

# INTEGRATION OF CLINICAL AND GENETIC DATA IN THE I2B2 ARCHITECTURE – ILLUSTRATED USE CASE



Shawn N. Murphy<sup>1</sup>, Henry C. Chueh<sup>1</sup>, David A. Berkowicz<sup>1</sup>, Michael E. Mendis<sup>1</sup>, John P. Glaser<sup>2</sup>, Isaac S. Kohane<sup>3</sup>

<sup>1</sup>. Laboratory of Computer Science, Massachusetts General Hospital and Harvard Medical School, Boston MA; <sup>2</sup>. Partners Healthcare System, Boston, MA; <sup>3</sup>. Informatics Program, Children's Hospital, Boston MA

## ABSTRACT

One of the goals of i2b2 is to provide clinical investigators broadly with the software tools necessary to collect and manage project-related clinical research data in the genomics age as a cohesive entity – a software suite to construct and manage the modern clinical research chart. The i2b2 team is developing an interoperable software framework for the research chart that can be extended for new and unanticipated data types as well as functionality. It is intended to serve the following users:

- Clinical investigators who want to use the software in as "shrink-wrapped" a way as possible,
- Bioinformatics scientists who want the ability to customize the flow of data and interactions, and
- Biocomputational software developers who want to develop new software capabilities that can be integrated easily into the computing environment.

One method for developing this new software and framework is to work hand and hand with current researchers to access their current and future needs. Our Asthma "Driving Biology Project" studies the interplay between environmental exposures and genetic variation in determining both individual airways disease risk and individual response to airways disease medications. Tools and methods currently available were successful for monogenic disorders, but have not yielded major breakthroughs in complex diseases to date. This work will lead to the development and implementation of methods and tools to improve genetic epidemiological and pharmacogenetic research in complex diseases.

In order to run queries and perform data mining to determine the differential affects of environmental exposures such as smoking on asthma patients, the Asthma DBP created a data mart containing 71 million clinical observations from over 97 thousand patients. Research specific data not routinely available in clinical data sets was loaded by running it through a set of web services. These services could be located within the local network or at a different remote network. Each of the services is either exclusively a web service or a web service wrapper is developed to encumber the native service. Because each of the i2b2 web services expect and produce data that adheres to specific xml schema, they can be interconnected in multiple different ways.

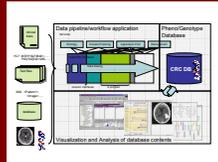
Supported by grant U54LM008748.

## The i2b2 Hive

An i2b2 hive is a collection of independently acting applications, referred to as the "cells" of the hive. They communicate with each other through a standard set of web protocols. The hive can be joined by any application that knows how to communicate through these standard communications protocols. For use cases that require a clinical researcher to work with data from patients or research subjects, a minimum number of core cells are required. These are illustrated in dark blue on the picture to the left. Other cells are added as greater functionality is required from the hive. The use case of this poster requires the colored cells.



## How One uses the Hive to do Clinical Research



The task of any investigator working on a clinical research project is to collect data on a patient population, to organize that data into a chart, and then to analyze that data using various statistical methods in order to further understand that patient population. Usually the investigator will have a specific hypothesis that drives the data collection, but that is not always the case. Even when there is a specific hypothesis, the results of an investigation may be negative, but insights from the data collected may lead to further hypotheses that will eventually be proven. Thus, clinical research is often an iterative process.

The process supported by the i2b2 hive is essentially this cycle of collecting data, organizing it, and then performing an analysis, but specifically focused on the clinical domain.

Illustrated above is the general plan for data moving into the hive. The data begins in some clinical or research system which is illustrated on the left, from which it will need to be cleaned, transformed, and loaded into the central database of the hive, named the Clinical Research Chart (CRC). Often the process of transforming the data can be complex and the workflow cell will need to put multiple applications together in order to achieve this task. This creates a data pipeline using various cells of the hive. Once the data is organized in the CRC, it can be analyzed by tools that have been created for, or transformed into cells of the hive, cells that understand the communication protocols into and out of the CRC.

Illustrated in this poster are specific examples of cells that are being developed within the i2b2 hive, and how they interconnect in a specific use-case where investigators are exploring the confusing effects of how smoking affects the severity of a patient's asthma, as measured by the number of hospital visits per patient.

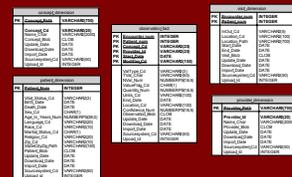
## Adding data to the Hive



Data is placed into the hive either from data collected as part of a clinical research study or from data that exists in a clinic/hospital medical record system. Core tools form part of the data repository (CRC) cell to facilitate this process, typically moving data in HL7 message formats, data available in tabular formats, and data available through database connections into the hive. Ultimately these tools put all data into a common XML schema as it moves through the hive and into the CRC. New cells can be built to extract, transform, and load new types of data. They will need to put the data into the XML schema that allows it to be bulk loaded into the CRC.

Illustrated in the figure above is the design of a workflow that uses the Kepler workflow system (available at <http://kepler-project.org>) as part of the workflow cell. In this example, the investigator has designed a system to import patient clinic notes that they have collected from the medical record system and load them into the CRC.

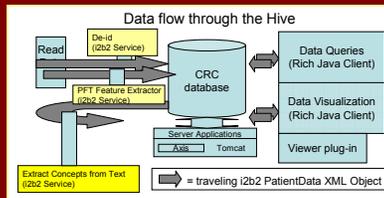
## The Center of the Hive - the Clinical Research Chart



The database that underlies the CRC is organized to accept many different kinds of data into a common format. The design of this database has been time-tested at Partners Healthcare for the past 7 years, serving as the schema of the Research Patient Data Registry, a large production-grade data warehouse with over 700 million rows of data on 2.7 million patients. The database is organized in a "Star Schema" – a popular approach to data warehousing. Fitting the data into a common schema allows queries and analyses that are built for one kind of data to be applied to others, unifying the concepts within the data, and allowing phenotypic data to be integrated with genotypic data.

In the CRC database, an investigator will have the note text stored by the patient number under the "note" concept code.

## Transforming Data in the Hive



Data that is in the CRC database schema can be sent to other cells that are interconnected through workflow applications. The interconnection of cells is enabled by the common i2b2 XML object that flows from the CRC database schema. Data from the medical record may require several different kinds of manipulation before it is useful for research, such as natural language processing (NLP) or the normalization of values. These manipulations are managed through the workflow cell that sends the data to various independent i2b2 cells that do the actual computational work.

In the use case on which we are focusing, the investigator wishes to query those patients who are smokers. But the concept of "smoker" has not yet been coded in the medical record system and will need to be extracted from the notes. The investigator will use an NLP cell to extract the following smoking concepts from the narrative text, "Current Smoker", "Past Smoker", "Never smoked", and "Denies smoking".

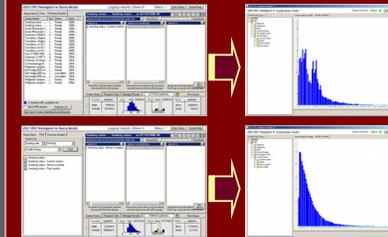
## Querying and Visualizing Data in the Hive

Data from medical record systems can present many challenges to the clinical researcher towards understanding how the data is organized and what is available. Furthermore, the data can often be conflicting within itself. Time stamps on the data can be non-intuitive as well, since they often represent when the data was collected rather than when an event actually occurred. One of the cells of the hive allows the data to be viewed on a timeline within different concept categories for each patient.



Illustrated to the right is the timeline display of patient data available through the i2b2 Navigator. The investigator is interested in exploring the data that was just calculated from natural language processing (NLP) in the last step to see if there are any obvious omissions or errors. What the investigator finds is that some of the patients are classified both to be "Current smokers" and to have "Never smoked". Although this may have been unexpected by the investigator, these kinds of discrepancies are fairly common in patient notes, and also may be errors of the NLP program itself. However, now that the investigator is aware of the issues, they can either choose to reprocess the notes, collect the data in a different way, or find a way to use the data as it has been processed.

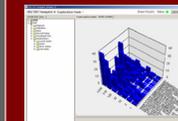
One of the most important functions of the CRC is to allow queries to be performed against it. Various query interfaces can be constructed to accommodate the different levels of database language skills that are expected from our users. However, several query paradigms have proven to be very popular with our users that lend themselves to fairly intuitive interfaces. One of the most popular query interfaces is shown on the left below and consists of coded concepts on the left, and a Venn-diagram-like set of panels on the right. This allows a set of patients to be specified, that can then be further analyzed using standard methods.



The investigator chooses to use the data as it has been processed by the NLP program, but to compensate for the issues discovered in the data. They do this by asking for a set of patients that have one attribute, but do not have the contradicting attribute. Queries are done to get a set of patients that smoke, shown at the top on the left, and a set that do not smoke, shown at the bottom on the left. The number of patients is then graphed for each set on the Y axis, with the number of visits on the X axis. Surprisingly, the graphs show that smokers and non-smokers have similar distributions of numbers of visits. However, the smokers (graph on the top) appear to be divided into two sub-groups, one group that appears to have more than the average numbers of visits, and one that may have less.

## Taking the Data to the Next Step with Correlation Analysis

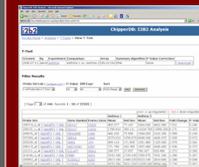
The distribution of the incidence of disease, medication use, adverse events, laboratory test values, or genomic expression within a patient population can provide valuable information on that set of patients. Even more value can be added by comparing the correlations between these incidences amongst each other in the data, such as how many people taking various kinds of medications are experiencing cardiovascular events. Another approach is to correlate the incidences among two groups of patients, such as what diseases are common in one group of patients but not the other. For example, if one compares the diseases that smokers and non-smokers get, one will find a higher incidence of small cell lung cancer in the smoker group. Although finding such correlations may appear to be just another way of performing many Chi Square or Mutual Information calculations at once, medical records present unique problems, because of the sparsity or absence of data that tends to place too much power in the dual-negative bins.



The phenotypic data of the two groups of smokers are compared by the investigator, those with the high number of patient visits, and those with the low number of patient visits. The phenotypes of the two populations are correlated to see if there are any differences in diseases, metabolic state (through laboratory test values) or any phenotypic variant, that can account for the differences in the number of visits between the two groups. As the numbers along the Y axis increase, so do the sophistication of the methods used for the calculation. The illustrations show that as the sophistications of the methods increase, some differences in disease incidence do become apparent in the two populations.

## Taking advantage of Genomic Analytic Resources

One of the most powerful aspects of the i2b2 hive is the ability to incorporate available standard analysis of the data by connecting to resources through the i2b2 web protocols. Illustrated are two hive cells that are used to analyze and annotated genomic data. The ChipperDB cell, illustrated on the left, can be used to perform T-tests on probe IDs from different patient sets to determine differences in expression between the two populations, similar to the correlation analysis above, but without the complexities of information from the medical record. The Genopia cell, illustrated on the right, can be used to explore and annotate various attributes present in some of the probe IDs, for example to determine what genes are represented by the probe ID oligomers and the role those genes may play in a specific protein's expression.



Cells for calculating differences in the expression profiles of various probe IDs on microarrays from patients in the different sets were used by the investigator to find specific genes that were differentially expressed in the two patient populations using the ChipperDB cell shown on the left. The characteristics of this set of genes may then be further investigated with the Genopia cell shown on the right.



The methods described in this poster serve to illustrate one of the many typical use cases around which the functionality of the i2b2 hive is built. The data do not represent actual results.