperscript{3}, Sachin B. Wakle BE\textsuperscript{3}, Sheetal Jain BE\textsuperscript{3}, Akshay Zagade BE\textsuperscript{3}, Rahul Patel MS\textsuperscript{4}, Shawn N. Murphy MD PhD\textsuperscript{1,2,4}

\textsuperscript{1}Harvard Medical School, Boston, MA; \textsuperscript{2}Massachusetts General Hospital, Boston, MA; \textsuperscript{3}Persistent Systems, Pune, India; \textsuperscript{4}Partners Healthcare Boston, MA

Introduction

I2b2 has been widely deployed to enable researchers to identify patient cohorts for clinical studies. Over 200 institutions have deployed i2b2 worldwide, and the deployments at several institutions in the United states are connected into networks for federated querying. However, despite the widespread use of the i2b2 platform, there remain substantial challenges for importing EHR data into i2b2 and for querying the data in i2b2.

To address the above shortcomings, building on our previous work we have implemented cloud-based services to augment patient cohort identification on the i2b2-data store. In our previous work we had described a framework to facilitate rapid development of apps on i2b2, which involved a semantic abstraction of the EHR data involving four facets: (1) transforming EHR data into csv files having a simplified schema to facilitate import into i2b2; (2) use of standard vocabularies to create a hierarchical concept catalogue; (3) mapping local codes to those from standard vocabularies; and (4) definition of clinically meaningful ‘derived concepts’ that involve computation on imported i2b2 facts, resulting in generation of “derived facts.” Together these four facets aim to facilitate IT teams to rapidly import EHR data into i2b2, and to create derived facts that are closer to the cognitive concepts of clinicians.

Methods

We have built a graphical user interface (GUI) to enable a knowledge engineer to define the artefacts needed to implement the framework, and for users to construct complex queries using Sankey-like flow diagrams. The GUIs are linked to the cloud for execution and housekeeping of the framework. Furthermore, we present the evaluation of the developed system using the real-world MIMIC III dataset.

Results

For selecting a cohort of Diabetic patients from the MIMIC III dataset, we created Logics for importing relevant data and for defining the concepts required for building a query for diabetes patients. Execution of the system resulted in the import of the diagnosis and labs tables from the MIMIC III data set. The results demonstrate that the Logic construct is useful to capture the knowledge required to import data from an external repository and to apply transformations that facilitate query, and that the query GUI successfully provides the functionality to create a cohort and to visualize the resulting flow diagram.

Concluding remarks and future work

We have implemented a system that allows the knowledge engineer to declare the key components needed for importing EHR data and for transforming it into clinically meaningful concepts that facilitate querying. As the defined components are automatically converted into processes managed by the cloud service to move and transform data, our system significantly reduces the need for manual programming and for managing the compute environment. The systems’s graphical interface allows creation of complex queries and provides a visualization of the results as a flow diagram, which can immensely simplify the identification of complex cohorts. Our future work includes evaluation of the system in a real-world setting and benchmarking the scalability of the system on larger datasets.

Acknowledgements

This work was supported by R00-LM011575, R01-HG009174, Partners HealthCare and Persistent Systems.

References