Integrating TCGA Clinical Data Using Metadata-driven Tools and NLP

Richard C. Kiefer1, Guoqian Jiang, MD, PhD1, Michael K. Davis2, Melissa Castine2, Girish Chavan2, Guergana Savova, PhD3, Rebecca Jacobson, MD, MS2

1Mayo Clinic, Rochester, MN; 2University of Pittsburgh, Pittsburgh, PA; 3Harvard Medical School/Boston Children's Hospital, Boston, MA.

The Integration Problem

Background: Integration of cancer clinical data from EMRs and cancer registries is a major impediment to the development of integrated data repositories (IDRs) for translational cancer research because of the wide range of data models used by these systems.

Objective: We sought to leverage work ongoing in multiple ITCR groups to transcend challenging formats silos that limit integration of The Cancer Genome Atlas (TCGA) clinical text, cancer registry data export files (North American Association of Central Cancer Registries - NAACCR standard based), images and clinical text. A goal of this collaboration was to be able to better integrate TCGA clinical data with molecular data to support translational research projects.

Implementations: This is a collaboration between two NCI ITCR projects: caCDE-QA U01 project, the DeepPhe U24 project and the TIES U24 project. We collaboratively developed a metadata-driven framework to facilitate the ETL process (Figure 1). The caCDE-QA team developed the cancer study common data elements (CDE) RESTful services whereas the DeepPhe team implemented an ETL script generator that consumes the metadata RESTful services (Figure 2) to load TCGA data and Cancer Registry data into tranSMART using the DeepPhe domain ontology to harmonize these datasets. The TIES team developed a TCGA integrated clinical repository that enables users to create cohorts using a combination of NLP features in the TCGA pathology report, as well as TCGA clinical data.

ETL Script Generator

- Metadata-driven ETL tools enable reusable ETL model bindings.
- DeepPhe ontology was sufficient to generate mappings among multiple different sources of data including Cancer Registry, TCGA clinical data, and phenotype extractions emanating from text.
- Tools and approaches developed can be shared and used by translational investigators
- Collaboration between ITCR groups helps diffuse ITCR innovations among groups, helps us develop trans-ITCR use cases that combine tools, facilitates program level achievements.

References

- caCDE-QA: https://github.com/caCDE-QA
- DeepPhe: https://github.com/DeepPhe
- TIES: https://sourceforge.net/projects/caCDE-QA
- Metadata Services: http://informatics.mayo.edu/ITCR

Project Websites:

- caCDE-QA: http://informatics.mayo.edu/caCDE-QA
- DeepPhe: http://cancer.healthnlp.org
- TIES: http://ties.dbmi.pitt.edu/live-demos

© 2016 Mayo Foundation for Medical Education and Research