*Quality of Life through Quality of Information*
*J. Mantas et al. (Eds.)*
*IOS Press, 2012*

*doi:10.3233/978-1-61499-101-4-1126*

# CARDIO-i2b2: integrating arrhythmogenic disease data in i2b2

Daniele SEGAGNI[a,1], Valentina TIBOLLO[a], Arianna DAGLIATI[b], Carlo NAPOLITANO[a,e], Silvia G. PRIORI[a,d,e], Riccardo BELLAZZI[c]
[a] *IRCCS Fondazione Salvatore Maugeri, Pavia, Italy,* [b] *Istituto Universitario di Studi Superiori (IUSS), Pavia, Italy,* [c] *Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy,* [d] *Dipartimento di Cardiologia, Università di Pavia, Pavia, Italy,* [e] *New York University School of Medicine, New York, USA*

**Abstract.** The CARDIO-i2b2 project is an initiative to customize the i2b2 bioinformatics tool with the aim to integrate clinical and research data in order to support translational research in cardiology. In this work we describe the implementation and the customization of i2b2 to manage the data of arrhytmogenic disease patients collected at the Fondazione Salvatore Maugeri of Pavia in a joint project with the NYU Langone Medical Center (New York, USA). The i2b2 clinical research chart data warehouse is populated with the data obtained by the research database called TRIAD. The research infrastructure is extended by the development of new plug-ins for the i2b2 web client application able to properly select and export phenotypic data and to perform data analysis.

**Keywords. 3-5** i2b2, arrhythmogenic disease, research database, data analysis

## Introduction

The CARDIO-i2b2 project supports the implementation and the deployment of the bioinformatics platform, developed by the “Informatics for Integrating Biology and the Bedside” (i2b2) [1,2] research center, in the University hospitals of Pavia, Italy. One of these hospitals, the IRCCS Fondazione S. Maugeri (FSM), is currently exploiting the i2b2 software to empower research activities in the cardiology and oncology areas. In particular, several efforts are ongoing in molecular cardiology research by integrating the i2b2 software with the Transatlantic Registry of Inherited Arrhythmogenic Diseases (TRIAD). TRIAD is a joint project of the Molecular cardiology department of the FSM and the Langone Medical Center of the New York University (NYUMC).

## 1. Methods

CARDIO-i2b2 integrates data coming from multiple sources and allows the users to jointly query them. The collected data are then stored in the i2b2 data warehouse, where facts are hierarchically structured as ontologies [3]. A web client application allows researchers to query the data collected through a user friendly interface.

[1] Corresponding Author: Daniele Segagni, email: daniele.segagni@fsm.it

CARDIO-i2b2 gathers data from the Molecular Cardiology Labs (MCL) databases and merges them with the clinical information from the TRIAD system.

TRIAD is an electronic medical record (EMR) created to collect data related to specific arrhythmogenic diseases like Brugarda syndrome, long QT syndrome, short QT syndrome, catecholaminergic ventricular tachycardia and arrhythmogenic right ventricular cardiomyopathy. In this database researchers and physicians collect information related to patients diseases, therapies, holter and ECG measurements, mutations, implanted devices information (for examples pacemakers or cardioverter defibrillator) and cardiac events such as cardiac arrests or syncopes.

There are two TRIAD instances currently activated: one inside the FSM molecular cardiology labs and another one at the NYUMC labs.

Genetics information related to affected patients are also collected, the main part of these data are related to specific gene mutations. MCL biologists collect blood samples for each patient and annotate the sequenced gene through specific software. In the system there are also stored information relating to the family of patients to reconstruct the family tree of the disease.

Our main effort was to provide a robust integrated research environment, giving a particular emphasis to the integration process and facing different challenges, consecutively listed: biospecimen samples privacy and anonymization [4]; synchronization of the TRIAD database with the i2b2 data warehouse through a series of Extract, Transform, Load operations [5].

## 2. Results

We exported the data contained into the TRIAD relational database and populated every i2b2 dimension table of its warehouse architecture concerning patients, visits and observations data. Currently, a total of 591 patients, 13987 visits, 367 concepts and 262512 observations have been exported in the i2b2 data warehouse. The observation set, in particular, is a unification of observations used for stratify the whole patient set and observations that represents specific measure values.

With the aim to be more helpful for the researchers during the data analysis and patients stratification, we developed new web plug-ins that exploits the data selected with the i2b2 query engine [6].

To allow researchers to dynamically perform survival analysis on selected patient sets, a dedicated i2b2-cell [7] has been developed in order to include the R statistical software inside the IT architecture [8]. We created a novel server side engine (called R Engine Cell that allows the communication between the i2b2 architecture and the R software. As survival analyses are routinely performed by cardiology researchers at FSM, firstly we have concentrated on making Kaplan Meier analyses available within the i2b2 web interface. To this aim, we have developed a web-client plug-in that allows users to easily select the patient set on which to perform the analysis and displays the results in a graphical, intuitive way.

Furthermore we developed a data export plug-in that allows users to select a specific patient set and export several values of interest into an Excel file. An XML configuration file correlates each project with the concepts that can be exported.

## 3. Discussion

CARDIO-i2b2 is a concrete example of an integrated ICT architecture implemented to support translational research.

Thanks to a well-integrated customization of the i2b2 software suite, it is now possible to analyze the data coming from a large database of cardiac diseases by using a single and unique web-based product which is able to support researchers from concepts selection to the data export and statistical analysis activities.

To reach this goal, we have faced several challenges. One of the most interesting task was the creation of dedicated ETL transformations to populate the i2b2 data warehouse with concepts related to cardiology research. Each applied ETL process is a reusable component that can be scheduled to perform different data transformation jobs, once the system will have to deal with new data coming from the NYUMC labs.

We also plan to continue extending the capabilities of the CARDIO-i2b2 architecture by implementing new plug-in devoted to data analysis; in particular, we are working on an extension of the i2b2 query engine by adding temporal query capabilities.

Following the example of the experience reported in [9], the future developments of the project will regard the integration of patient's genotype data, including Next Generation Sequencing derived information, which will require careful evaluation both in terms of the data representation and storage and of data security and privacy.

## References

[1] Murphy SN, Mendis M, Hackett K et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc. 2007, 548-52.

[2] https://www.i2b2.org (last accessed 30th January 2012)

[3] Adamusiak T, Burdett T, Kurbatova N, Joeri van der Velde K, Abeygunawardena N, Antonakaki D, Kapushesky M, Parkinson H, Swertz MA: OntoCAT-simple ontology search and integration in Java, R and REST/JavaScript. BMC Bioinformatics. 2011 May 29, 12:218.

[4] Malin B, Loukides G, Benitez K, Clayton EW. Identifiability in biobanks: models, measures, and mitigation strategies. Hum Genet. 2011Sep, 130(3):383-92

[5] R. Bouman and J. van Dongen. Pentaho Solutions. Wiley; 2009.

[6] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr, 17(2):124-30.

[7] Mate S, Bürkle T, Köpcke F, Breil B, Wullich B, Dugas M, Prokosch HU, Ganslandt T. Populating the i2b2 database with heterogeneous EMR data: a ‚semantic network approach. Stud Health Technol Inform. 2011, 169:502-6.

[8] Segagni D, Ferrazzi F, Larizza C, Tibollo V, Napolitano C, Priori SG, Bellazzi R. R engine cell: integrating R into the i2b2 software infrastructure. J Am Med Inform Assoc. 2011 May 1, 18(3):314-7.

[9] Lori C. Phillips et al. Use Genomic Variants in Informatics for Integrating Biology and the Bedside (i2b2). In AMIA Annu Symp Proc: 48-52 T2011