Exploring Complexity in Health: An Interdisciplinary Systems Approach
A. Hoerbst et al. (Eds.)

doi:10.3233/978-1-61499-678-1-572

Beyond Cohort Selection: An Analytics-Enabled i2b2

Matteo GABETTA[a,b,1], Alberto MALOVINI[c], Mauro BUCALO[a], Elisa ZINI[b], Valentina TIBOLLO[c], Silvia G PRIORI[c], Simone VETTORETTI[d], Cristiana LARIZZA[b], Riccardo BELLAZZI[b] and Nicola BARBARINI[a]
[a]*BIOMERIS s.r.l., Pavia, Italy*
[b] *Center for Health Technologies, Università di Pavia, Pavia, Italy*
[c]*Fondazione IRCCS S. Maugeri, Pavia, Italy*
[d]*Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy*

Abstract. The i2b2 software is a widely adopted solution for secondary use of clinical data for clinical research, specifically designed for cohort identification. i2b2 is still lacking functionalities for data analysis. The aim of this work is to empower the i2b2 framework enabling clinical researchers to perform statistical analyses for accelerating the process of hypothesis testing. To this aim we have developed a flexible extension of i2b2 able to exploit different statistical engines. We have implemented some first applications for basic statistics and survival analyses, exploiting this extension and accessible through suitable user interfaces designed with a special consideration for usability.

Keywords. i2b2, data analysis, statistics, data warehouse

1. Introduction

The i2b2 software has proven, over the years, to be an effective and widely adopted solution to reuse clinical data for research purposes [1,2]. Specifically, i2b2 helps in identifying cohorts of de-identified patients meeting certain clinical criteria. To date in order to perform even simple statistical analyses on a cohort selected with i2b2, a clinical researcher needs to extract the interested data from i2b2 and then to run data analysis with other statistical tools, often with the support of statisticians. In the past, efforts have been made to address this issue but, although serviceable, the solutions proposed lacked in flexibility and usability [3]. Other knowledge management platforms for translational medicine have been developed through the years; for example, TranSMART extends the i2b2 data model providing advanced analytics facilities, which do require some statistics expertise to be exploited fruitfully [4]. The aim of the work is to equip i2b2 with analytics capabilities through a novel extension flexible enough to be interfaced with different statistical engines and to execute a wide range of algorithms (like R packages [5]), exploiting a common framework to access the data stored within the i2b2 data warehouse (DW). A further objective is to provide graphical user interfaces (i2b2 Webclient plugins) enabling clinicians with basic statistical and informatics skills to perform fast hypothesis testing within the i2b2 environment.

[1] Corresponding Author: matteo.gabetta@biomeris.com

2. Methods

2.1. System Architecture

We have realized an extension of the i2b2 framework composed by three main components (Fig. 1): (i) the Analytics Cell is a web-service which receives a request from the GUI, retrieves the data from the i2b2 database, executes the analysis on the Analytics Engine and returns the results to the GUI; (ii) the Analytics Engine, performing the actual analysis, is a Java API that runs inside the cell or provides connection to an external computation engine (e.g. Rserve); (iii) the Graphical User Interface (GUI), i.e. the i2b2 Webclient plugins, allows to set the input parameters for the specific analysis and to visualize and export the outcomes in a suitable way.

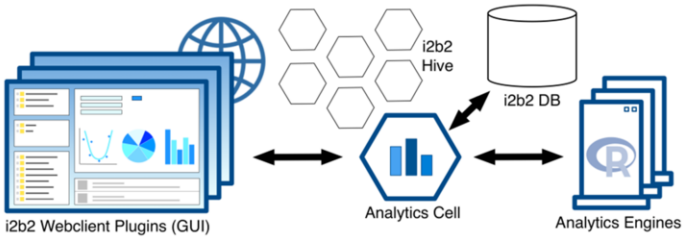

Figure 1. Overall architecture of the i2b2 extension for analytics.

2.1.1. The Analytics Cell

The Analytics Cell, like the other cells belonging to i2b2 Hive, addresses input/output communication by means of the i2b2 messaging standard; yet our choice was not to extend this model, but to use its "observation blob" field to contain all the exchanged information. We have chosen the JSON to represent this information since it facilitates the communication with the GUI, written is JavaScript language.

We have also developed a session-aware client/server communication model in four main steps in order to manage multiple potentially complex operations at the same time and to store raw data within the cell to speed up consecutive analyses (Fig.2):

1. The handshaking call associates the client/server interaction with a new session.
2. The submission call, after the user has defined all the analysis parameters on the GUI, sends them to the cell where the real execution starts. The cell, after returning the identification number of the requested analysis:
   A. Downloads the data directly from the i2b2 database. We chose to avoid the use of i2b2 CRC service to speed up the process for large amounts of data.
   B. Once the data are available, the cell, depending on the analysis to be carried out and on the analytics engine on which this has been implemented, adapts them to the input format of the algorithm and starts the analysis. When the analytics engine returns the results, these are processed in the cell in order to create data structures that will be returned to the plugins.
3. The polling call, performed periodically, updates the client about the status (e.g. *working*, *done*, *error*) of a specific analysis running on the server.

4. Once an analysis is completed, the download call asks for the results of the analysis that are finally processed by the GUI in order to be presented to the user. The results of each analysis persist in the cell as long as the session lasts.

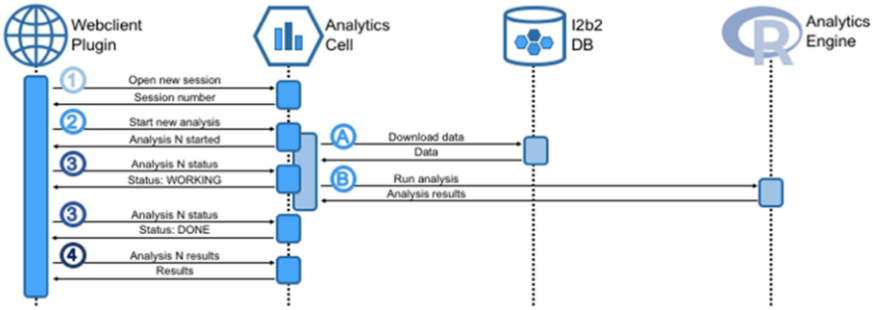

Figure 2. Sequence diagram of the communication model between the components of the architecture.

2.1.2. The Analytics Engine

The communication between the cell and the specific algorithm takes advantage of the wide range of Java libraries available for this purpose and is, of course, dependent on the analytics engine and by the algorithm itself.

To date, the analyses integrated into the cell exploit R project as analytics engine (see "Applications"). R project is chosen since it is one of the reference instrument for statistical data analysis, implementing several analytical packages. The communication between the cell and R is via the "Rserve" package and the associated Java library.

2.1.3. The Graphical User Interface

The GUI, implemented as a set of extensions (i.e. plugins) of the i2b2 Webclient, is implemented by using JQuery for DOM manipulation and JavaScript libraries to draw graphs. Each implemented application of the cell needs a dedicated plugin.

The common principle which the plugins are built on, is providing usable tools enabling intuitive and dynamic interaction with the user.

2.2. Applications

So far we have implemented three applications to perform specific types of analyses, accessible from devoted plugins (available at i2b2.biomeris.com). These applications exploit customized R scripts accessed via Rserve and returning respectively:

- *summary statistics* for a patient set on the observations belonging to a single concept/variable. Categorical distributions are described by counts and frequencies, continuous distributions are described by mean ± standard deviation or median (25^{th} – 75^{th} percentile) if deviating from the normality assumptions. Suitable plots to graphically represent the variables distribution are also generated.
- *results from statistical tests* between two variables for a patient set, by automatically applying appropriate analytical methods (t-test, Wilcoxon, Kruskal-Wallis, Fisher test, etc.) showing also some useful plots (box-plot, heat-map and scatterplot). As an example, Fig. 3 reports the results from the analysis of the correlation between two continuous variables: in this case the

Spearman's rank correlation test is applied since one of the two variables deviates from normality and the deriving rho and p-values are reported. The correlation between the two variables is graphically represented by scatterplots, box plots and heat map tables.

- *survival curves and related statistics*, estimated by the Kaplan-Meier method, starting from a patient set and a specific concept representing the events of interest.

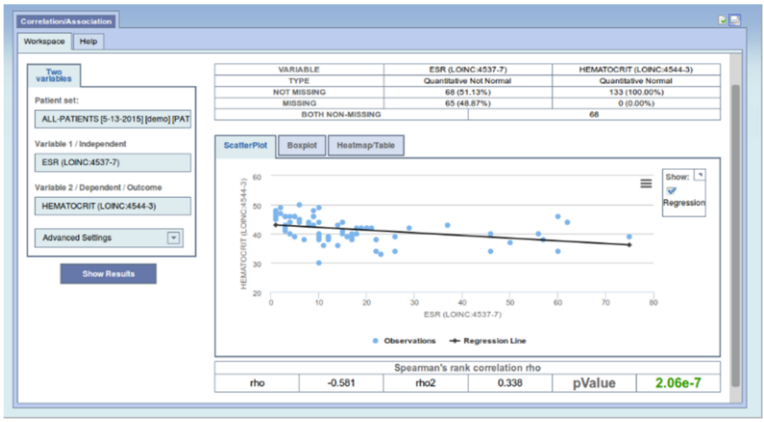

Figure 3. Correlation plugin representing the correlation between two numeric concepts.

3. Results

3.1. Performance test

In order to test the performance of each application we simulated the DW of two different research organizations that could benefit from the use of statistical features in i2b2: (a) a medium organization counting around 150.000 patients and more than 100 million observations; (b) a large organization counting more than 1.000.000 patients and around 800 million observations. Each dataset is created by properly replicating the standard "demo" dataset coming along with i2b2.

Table 1. Execution times achieved by the three applications run on two simulated data warehouse (DW) each on two subsets of patients (pat.). The results calculated as the mean execution time in 5 runs.

	Medium DW		***Large DW***	
	1.5%(2448 pat.)	**10%(15912 pat.)**	**1.5%(17136 pat.)**	**10%(111384 pat.)**
Summary statistics	1.6 sec.	3.8 sec.	3.6 sec.	18 sec.
Correlation/association	1.8 sec.	4.7 sec.	5.2 sec.	23 sec.
Survival analysis	1.5 sec.	4.8 sec.	4.8 sec.	25 sec.

All tests have been performed on Amazon Cloud resources, by deploying each DW on a different architecture and keeping the costs proportional to the number of stored observations. The medium organization was simulated on a *c4.xlarge* EC2 server with a *db.m4.large* database; the large organization was simulated on a *m4.2xlarge* server with a *db.r3.4xlarge* database. Each application has been tested on two subsets composed by 1,5% and 10% of the entire population; these ratios are chosen to fit typical cohort selection tasks in a DW. The results of the performance test (Table 1) are

encouraging since, even on large datasets, the execution times remain acceptable to support research activity.

3.2. Usability test

Usability has been evaluated with a System Usability Scale (SUS) questionnaire [6]. A group of 10 clinical researchers with a wide range of statistical and informatics skills tested all the three implemented applications. Each researcher, before evaluating the applications, is asked to complete some typical and frequent analytics tasks: given a certain patient set, (i) find summary statistics of a variable; (ii) calculate the correlation p-value between two variables; (iii) show the survival curve of cohort given certain events of interest. The mean of normalized SUS scores is 79/100, which demonstrates the good usablity of the system. Moreover the results show that differences in the statistical and informatics skills of the users do not impact on the usability of the applications. Recently the analytics extension has been deployed and is in use on real data within two research institutes:

- Fondazione IRCCS Salvatore Maugeri - Pavia (Italy), deployed on an i2b2 project containing around 53.000 patients and more than 7 million observations;
- Fondazione IRCCS Ca' Granda - Ospedale Maggiore Policlinico - Milan (Italy), deployed on an i2b2 project collecting in just one year around 10 million observations for more than 150.000 patients.

4. Conclusions

The implemented i2b2 extension empowers i2b2 with analytics capabilities which are crucial to provide clinical researchers with an effective tool for accelerating the process of hypothesis testing. In fact the reported results show good performance in terms of execution time and high usability scores. Moreover the analytics extension has been already succefully deployed within two research institutes. The source code of the analytics extension is in the process of being shared with the i2b2 developer community, in order to garantuee further tests and deployments on the current i2b2 instances worldwide, as well as to be further enriched with more analytics applications.

References

[1] I.S. Kohane, S.E. Churchill, S.N. Murphy. A Translational Engine at the National Scale: Informatics for Integrating Biology and the Bedside. *JAMIA*,**19.2** (2011): 181-85.

[2] H.U. Prokosch, M. Ries, A. Beyer, M. Schwenk,C. Seggewies, F. Köpcke, S. Mate, M. Martin, B. Bärthlein, MW. Beckmann, M. Stürzl, R. Croner, B. Wullich, T. Ganslandt, T. Bürkle. IT infrastructure components to support clinical care and translational research projects in a comprehensive cancer center. *Stud Health Technol Inform*, **169** (2011):892-6.

[3] D. Segagni, F. Ferrazzi, C. Larizza, V. Tibollo, C. Napolitano, S.G. Priori, R. Bellazzi. R engine cell: integrating R into the i2b2 software infrastructure. *JAMIA*, **18.3** (2011): 314-317.

[4] B.D. Athey, M. Braxenthaler, M. Haas, Y. Guo. tranSMART: an open source and community-driven informatics and data sharing platform for clinical and translational research. *AMIA Summits on Translational Science Proceedings* **2013**: 6.

[5] R Development Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. 2013.

[6] J. Brooke. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, **189.194** (1996): 4-7.