Clinical and research data integration: the i2b2 FSM experience

Laboratory of Biomedical Informatics for Clinical Research
Fondazione Salvatore Maugeri - FSM - Hospital, Pavia, italy

Laboratory of Biomedical Informatics "Mario Stefanelli"
University of Pavia, Pavia, Italy

Daniele Segagni, MS, Biomedical Engineer
daniele.segagni@fsm.it

FSM institutes

18 FSM institutes in Italy

3 i2b2 project:

- Pavia: Oncology data & Cardiology data
- Puglia: Administrative data
- Sources and CRC in Pavia

The Hospitals network

I2b2 instances

Oncology

Active since	Patients	Visits	Observations	Genetic data	NLP	Plug-ins	Cells
2010	28.838	142.464	2.341.771	Y	y	6	1

Sub-project: biobank

Active since	Patients	Visits	Observations	Genetic data	NLP	Plug-ins	Cells
2011	786	-	6.855	-	y	1	0

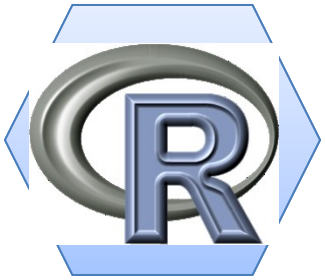
Cardiology

Active since	Patients	Visits	Observations	Genetic	NLP	Plug-ins	Cells
2009	6.334	15.094	205.418	y	n	5	1

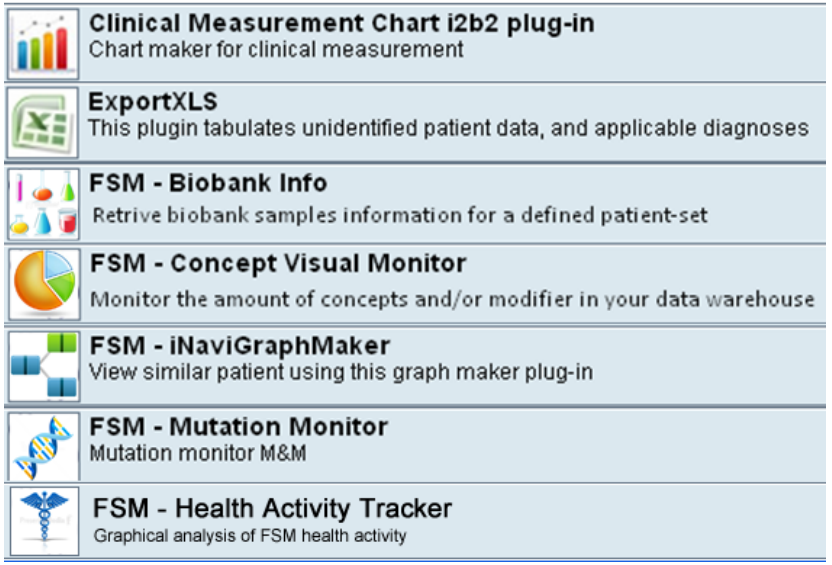
Adminstration

Active since	Patients	Visits	Observations	Genetic	NLP	Plug-ins	Cells
2011	5.611	7.726	23.175	n	n	2	0

I2b2 development



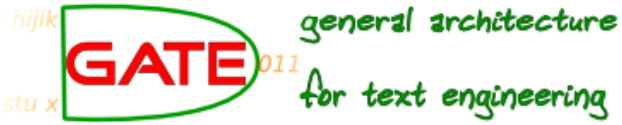
7 web plug-in



1 server side cell

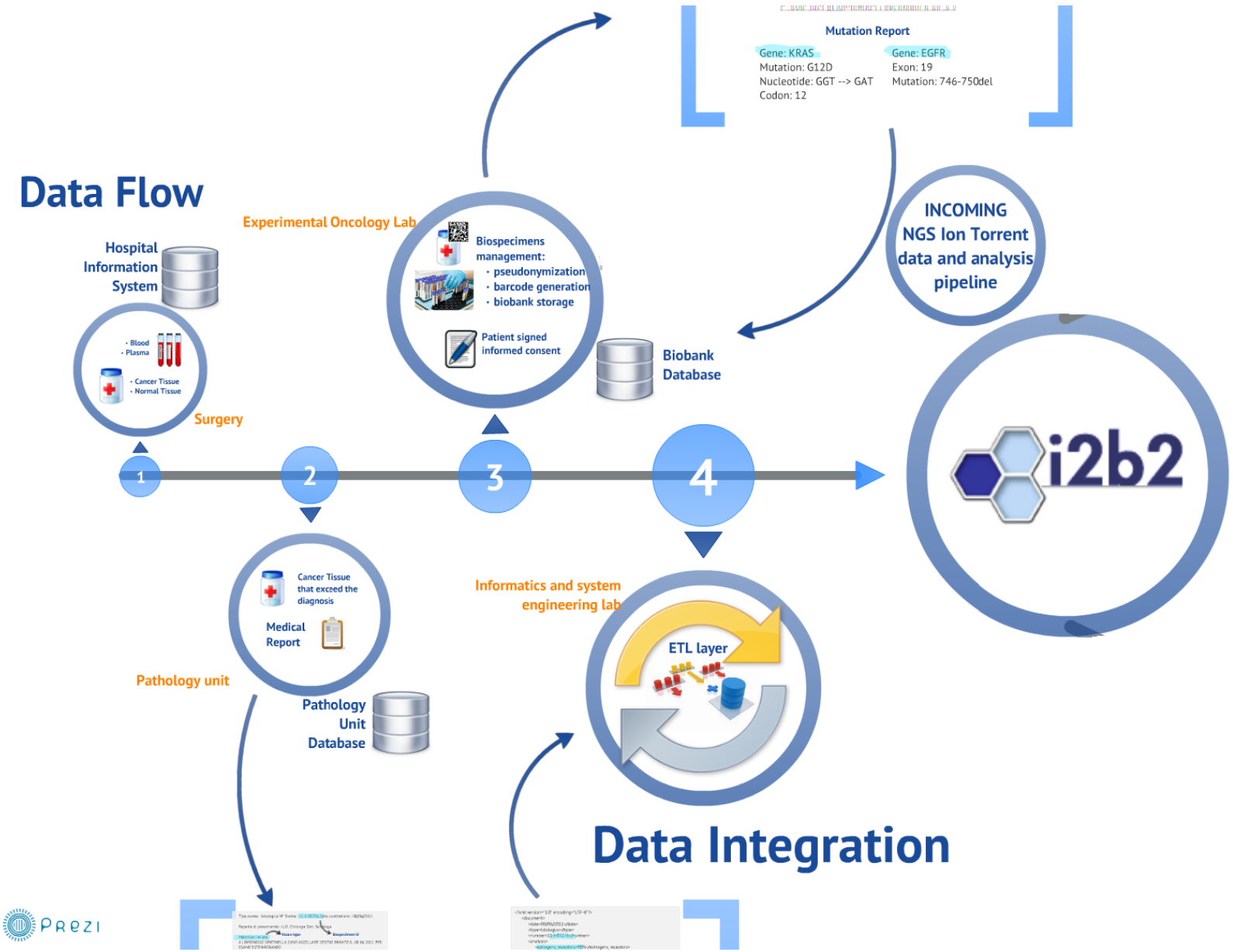
2 new cells in development

1 NLP module

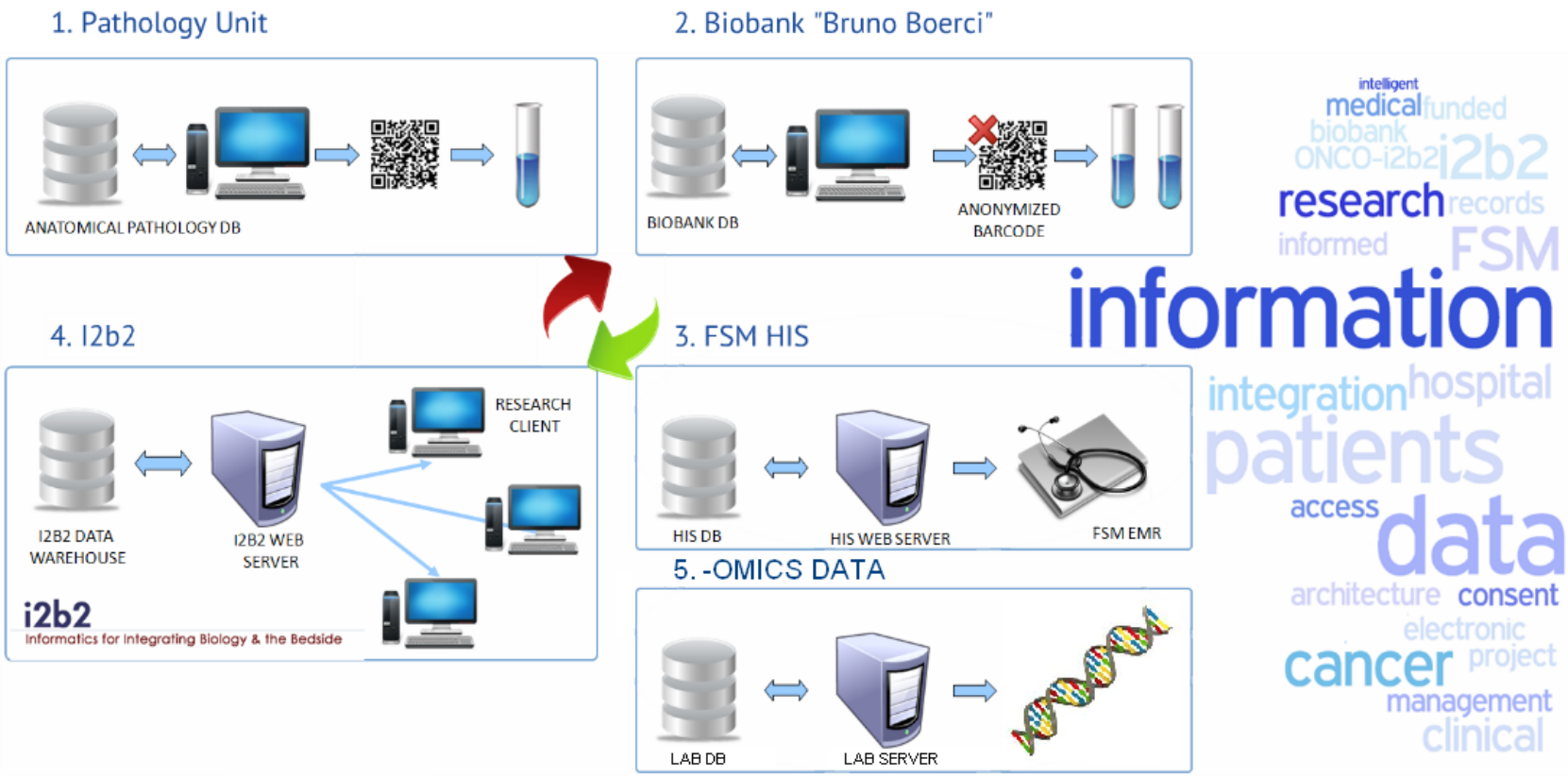


Many ETL jobs

i2b2 FSM data flow



ONCO-i2b2: System Architecture



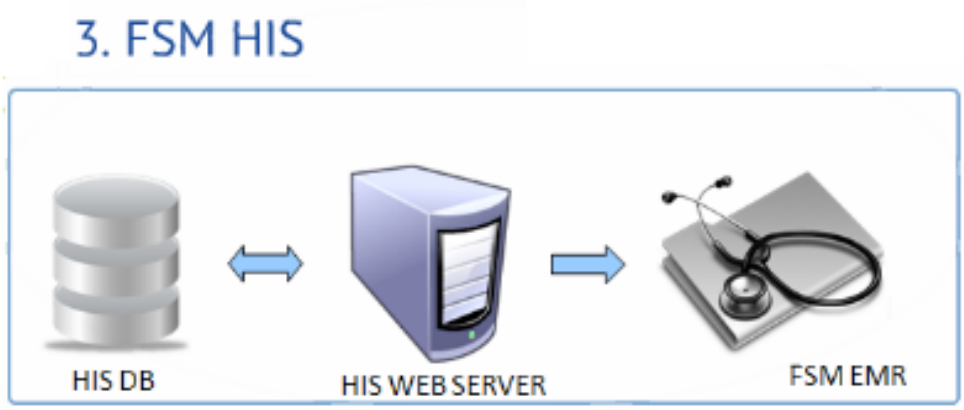
Research Data

Clinical Data

"-omics" data
[in progress]

Data Sources

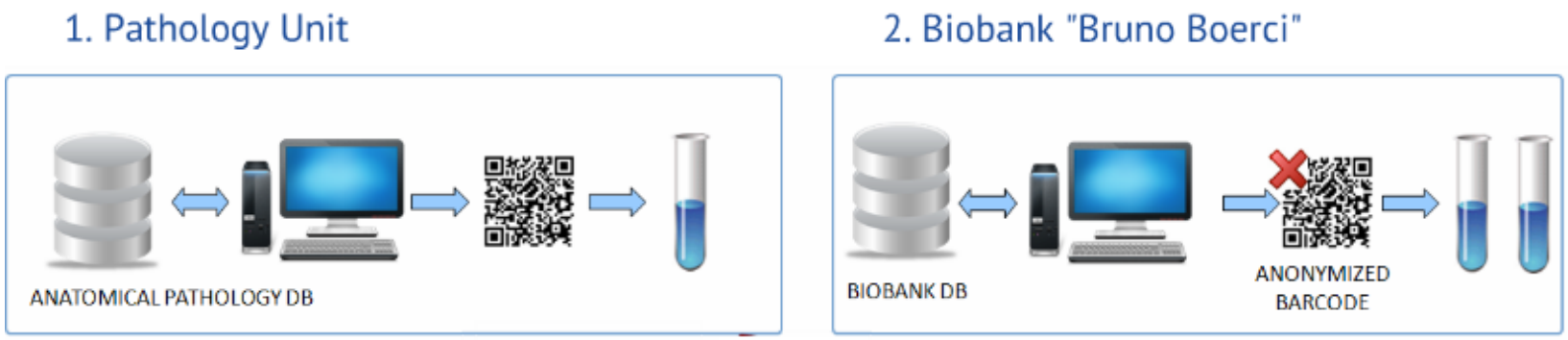
Integration process: clinical data



Integration in the i2b2 data warehouse of medical concept related to the day by day clinical practice, for example:

- Anamnesis
- Therapies
- Diagnosis (ICD9)
- Procedures (ICD9)
- Haematochemical blood analysis

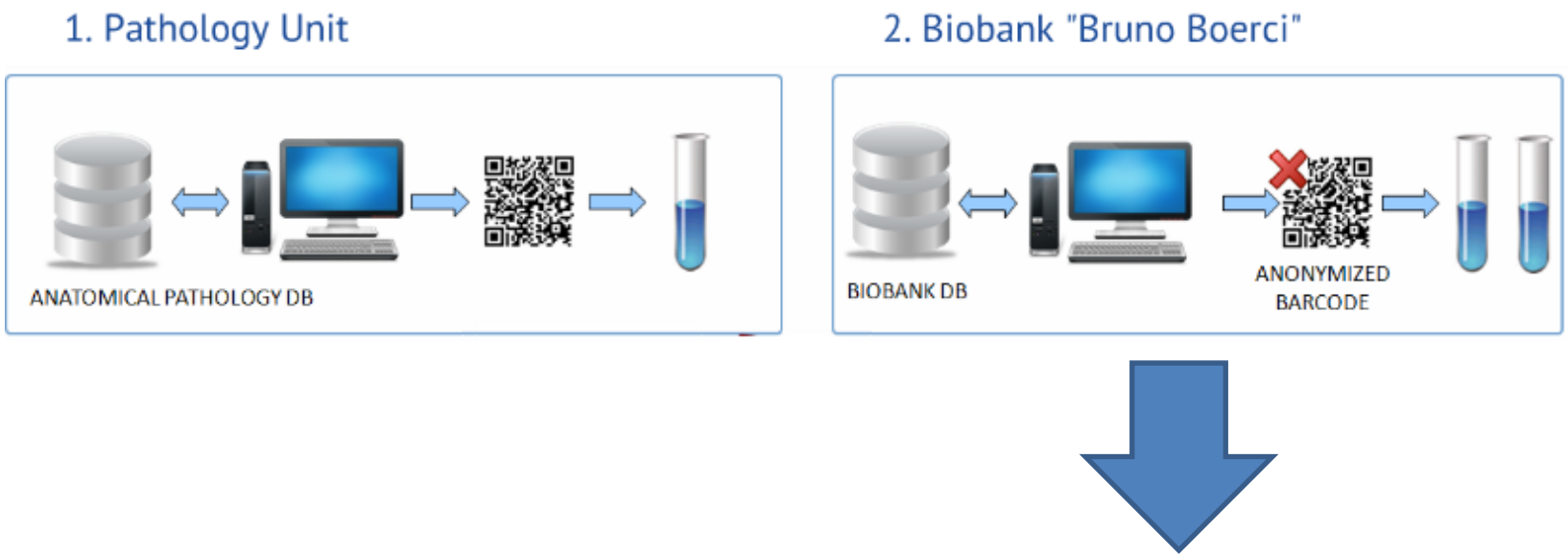
Integration process: research data



Natural language processing (NLP) identifies various concept types in the istological textual records that are associated with each patient for each medical record.

Automated tracking for samples coming from PU to biobank

Integration process: research data



Biobank sample anonymization

Once a patient is hospitalized he/she may sign an informed consent to donate samples, specimens and data collected for clinical reasons to research

NLP in action

Patology Medical Report In italian

Tipo esame : Istologico N° Esame : 11-I-03276 Data accettazione : 08/06/2011

Reparto di provenienza : U.O. Chirurgia Gen. Senologia

tissue type

biospecimen id

Materiale inviato
A LINFONODO SENTINELLA CAVO ASCELLARE DESTRO INVIATO IL 08 06 2011 PER ESAME ESTEMPORANEO
B QUADRANTE SUPERIORE MAMMELLA DESTRA

tissue description

Descrizione Macroscopica
A) Due linfonodi rispettivamente di cm. 1.6 e cm. 1.1 esminati completamente con sezioni multiple seriate.
B) Tessuto mammario di cm. 7 x 6 x 3, orientabile (peso gr. 73 dopo fissazione). Al taglio, prossima al margine infero esterno (distanza minima cm. 1.2), neoplasia a margini infiltrativi di cm. 1.

diagnosis

Diagnosi intraoperatoria
A) Due linfonodi indenni da metastasi.
Reperti e Conclusioni
B) Carcinoma duttale infiltrante della mammella, G2 SBR, a crescita di tipo infiltrativo, con scarsa reazione linfoplasmacellulare e minima componente (10%) peritumorale di carcinoma intraduttale di grado nucleare intermedio ed alto (DIN 2-DIN 3), prevalentemente di tipo solido e cribroso con focale necrosi.

Valutazione assetto recettoriale, frazione proliferante e c-erbB2:
Recettori estrogeni: 90%
Recettori progesterone: 70%
Ki67: 15%
c-erb B2:punteggio DAKO HerceptTest: 3+ (positivo); intensa e completa colorazione delle membrane citoplasmatiche nell'80% delle cellule neoplastiche.

hormone receptors values

T-D8100 - M-09410
T-04000 - M-85003
T-04020 - M-09410
T-04000 - M-09410
T-04000 - M-09410
T-04000 - M-09410
T-04000 - M-09410

SNOMED codes

Stadio : pT1b-pN0(sn)-pM Grado: G 2

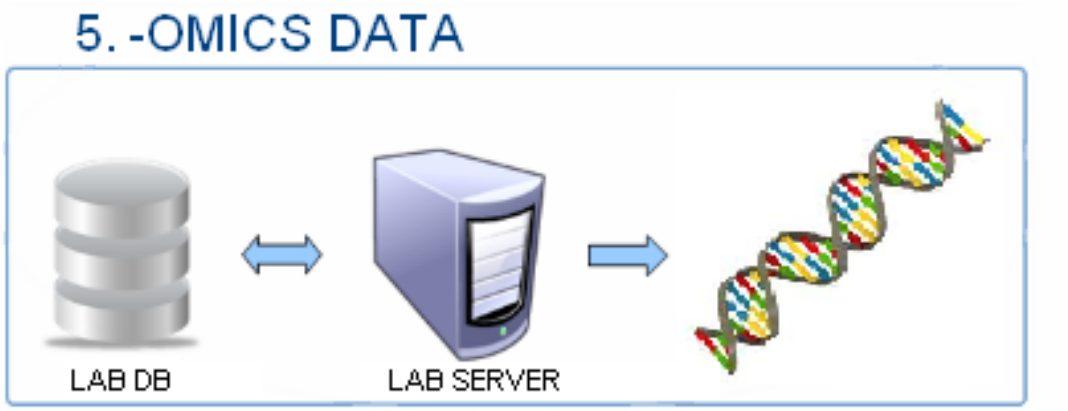
TNM classification

NLP Module

NLP output - XML format

```
<?xml version="1.0" encoding="UTF-8"?>
  <document>
    <date>08/06/2011</date>
    <type>Istologico</type>
    <number>11-I-03276</number>
    <analysis>
      <estrogens_receptors>90%</estrogens_receptors>
      <progesterone_receptors>70%</progesterone_receptors>
      <Ki67>15%</Ki67>
      <c-erb_B2>3+</c-erb_B2>
    </analysis>
    <grade>G 2</grade>
    <state>pT1bÂ-pN0(sn)Â-pM</state>
    <snomed_codes>
      <snomed consistent="true">
        <code1>T-D8100</code1>
        <code2>M-09410</code2>
        <nome1>Axilla structure (body structure)</nome1>
        <nome2>No evidence of neoplasm (finding)</nome2>
      </snomed>
      <snomed consistent="true">
        <code1>T-04000</code1>
        <code2>M-85003</code2>
        <name1>Breast structure (body structure)</name1>
        <name2>Infiltrating duct carcinoma (morphologic abnormality)</name2>
      </snomed>
      [.....]
      <snomed consistent="true">
        <code1>T-04020</code1>
        <code2>M-09410</code2>
        <name1>Right breast structure (body structure)</name1>
        <name2>No evidence of neoplasm (finding)</name2>
      </snomed>
      <snomed consistent="true">
        <code1>T-04000</code1>
        <code2>M-09410</code2>
        <name1>Breast structure (body structure)</name1>
        <name2>No evidence of neoplasm (finding)</name2>
      </snomed>
    </snomed_codes>
</document>
```

Integration process: -omics data



Data coming from High-Throughput Screening (HTS) experiments

The results of these experiments provide high amount of data that will be analyzed with the purpose of verifying the existence of genetic mutations

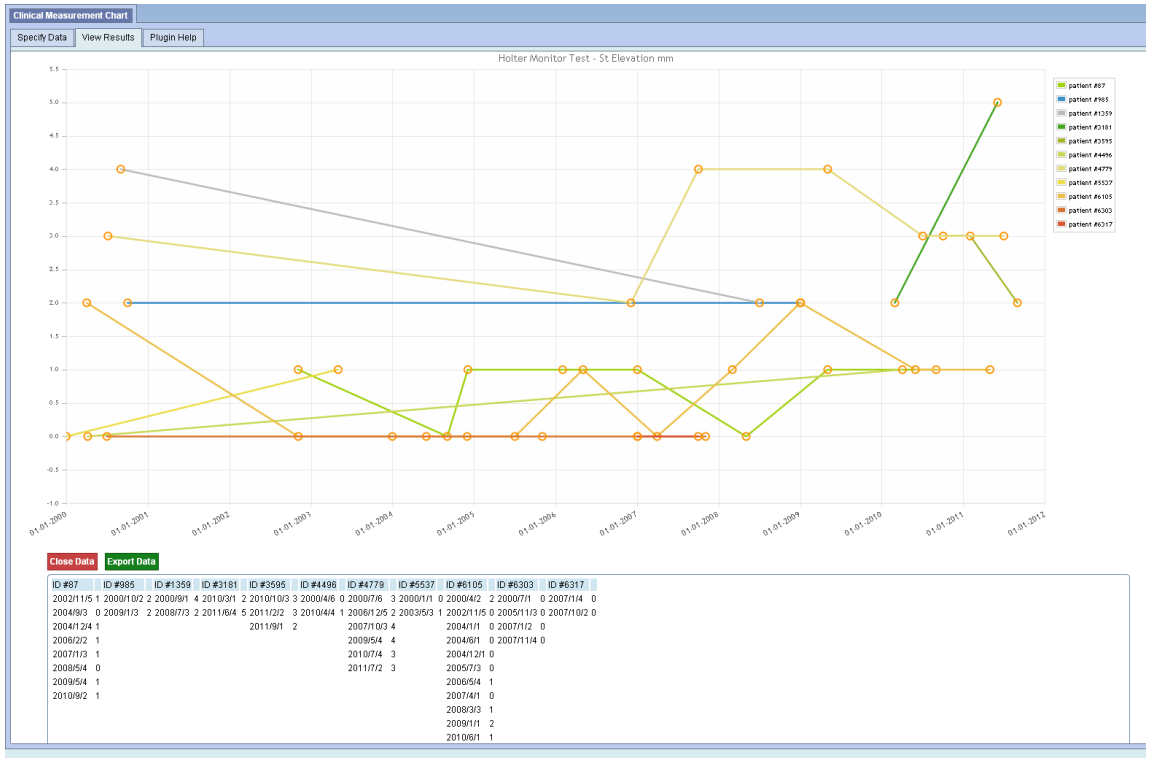
Future developments:
I2b2-noSQL database integration

Ion Proton™ Sequencer

Data Sources

I2b2 web plug-ins

- Shows trend for clinical concepts and modifiers (ecg, holter values or therapy dose) for each patient present in the patient set
- Used in oncology and cardiology



I2b2 web plug-ins

ExportXLS
This plugin tabulates unidentified patient data, and applicable diagnoses

i2b2 WIKI **ExportXLS Home**

1 Last modified on Mar 21, 2012 (view change)

This is the home of the Related Project - ExportXLS space.

- Export patient data in Excel Format V3
- Project developed by Mauro Bucalo, University of Pavia and Wayne Chan from the BMI-Core, University of Massachusetts Medical School, Worcester
- https://community.i2b2.org/wiki/display/ExportXLS

ExportXLS

Specify Data | View Results | Plugin Help

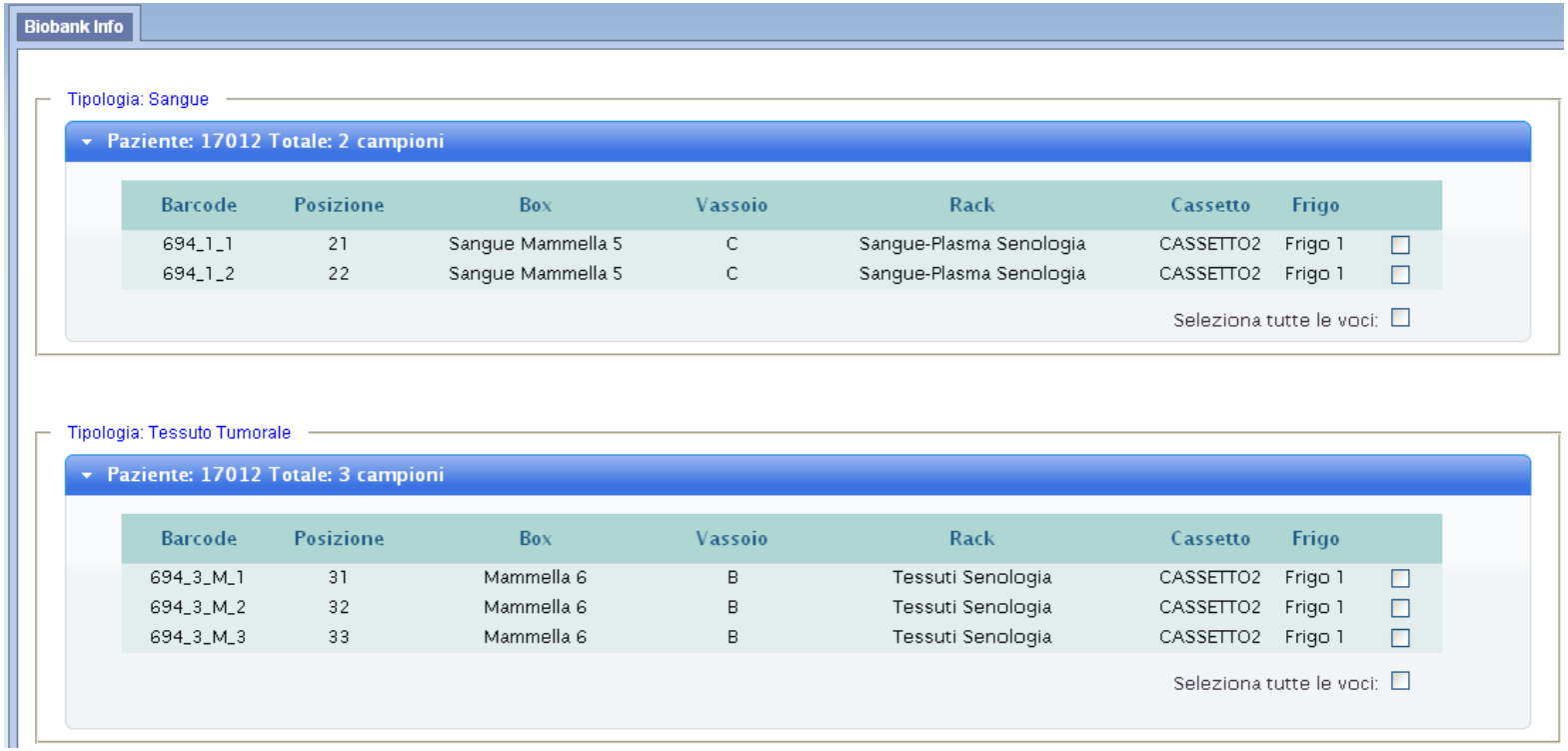
Click the [Excel Export] button to download the following table into an Excel file. Excel Export

Patient Information
for Patient Set *'Tripli_-TX - Pr@16:22:40 [10-17-2012] [daniele] [PATIENTSET_981]'*

	PATIENT_ID	Vital Status	Birth Year	Sex	Age	Language	Race	Religion	Marital Status	State\City	Income	Age	Procedures
1	284	N	1965	F	40	@	@	@	@			40	
2	705	N	1952	F	59	@	@	@	@			59	ALTRA TERAPIA FISICA
3	3637	N	1963	F	43	@	@	@	@			43	Altra riparazione o ricostruzione del capezzolo
4	8470	N	1974	F	32	@	@	@	@			31	Impianto di protesi monolaterale
5	9684	N	1955	F	52	@	@	@	@			12	Altro innesto di cute su altre sedi
6	10246	N	1951	F	60	@	@	@	@			2	Quadrantectomia della mammella
7	14103	N	1975	F	36	@	@	@	@			36	ALTRA TERAPIA FISICA
8	16680	N	1959	F	52	@	@	@	@			52	IMPIANTO DI PROTESI MONOLATERALE
9	16752	N	1947	F	62	@	@	@	@			62	Visita generale
10	17214	Y	1953	F	57	@	@	@	@			57	INIEZIONE O INFUSIONE DI SOSTANZE CHEMIOTERAPICHE PER TUMORE
11	17925	N	1957	F	54	@	@	@	@			54	Quadrantectomia della mammella
12	19128	N	1952	F	59	@	@	@	@			59	
13	27664	N	1954	F	56	@	@	@	@			56	INIEZIONE O INFUSIONE DI SOSTANZE CHEMIOTERAPICHE PER TUMORE
14	28268	N	1959	F	51	@	@	@	@			51	INIEZIONE O INFUSIONE DI SOSTANZE CHEMIOTERAPICHE PER TUMORE

I2b2 web plug-ins

- Retrive bio-speciments information (tube position, tissue desription, blood aliquots) for a selected patient set
- Used in oncology

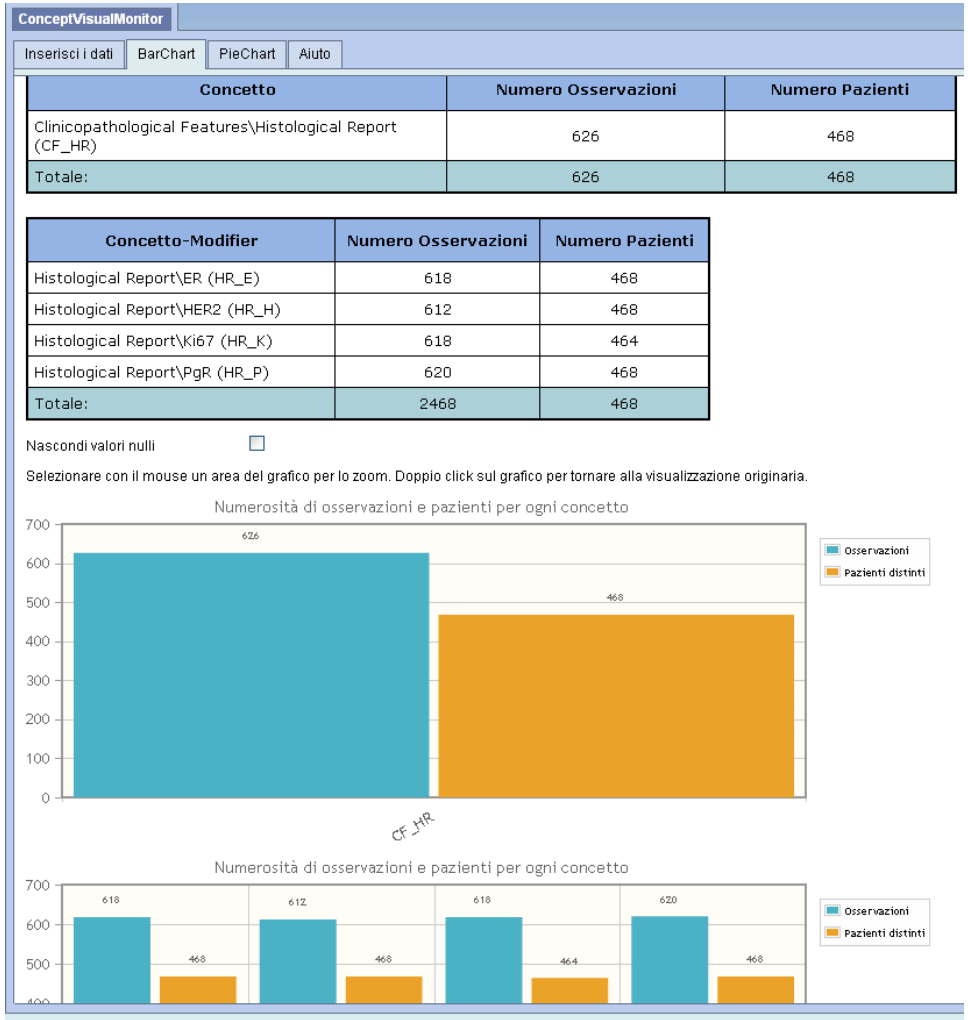


I2b2 web plug-ins

FSM - Concept Visual Monitor

Monitor the amount of concepts and/or modifier in your data warehouse

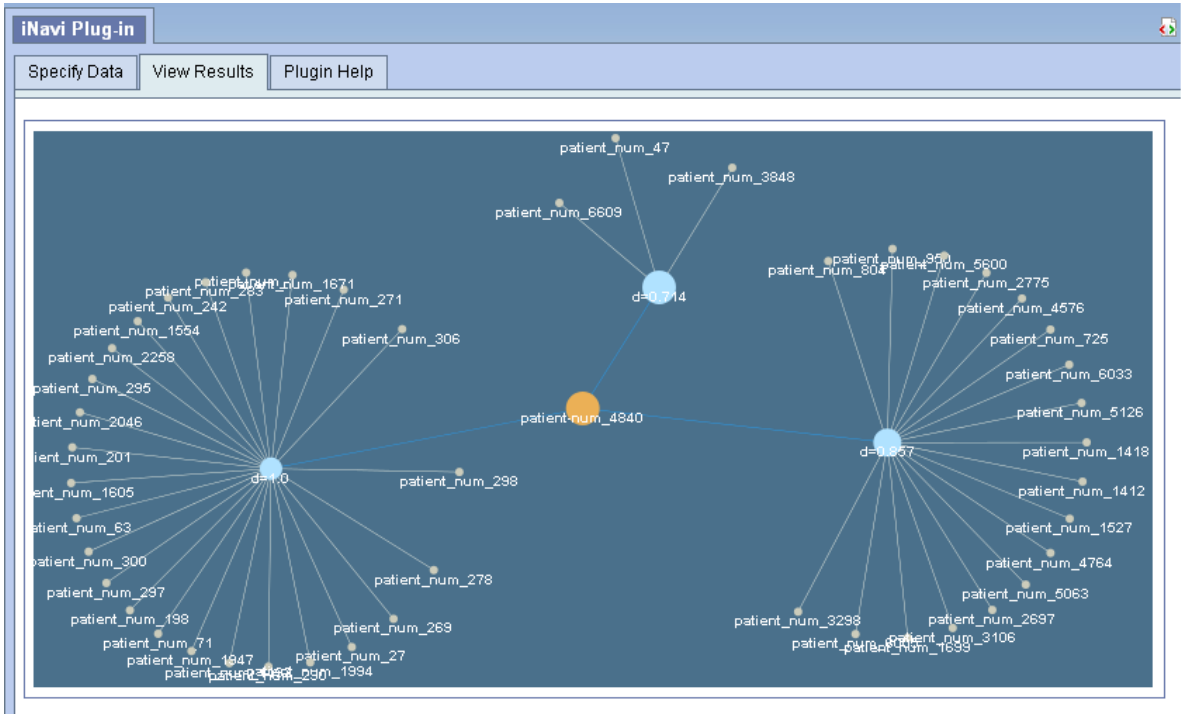
- Monitor the status of the i2b2 data warehouse by creating charts that represent the number of observations and the number of patients related to each concept or modifier selected



I2b2 web plug-ins

FSM - iNaviGraphMaker
View similar patient using this graph maker plug-in

- Using a CBR procedure, researchers are able to enhance the patient selection with a process based on the medical concept space related to a patient set to identify a group of similar patients
- Analysis based on binary, nominal and continuous variables
  - Binary variables: presence/absence of diseases or signs/symptoms
  - Nominal and continuous variables: discrete/continuous values of clinical observations

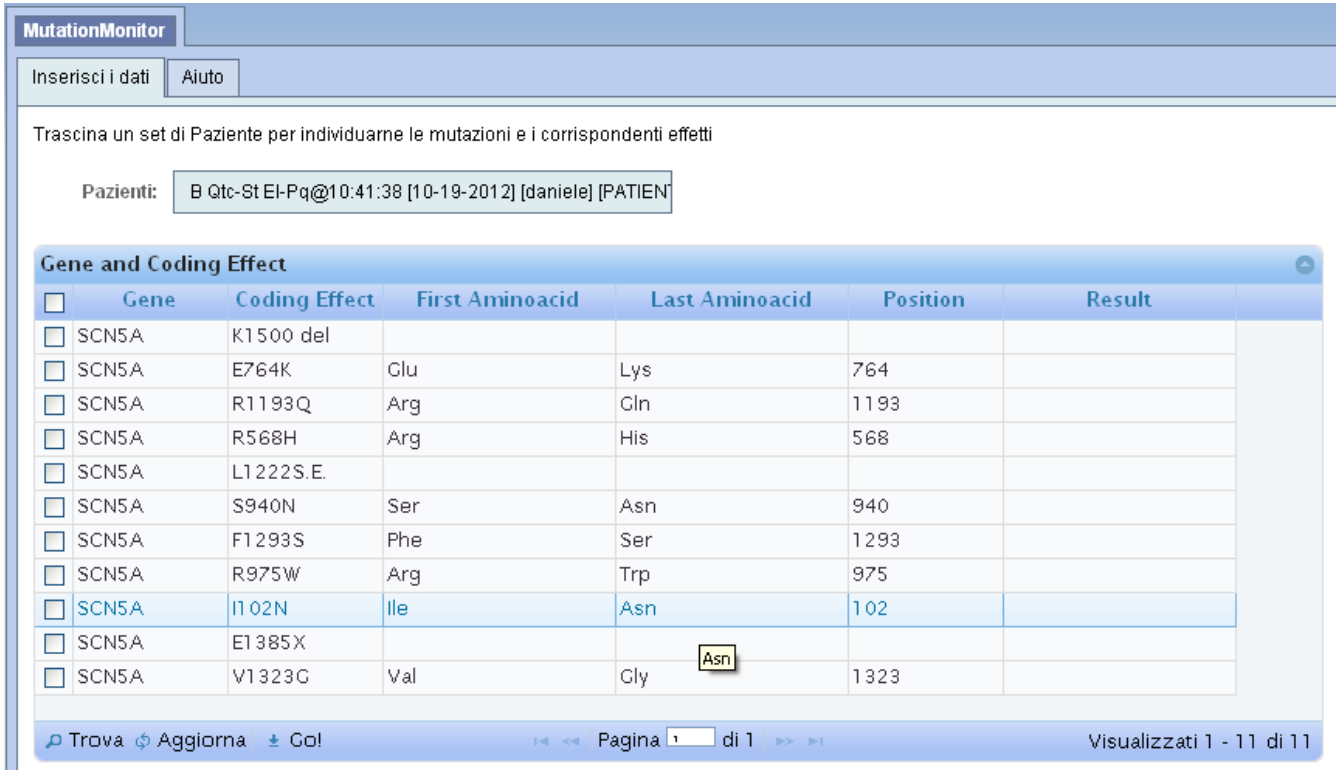


- Concept: Concept Unique Identifier (CUI) in UMLS Metathesaurus
- Distance between cases : semantic similarity between concepts in the UMLS ontology

I2b2 web plug-ins

FSM - Mutation Monitor
Mutation monitor M&M

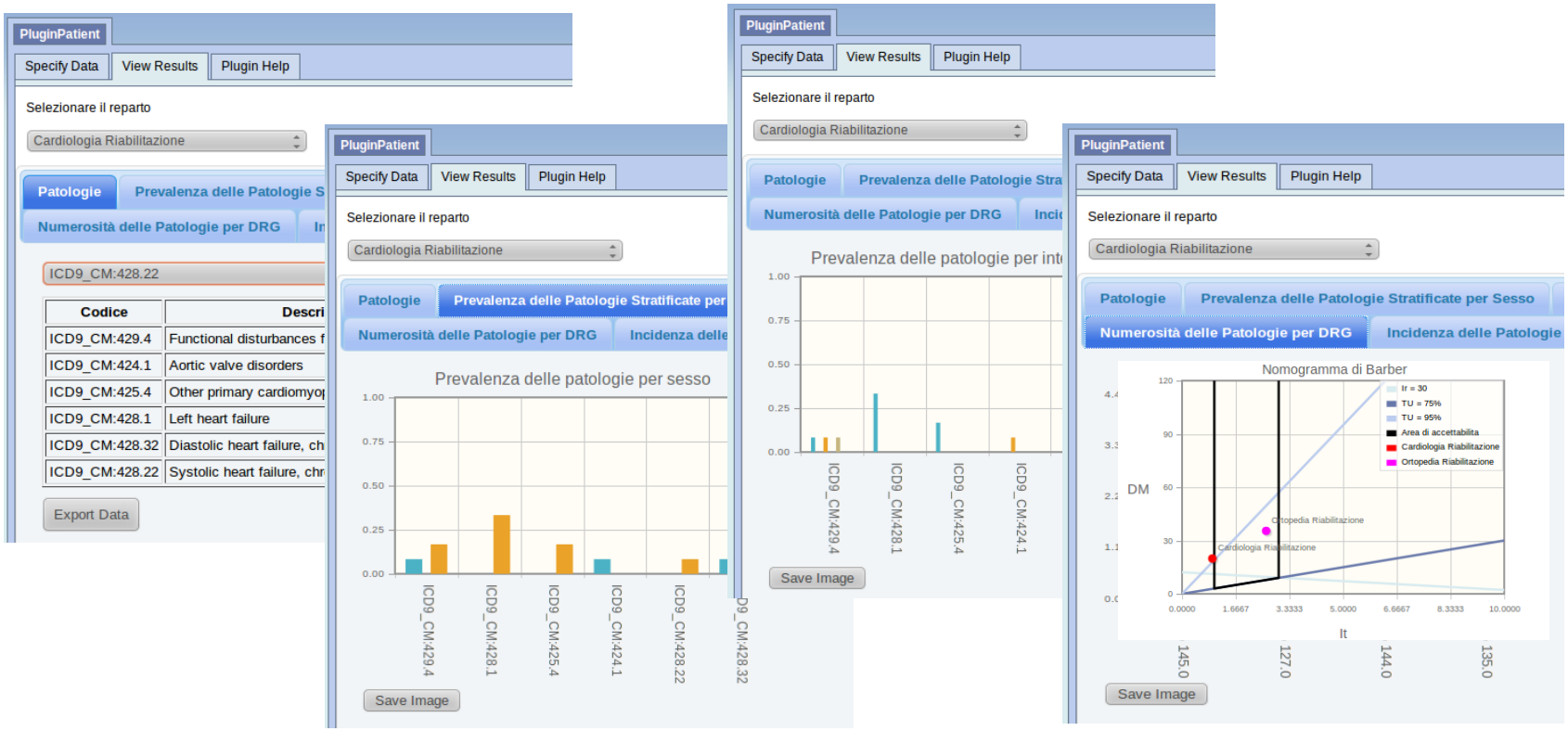
- View the genetic information and search the available publications on Exome Variant Server http://evs.gs.washington.edu/EVS/
- Used in cardiology
- Still in development



I2b2 web plug-ins

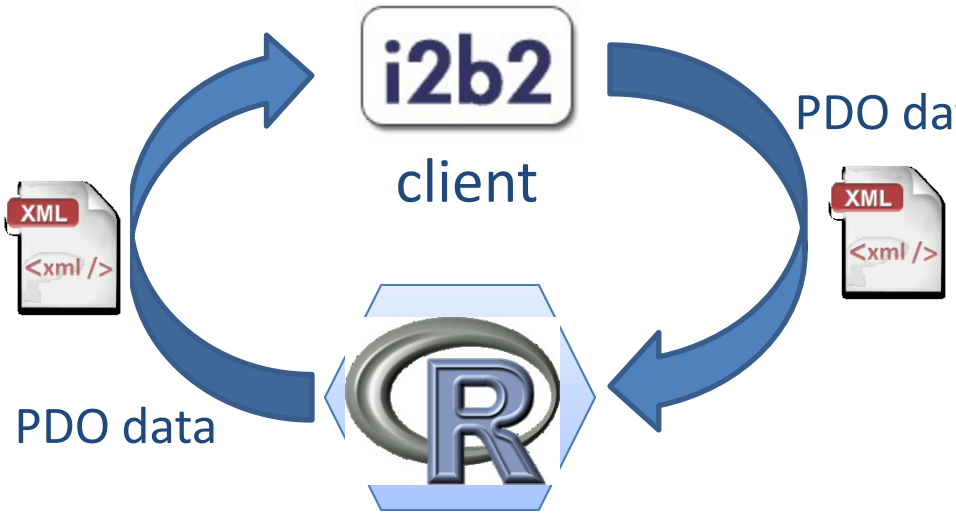
FSM - Health Activity Tracker
Graphical analysis of FSM health activity

- Developed to support the generation of reports that explain the performance of the hospital medical units
- Used only in administration



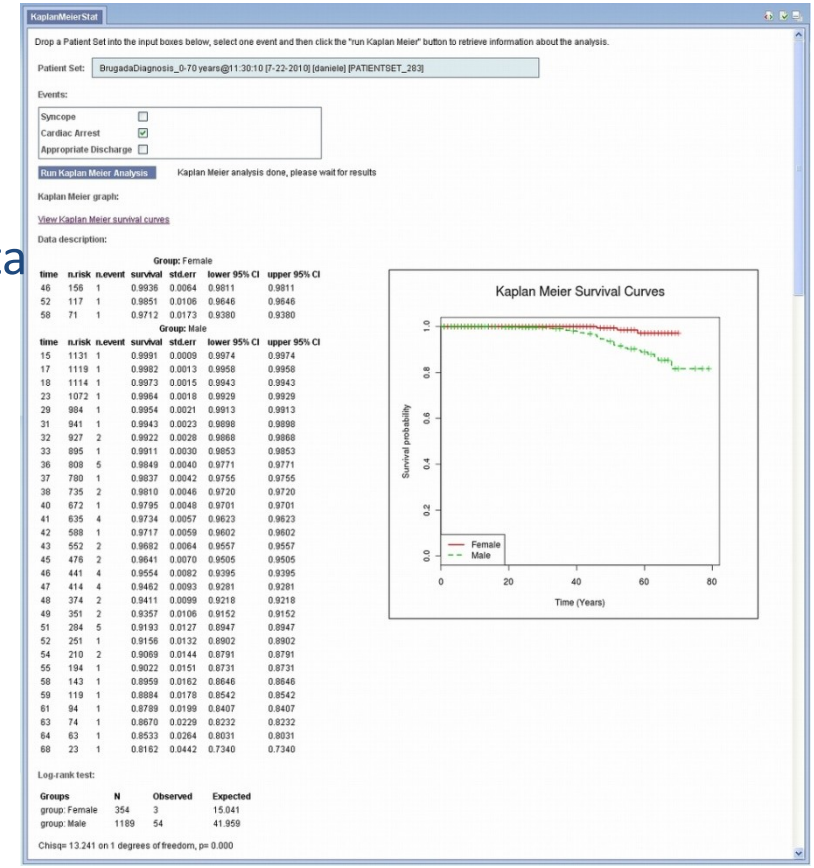
Server side developed

- Integated R statistical computing environment to perform survial analysis on demand by selecting
  - the patient set of interest
  - the event related to the survival
  - the stratification variable
- Used in cardiology
  - running on I2b2 version 1.5
  - currently in test for version 1.6



JRI - Java/R Interface
http://www.rforge.net/JRI/

rJava - Low-level R to Java interface
http://www.rforge.net/rJava



I2b2 server side development

pentaho™
open source business intelligence™
Pentaho Data Integration
Previously Kettle

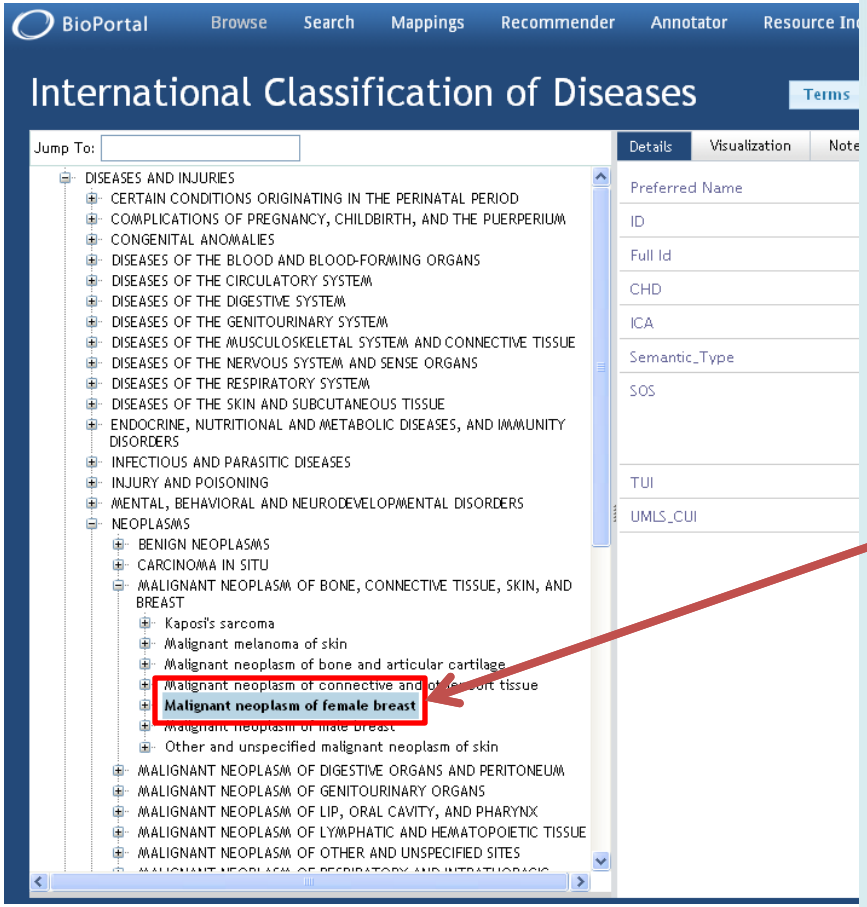
- Insert the **Pentaho Data Mining Community Edition (CE)** also known as **Weka** (original project of university of Waikato, NZ) http://weka.pentaho.com/
- tools for machine learning and data mining in a server side i2b2 cell
  - Suite of classification, regression, association rules and clustering algorithms
  - Design a dedicated plug-in for temporal abstraction

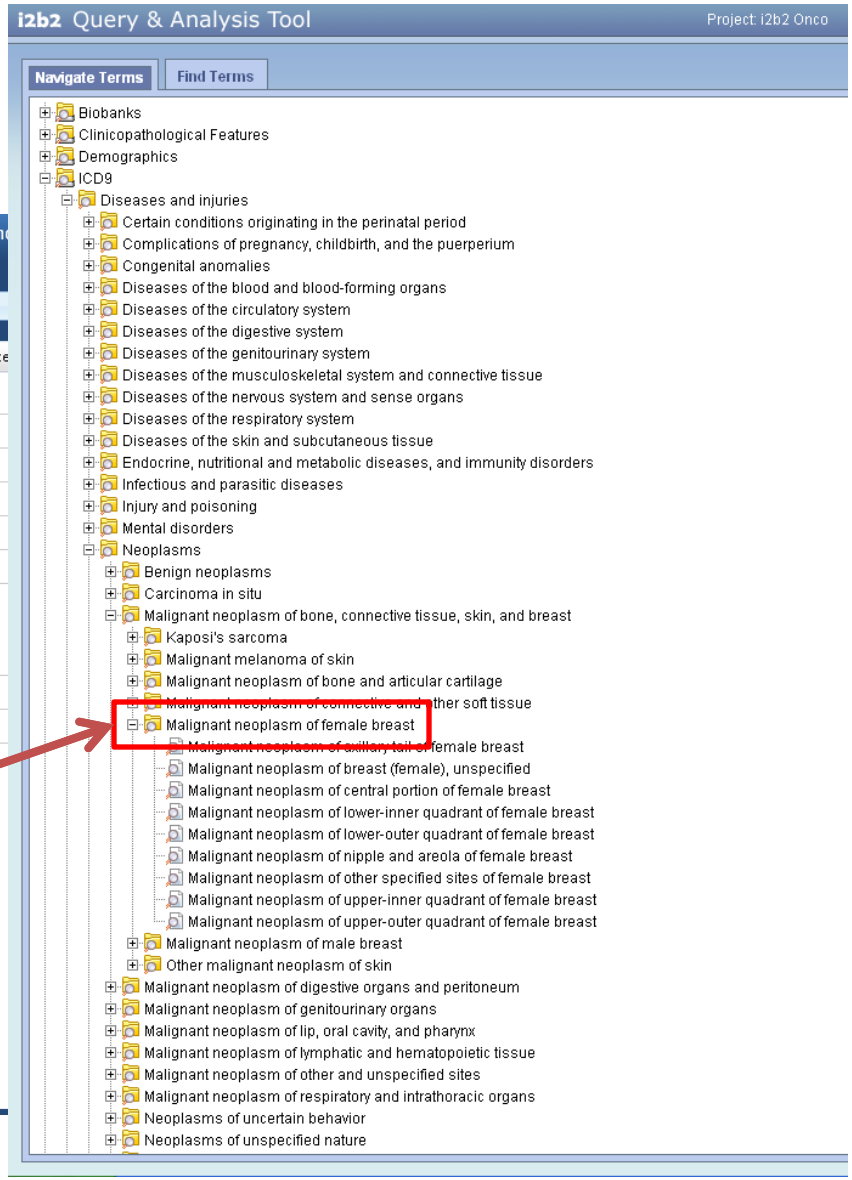
- Insert the **NHLBI Exome Sequencing Project batch query** program in a dedicated i2b2 cells, in order to retrive:
  - information about mutations
  - pubblications related to mutations
- Batch query program:
  - Based on Java 6
  - java -jar YOUR_DOWNLOADED_EVS_CLIENT_JAR_FILE -h

International ontologies

International Ontologies:

- ICD9-CM
- SNOMED
- TNM





Internationalization

Results

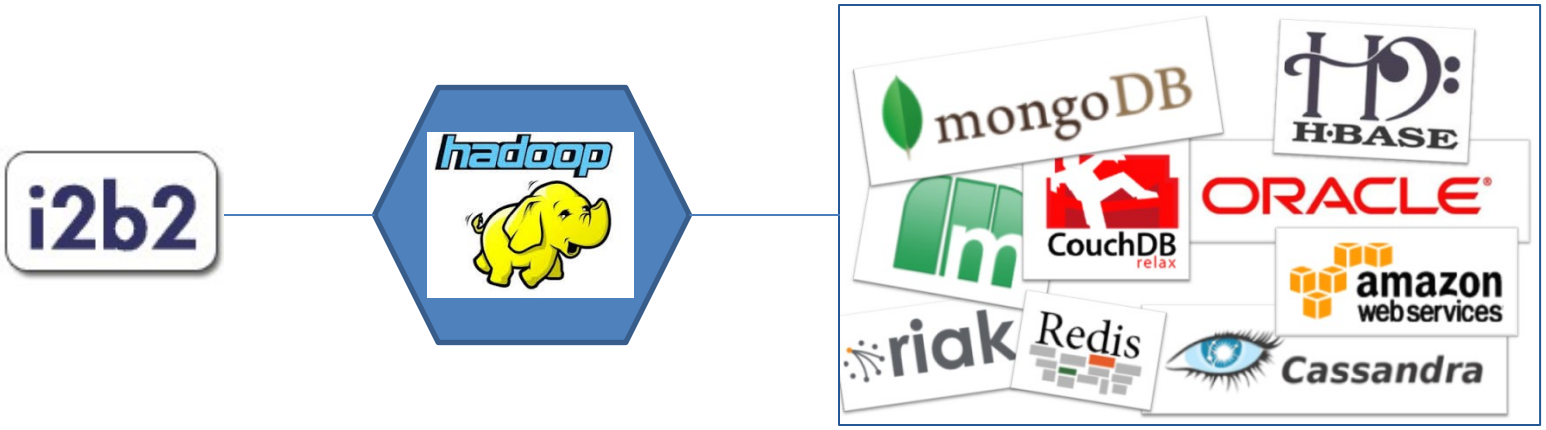
- i2b2 personalization, both on the client side and on the server side
- NLP modules to enrich the clinical information and provide in this way the possibility to query textual reports
- ETL procedures for automatic population and update of the i2b2 CRC
- Implement no/off line data analysis tools for analyze data mart instances

The IT architecture created at FSM is a concrete example of how this type of integration can be implemented and made available for increasing the quality of clinical practice as well as improving the scientific results

Future ideas: i2b2 and noSQL

- Manage noSQL databases for storage of BIG DATA coming from
  - Continuous patient monitoring systems (Cardiac Holter, Glycaemic Holter,…)
  - Next Generation Sequencing experiments
  - Exposome: mapping the environment to understand the risk of diseases

The idea is to integrate this type of non relational data source with i2b2 using server side cells and/or web plug-ins.

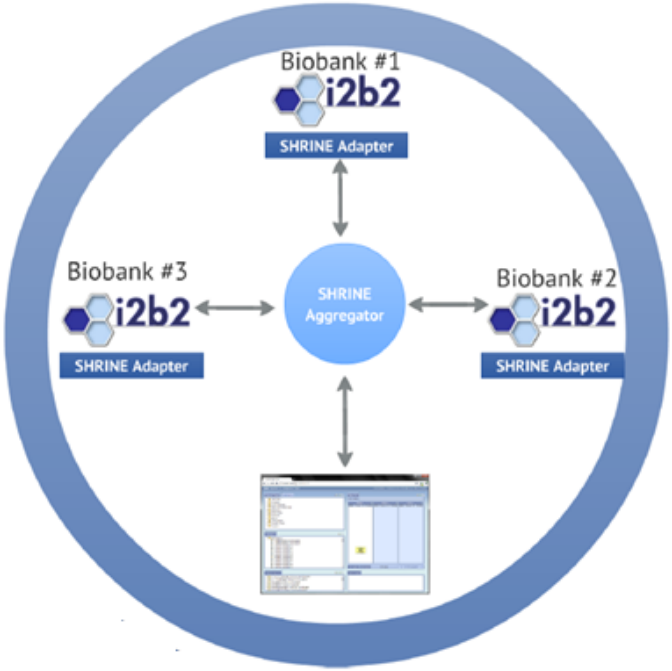


Future ideas: Shrine

Shared Health Research Information Network (SHRINE)

- Developed by Harverd Catalyst: the Harvard Clinical and Translational Science Center
- Compiling large groups of well-characterized patients
- Determine the aggregate total number of patients at partecipating hospitals who meet a given set of inclusion and exclusion criteria
- Because counts are aggregate, patients privacy is protected

http://catalyst.harvard.edu/spotlights/shrine.html



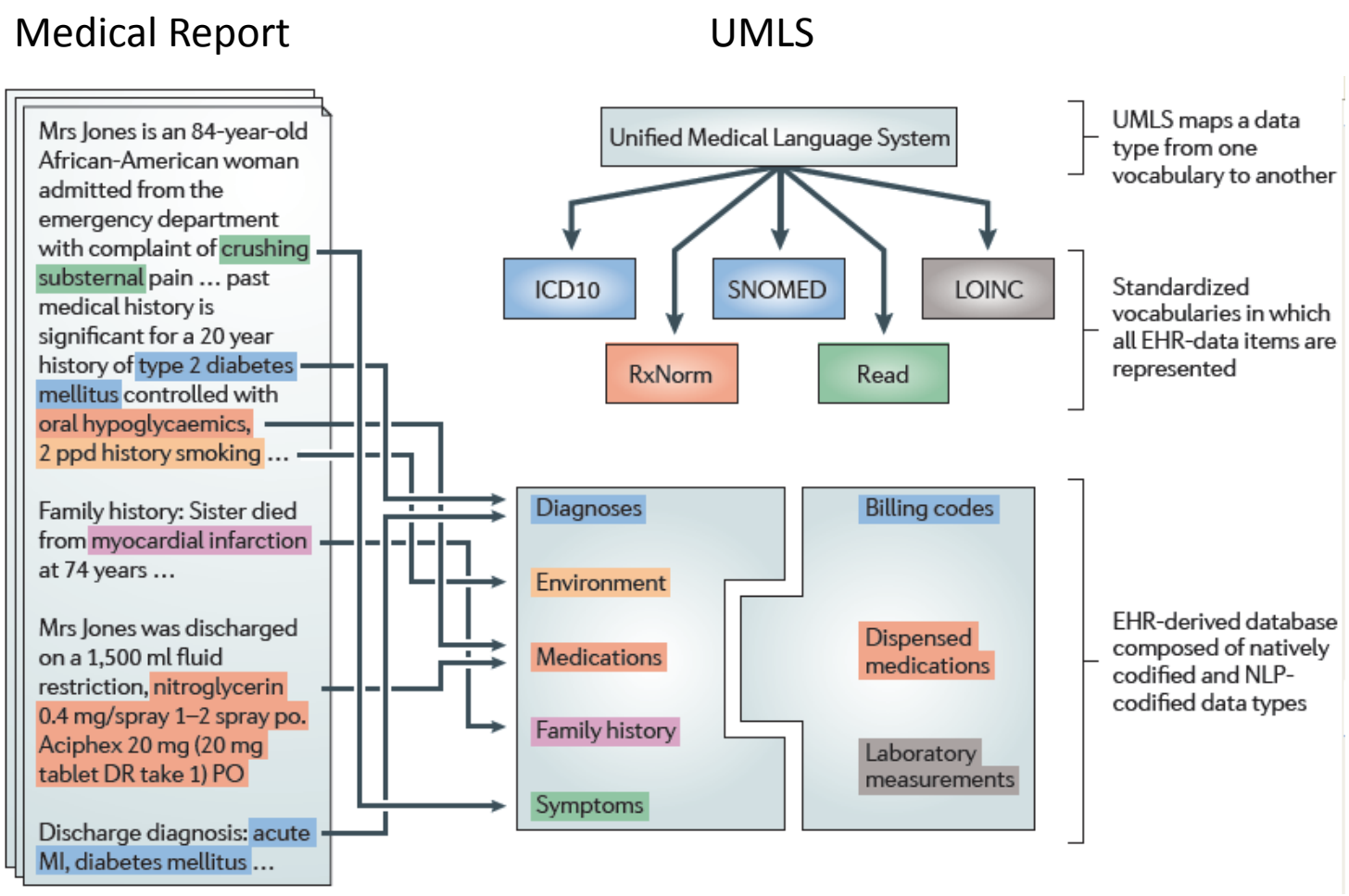
- Connections between FSM institutes in the north part of Italy (e.g. biobank network: a prototype could be ready in the coming months)
- Connections between all FSM institutes
- Open to other national and international research institutes

Clinical and research data integration: the i2b2 FSM experience

Questions?
THANK YOU

Daniele Segagni, MS, Biomedical Engineer
daniele.segagni@fsm.it

Phenotypes from clinical notes



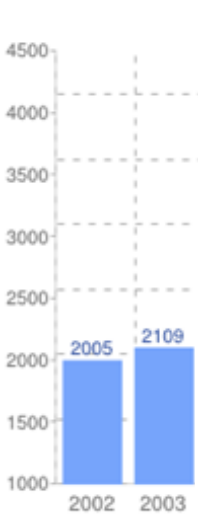
Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011 Jun;12(6):417-28. Epub 2011 May 18. Review. PubMed PMID: 21587298.

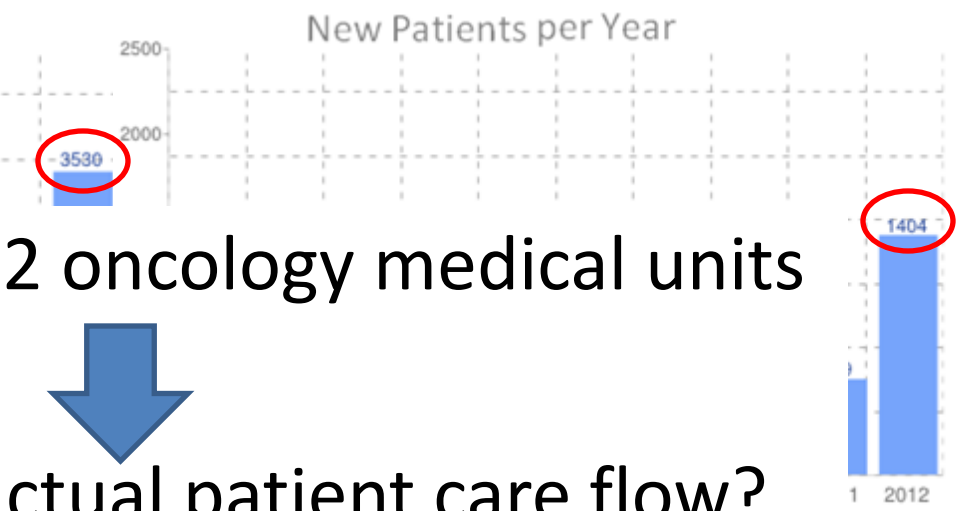
I2b2 data for clinical issues

- i2b2- ontology relying NCBO BioPortal http://bioportal.bioontology.org
- Analysis based on ICD9-CM codes related to
  - diagnosis
  - procedures
- Selection of patients with diagnosis of
  - malignant neoplasm of female breast (ICD9-CM code 174.0-174.9)
  - personal history of malignant neoplasm (ICD9-CM V10.3)
- Starting from 28.838 patients we actually focused our attention on 8.969

I2b2 data for clinical issues

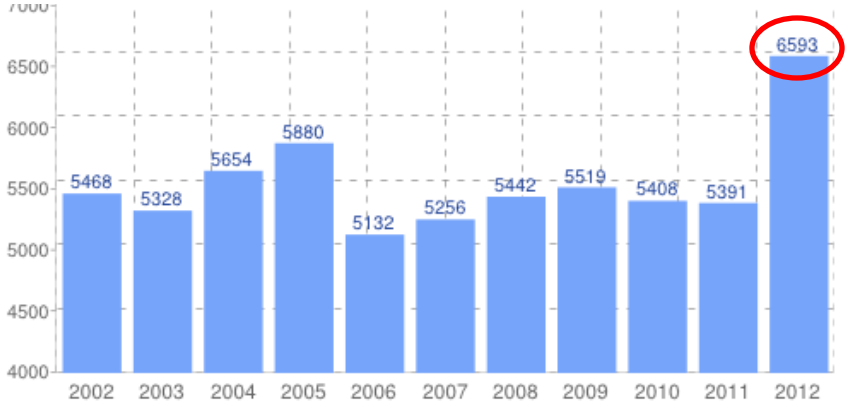
Oncology Medical Unit





2012: union of 2 oncology medical units

Which is the actual patient care flow?
Which is the adherence with guidelines?



I2b2 data for clinical issues

Which is the actual patient care flow?
Which is the adherence with guidelines?

- Off line analysis on clinical procedures and day-hospital/inpatient treatment using
  - Temporal data mining
  - Process mining
- Extract the common care flows in FSM oncology division
  - the most common pathway is hospitalization in breast surgery division followed by day-hospital treatements
- Extraction of most frequent sequences of procedures
  - The most common patterns extracted matched with the implementation of userguide
  - Example: pattern describing control ECGs that are performed due to the potential cardiotoxicity of some chemotherapies