Admin Supplement Proposal:

Tools for Creation of Integrated Data Repository for Cancer Research

Parent Projects:

DeepPhe (PIs: Savova and Jacobson)

caCDE-QA (PI: Jiang)

Motivation

- Future advances in translational cancer research will be increasingly dependent on the creation of patient cohorts encompassing both highly detailed phenotypic and molecular data.
- Integrated Data Repositories (IDRs) are needed to combine molecular and phenotypic data, making data available with analytic tools.
- This is especially important for cancer investigators who do not have their own computationally-focused labs.
- Extraction, transformation and loading (ETL) of cancer clinical data from EMRs and cancer registries is a major impediment to the development of IDRs for translational cancer research because of the wide range of data models used by these systems

Significance

- This supplement will leverage existing technologies with tools developed in our ITCR projects to create an open-source integrated data repository (IDR) and associated cancer data services, advancing data analytic capabilities at cancer centers.
- The proposed collaboration will develop a common cancer research clinical data model and associated ETL tools to enable cancer research data to be automatically loaded into an IDR along with the associated molecular data.
  - The Cancer Genome Atlas (TCGA) clinical data
  - Cancer registry data (NAACCR standard based)
- The project builds on software already widely used in translational research environments
  - I2b2/tranSMART

Specific Aims (Common)

- **Aim 1: Create methods for loading i2b2/tranSMART repository with TCGA clinical data and cancer registry data.**
  - We will perform the source data extraction and initiate the data curation process, and create syntactic and semantic mappings via standard templates.
  - We will then transform the source data into standard format and load the output into tranSMART.
- **Aim 2: Create methods for building HL7 FHIR-based clinical cancer data services on top of i2b2/tranSMART repository.**
  - We will first create mappings between the i2b2 Star Schema and HL7 FHIR resources.
  - We will then invoke the tranSMART programming APIs to extract clinical cancer data and use open-source FHIR APIs to build FHIR-based messaging web services.

Tasks for Each Project

- **Aim 1: Create methods for loading i2b2/tranSMART repository with TCGA clinical data and cancer registry data.**
  - ***DeepPhe***
    - ***Create methods for clinical cancer data extraction .***
    - ***Create methods for clinical cancer data transformation and loading .***
  - ***caCDE-QA***
    - ***Develop a semantic cancer study metadata repository informed by ISO/IEC11179 standard.***
    - ***Create methods for enhancing tranSMART ETL process using the semantic metadata repository***
- **Aim 2: Create methods for building HL7 FHIR-based clinical cancer data services on top of i2b2/tranSMART repository.**
  - **DeepPhe**
    - **Examine domain-specific template generation tool for creation of a DeepPhe breast cancer phenotype model .**
    - ***Create HL7 FHIR model ontologies in i2b2/tranSMART to support data transformation.***
  - ***caCDE-QA***
    - ***Create tools for supporting creation of FHIR-based clinical cancer data models .***
    - ***Create tools for building HL7 FHIR-based clinical cancer data services.***