Development of an open-source, flexible framework for complex inter-institutional disparate data sharing and collaboration

Development of an open-source, flexible framework for complex inter-institutional disparate data sharing and collaboration

Chaim Kirby[1,2], Peter F. Ambros[3], David Billiter[4], Wendy B. London[5], Eneida Mendonca[6], Tom Monclair[7], Andrew D. J. Pearson[8], Susan L. Cohn[1], Samuel L. Volchenboum[1,2,9]
[1]Department of Pediatrics, [2]Center for Research Informatics, and [9]Computation Institute, University of Chicago, Chicago, IL; [3]Children's Cancer Research Institute, St. Anna Kinderspital, Vienna, Austria; [4]The Research Institute at Nationwide Children's Hospital Center for Childhood Cancer, Columbus, OH; [5]Dana-Farber Cancer Institute/Harvard Cancer Care and Children's Hospital Boston, Boston, MA; [6]University of Wisconsin, Madison, WI; [7]Section for Pediatric Surgery, Division of Surgery, Rikshospitalet University Hospital, Oslo, Norway; [8]Institute of Cancer Research and Royal Marsden Hospital, Sutton, UK

Abstract - Clinical information, "-omic" datasets, and tissue samples are difficult to harmonize and manage for data mining. We have developed a platform for storing clinical research data while providing access to associated data from other information stores. Data on 34 metrics from 11,000 neuroblastoma patients were instantiated into a database. The Django web framework was used to create a model for rapid development of tools and views with a front-end interface for generating complex queries. Working with Nationwide Children's Hospital, we can now consume their tissue inventory data through an API. The end-user sees the number of patients who both match their search and have tissue available. Since initial implementation, the current tasks revolve around developing a governance structure and the necessary data use agreements. Efforts now are to (1) update the data with 5000 more patients, and (2) link to genomic data stores, facilitating disparate data acquisition for research studies.

Background - Clinical information, "-omic" datasets, and tissue samples are becoming more difficult to harmonize and manage for advanced data mining. We believe that clinical research data can be centralized, providing direct access to sample availability and associated data from a variety of information stores. The current process of querying tissue sample availability and finding associated genomic data is time-consuming, cumbersome, and inefficient. Rapid and accurate availability of information on intergrated disparate data sources will provide the researcher with unprecedented access to clinical research data.

Methods - A standardized set of more than 11,000 anonymized patient records was obtained from the International Neuroblastoma Risk Group (INRG). The data are from patients diagnosed between 1974 and 2002 and consist of 34 metrics, such as age at diagnosis, stage of tumor, and other clinical and biological markers. We instantiated the dataset into a PostgreSQL database, and using the Django web framework, created a data model for rapid development of tools and views and built a front-end interface for generating complex queries. To test the feasibility of accessing information on disparate and geographically distinct data samples, we entered into a formal agreement with the Children's Oncology Group Tumor Bank at The Research Institute at Nationwide Children's Hospital. Based on local query results from the clinical data, we can consume the Tumor Bank tissue inventory data through a web-facing application programming interface. We have further expanded the application to provide an easy means for tissue request based on query results. Lastly, we have engaged counsel and are establishing a governance structure to advise on the necessary data use agreements for participating institutions.

Results - We have completed our initial implementation and have collaborative agreements with other international consortium groups. We have created a paradigm for statisticians to securely update and add data, and a verification system checks for internal validity and provides a report of the transaction. Our system can initiate queries and accept results in a variety of standards-compliant formats, and is currently available in demonstration-form. Once the query is performed, the end-user is presented with the number and geographic region of patients who match their search terms and for whom tissue samples are available. The user is finally offered the option of populating a tissue request form to be sent to the INRG.

Discussion: Querying patient data while interrogating external sources allows researchers to observe which ancillary data and samples are available and permits them to quickly download data or generate a request. Current procedures for identifying study cohorts and tissue samples are time consuming and inefficient, often taking weeks to initiate requests. We are currently moving forward with the next phase of development, incorporating genomic data stores into our model. While designed around neuroblastoma, our system can be applied to a variety of clinical scenarios and data sources and will be made available through an open-source license.