ONCO-i2b2 Project: a Bioinformatics Tool to Integrate Biobank Information and Clinical Data to Support Translational Research in Oncology

SEGAGNI D (1), TIBOLLO V (1), DAGLIATI A (3), PERINATI L (1), ZAMBELLI A (1), PRIORI S (1), BELLAZZI R (1, 2)

(1) IRCCS Fondazione S. Maugeri, Pavia, Italy
(2) Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy
(3) Insitute for Advanced Studies, Pavia, Italy

Motivation
The University of Pavia and the IRCCS Fondazione Salvatore Maugeri of Pavia (FSM), thanks to a grant of the Lombardia, has recently started an IT initiative to support clinical research in oncology called ONCO-i2b2. The ONCO-i2b2 project aims at supporting translational research in oncology and exploits the software solutions implemented by the Informatics for Integrating Biology and the Bedside (i2b2) research centre, an initiative funded by the NIH Roadmap National Centres for Biomedical Computing and headed by Partners HealthCare Center in Boston. The ONCO-i2b2 software is designed to integrate the i2b2 infrastructure with the FSM hospital information system (HIS) and with a cancer biobank that manages both plasma and cancer samples, taken with the informed consent of healthy individuals and oncologic patients. The i2b2 infrastructure, integrated with the HIS, provides a web-based access to all the electronic medical records of cancer patients and allow researchers querying and analyzing the vast amount of information coming from the clinical practice, relying on a user-friendly interface.

Methods
Using i2b2 and new software modules purposely designed, data coming from multiple sources are integrated and jointly queried. The integration process is based on an ontology of the problem domain and on open-source software integration modules. The integration process could be described through three main steps. The first one concerns the upload of data stored inside the biobank from the hospital pathology unit. To achieve this goal a semi-automatic procedure has been implemented. Each cancer biospecimen is anonymized, by creating a two-dimensional DataMatrix barcode, and the biobank database is periodically synchronized in order to keep biobank samples data constantly updated. The information on the biological samples contained in a biobank is loaded into the i2b2 data warehouse through a series of Extract, Transform, Load (ETL) operations, performed relying on the KETTLE, a tool developed within the Pentaho project. The second step is designed to make the information collected in the FSM
HIS available to the I2B2 service. An ETL process transforms the medical information of interest in concepts that will be queried afterwards. The majority of the data collected in the FSM HIS is made of textual reports. It was therefore necessary to develop and integrate inside the ICT architecture a Natural Language Processing (NLP) module in order to extract information and clinical tests results from unstructured medical records. In particular, the NLP system, based on the GATE framework, is designed to process the textual report by looking at the SNOMED codes of the samples as well as at a set of numeric data about the tumour sample. The last step deals with managing data within the i2b2 architecture. The i2b2 data warehouse, called Clinical Research Chart (CRC), allows the management of data from clinical trials, medical record systems and laboratory systems, and is structured following the so-called star schema. This architecture is based on the notion of ‘facts’; specifically in healthcare, a logical fact is an observation on a patient. Data information are stored in the i2b2 data warehouse through complex ETL transformations following a cancer-specific ontology that combines atomic information to create a well defined medical observation. The extracted information can be analyzed through the i2b2 web client with appropriate plug-ins specially configured.

Results
Since December 2010 the entire software system has been installed and is currently running at FSM. Up to the present days, the i2b2 instance contains data of 2.214 patients (312 of them have at least one biological sample in the cancer biobank), 25,826 visits, 163 concepts (26 of them related to diagnosis, 24 related to clinical measurements and 31 related to histological reports) and 93,680 observations. In the next steps, we will empower the system implemented to far with a web-service that allows communication between i2b2 and the R statistical software. Relying on such module, we will develop a set of ONCO-i2b2 dedicated plug-in to perform further relevant analysis, with the goal of integrating patient’s genotype data, too.

Contact email
daniele.segagni@fsm.it