UNIVERSITA’ DEGLI STUDI DI PAVIA

FACOLTA’ DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL’INFORMAZIONE

DOTTORATO DI RICERCA IN BIOINGEGNERIA E BIOINFORMATICA
XXXII CICLO - 2019

AN IT INFRASTRUCTURE TO SUPPORT ONCOLOGICAL RESEARCH EMPOWERED BY NLP AND TEMPORAL DATA MINING

PhD Thesis by
LORENZO CHIUDINELLI

Advisor:
Prof. Riccardo Bellazzi

PhD Program Chair:
Prof. Silvana Quaglini

Acknowledgments

I want to thank my supervisor Professor Riccardo Bellazzi for having accompanied me in all these years and for always being a source of motivation and inspiration in my academic life.

I would like to thank also Doctors Alberto Zambelli and Carlo Tondini from the Hospital Papa Giovanni XXIII, as well as my colleagues of the Hospital Information Systems, for having involved me in this collaboration and introduced to the more applicative aspect of my work.

My research would have been impossible without the collaboration with Biomeris team, Engeneers Mauro Bucalo, Nicola Barbarini and. Matteo Gabetta, and the fundamental support of PhD Natalia Viani and Arianna Dagliati and Professors Lucia Sacchi and Niels Peek.

I also want to thank all the members of the BMI laboratory, because working with people you are comfortable with makes work much more pleasant.

Abstract (Italiano)

Al giorno d'oggi i dati elettronici ospedalieri vengono mantenuti in strutture dati che spesso sono organizzati per conservare l'informazione ma non la rendono facilmente fruibile per ulteriori applicazioni. Inoltre, all'interno di questi dati si nota una suddivisione netta dell'informazione basata spesso sul dipartimento di provenienza. Manca dunque una visione integrata dei dati che potrebbe dimostrarsi una sorgente inesplorata di informazione.

In questa tesi viene presentata l'applicazione di i2b2 come piattaforma dati adatta alla aggregazione di dati provenienti dai vari database ospedalieri e al contempo al loro potenziale utilizzo in applicazioni di ricerca.

Vengono presentate inoltre due tecniche per sfruttare l'informazione dei dati precedentemente immagazzinati nel data warehouse di i2b2. In particolare, si mostra una tecnica di Natural Language Processing (NLP) basata su ontologie, per l'estrazione di informazione da referti di anatomia patologica e l'utilizzo di un metodo di Care Flow Mining (CFM) per rendere fruibile l'informazione derivante dai pattern di cura dei pazienti.

Questo lavoro è basato sulla collaborazione con il reparto di Oncologia dell'ospedale ASST Papa Giovanni XXIII di Bergamo, in cui è stata installata la piattaforma i2b2.

L'attività di tesi di seguito mostrata emerge da diverse collaborazioni.

Assieme alla spin off Biomeris dell'università di Pavia sono stati gestiti l'installazione del database i2b2 all'interno dell'ospedale, il successivo popolamento coi dati prelevati dai sistemi informativi e i successivi studi per espandere le funzionalità utili alla ricerca in ambito oncologico. La procedura di inserimento viene automaticamente ripetuta settimanalmente per garantire il costante aggiornamento dell'informazione ed è frutto di un attento studio e implementazione di procedure di ETL (Extract, Transform, Load) tramite l'uso del tool Mirth Connect per la manipolazione dei dati ospedalieri tramite query SQL.

In parallelo con la fine del popolamento principale del database, si è avviato un progetto basato interamente sui pazienti oncologici, permesso dalla flessibilità della struttura di i2b2. Mantenendo tutti i dati già importati dai vari database ospedalieri per questi pazienti, si è proceduto ad arricchire il progetto con dati derivanti dal database interno al reparto di oncologia, Oncosys.

Oltre a queste informazioni aggiuntive, su questo progetto è stata testata l'immissione di nuovi dati originati da NLP ed è stato sviluppato un progetto per l’implementazione di un plugin di i2b2 dedicato all’esplorazione dei pattern di cura dei pazienti oncologici.

Una precedente pipeline NLP è stata riadattata per l’estrazione di informazioni dai referti di Anatomia Patologica seguendo una tassonomia creata ad hoc, sul caso di tumore alla mammella.

Si è presentata inoltre l’opportunità di sviluppare un plugin i2b2, sfruttando un precedente algoritmo per il Careflow Mining, derivato dallo studio e dall’approccio tipico del process mining. Questo metodo crea un grafo aciclico che raggruppa i pattern di eventi più frequenti compiuti dai pazienti. Sfruttando la possibilità di interazione con il database i2b2 si generano dei logs file da cui è possibile evidenziare le relazioni temporali tra eventi e l’identificazione di pattern clinici differenti in pazienti che percorrono eventi diversi.

L’attività svolta risponde al bisogno dei clinici di avere accesso facilitato allo storico dei trattamenti dei pazienti e al bisogno dei ricercatori di poter definire coorti di pazienti per approfondirne la ricerca oncologica.

Abstract (English)

Nowadays, digital health care data are stored in structures that are often organized to consistently preserve information but do not make it easily accessible for further applications.

Moreover, the collection of these data is often fragmented, for example on the basis on the department of origin. As a consequence, there is a lack of a comprehensive data overview that could be an unexplored source of information itself.

This thesis presents an application of i2b2 (Informatics for Integrating Biology & the Bedside) as a data platform for data aggregation from different databases across the hospital and at the same time for their potential use in research applications.

Two techniques are presented to exploit the data previously stored in the i2b2 data warehouse. First, a Natural Language Processing (NLP) pipeline based on ontologies is exploited for the extraction of information from anatomical pathology reports and, then, a CareFlow Mining (CFM) algorithm is used to embed the information extracted by NLP in the process of discovery patterns of care.

This work has been carried on thanks to the collaboration of the Oncology ward of the Hospital ASST Papa Giovanni XXIII in Bergamo.

The i2b2 platform was installed within the hospital, in collaboration with Biomeris s.r.l., a spin-off company of the University of Pavia. i2b2 was populated with data taken from different database, and a novel set of software solutions has been implemented to expand the i2b2 functionalities and to support oncology research.

The i2b2 data warehouse is automatically updated weekly to guarantee its alignment with clinical practice. The updating procedure includes a careful study and implementation of ETL (Extract, Transform, Load) procedures through the use of the Mirth Connect tool for the manipulation of hospital data via SQL queries.

Aside of the main i2b2 project, which includes all patients in the hospital, a second vertical Oncology project has been developed for oncology patients only. This vertical project imports all the oncology patients' data available in the main i2b2 project and merges them with an additional source of data derived from the internal oncology ward database, Oncosys.

Within the vertical Oncology project, a previously implemented NLP pipeline has been adapted for extracting information from the anatomical

pathology reports, based on an ontology of breast cancer, and the information originating from NLP was included.

Taking advantage of the i2b2 data warehouse augmented by NLP, it was possible to apply a careflow mining algorithm (CFM), which highlights the temporal relationships between events and identifies different clinical patterns in patients. An i2b2 plugin has been designed, in order to embed CFM as a commodity for clinicians and researchers during data exploration.

This research responds to the need of clinicians to have easy access to the history of patient treatments and to the need of researchers to be able to define cohorts of patients to support cancer research.

Contents

Abbreviation list

AIOM Associazione Italiana Oncologia Medica
AJCC American Joint Committee on Cancer
ASST Azienda Socio-Sanitaria Territoriale
CFM CareFlow Mining
CRC Clinical Research Chart
DAG Direct Acyclic Graph
DICOM Digital Imaging and COmmunications in Medicine
EAV Entity Attribute Value
EDI X12 Electronic Data Interchange X12
EHR Electronic Health Records
ETL Extract, Transform, Load
FN False Negative
FP False Positive
FROM Fondazione per la Ricerca Ospedale Maggiore
GUI Graphic User Interface
HDR Hospital Discharge Register
HER2 Human Epidermal growth factor Receptor 2
HIS Health Information Systems
HL7 Health Level Seven International
HPG23 Hospital Papa Giovanni XXXIII
HTTP HyperText Transfer Protocol
i2b2 Informatics for Integrating Biology & the Bedside
ICD9CM International Classification of Diseases 9 Clinical Modification
ICS Istituto Clinico Scientifico
IE Information Extraction
IRCCS Istituto Ricovero e Cura Carattere Scientifico
IT Information Technology
JMS Java Message Service
JSON JavaScript Object Notation
LOINC Logical Observation Identifiers Names and Codes
NCPDP National Council for Prescription Drug Programs
NLP Natural Language Processing
OWL Web Ontology Language
REST Representational State Transfer
SISS Sistema Informativo Socio-Sanitario
SMTP Simple Mail Transfer Protocol
SNOMED-CT Systematized NOmenclature of MEDicine Clinical Terms
SOAP Simple Object Access Protocol
SPU Short Procedure Unit

TCP Transmission Control Protocol
TCP/IP Transmission Control Protocol /Internet Protocol
TM Topic Modeling
TNM Classification of Malignant Tumors, lymph Nodes, Metastasis code
UIMA Unstructured Information Management Architecture
UMLS Unified Medical Language System
URI Uniform Resource Identifier
W3C World Wide Web Consortium
XML eXtensible Markup Language

Chapter 1

Introduction

This PhD project is focused on the study and the development of Information Technology solutions to improve clinical and oncological research. The project has been developed in partnership with the Oncology unit of the Hospital ASST Papa Giovanni XXIII [1] in Bergamo.

The activities presented in this thesis have been discussed and implemented in collaboration with the Biomeris [2] group, an academic spin-off of the University of Pavia, specialized on informatic solutions to support health research and data management.

This chapter presents a brief introduction to the data assets of the hospital, a subsequent explanation of the oncology domain, focusing on breast cancer, and an introduction to the problems that motivated this research activity.

1.1. Hospital Structure and Data

The Papa Giovanni XXIII Hospital (HPG23) [1] is a multidisciplinary center of care in the city of Bergamo, in Lombardy. It centralizes and offer care services to the entire Bergamo district, as a territorial health authority called ASST (Azienda Socio Sanitaria Territoriale).

The hospital, although not formally being an IRCCS (research hospitals), fosters health research, and health care innovation, including a strong focus on research projects in oncology.

In the area of oncology, all solid and hematologic tumors are treated, and patients are followed in every phase of the disease, providing accurate diagnosis and personalized treatment plans, including surgery, radiotherapy, various types of chemotherapy and palliative care.

HPG23 is a leading hospital in both the national and international scenario in several clinical areas. During 2018, it performed more than 40’000 hospitalizations (ordinary and day hospital), 35’000 surgery procedures and over 4 million clinical services for outpatients.

HPG23 health services characteristics (such as volume and heterogeneity) are crucial design elements of the data management solutions and implementations provided in this thesis.

1.2. Breast Oncology Care

The PhD work described in the thesis is mainly focused on breast cancer. This chapter provides a short overview of the breast cancer cycle of care [3] and a description of the main data management issues considered in this work.

Breast cancer starts when cells in the breast tissue begin to grow out of control and develops invading the surrounding tissues or spreading (metastasize) to distant areas of the body through blood or lymph system.

Different types of breast cancer can be distinguished based on the specific affected cells in the breast: carcinomas start from epithelial cells; adenocarcinomas start in the ducts (ductal carcinoma) or the lobules (lobular carcinoma) tissue of the breast. It is also possible to distinguish non-invasive (e.g. carcinoma in situ) and invasive breast cancer (e.g. invasive carcinoma).

Breast cancer detection typically occurs when a woman or her doctor discover a mass or abnormal calcification on a screening mammogram, or during a clinical or self-examination. Then, the most accurate method to make a definitive diagnosis is a microscope examination of a small amount of tissue collected with a biopsy.

Once the specimen is analyzed and the diagnosis of breast cancer is determined, oncologists try to figure out if it has spread, and if so, how far in the organism. This process is called staging and often it is based on the American Joint Committee on Cancer (AJCC) [4] definition, including the information from estrogen/progesterone/growth factor receptors, cancer Grade and TNM classifications.

The biopsy specimen is tested for the presence of estrogen (ER) or progesterone (PR) receptors. Based on these investigations, cancers are called hormone receptor-positive (ER+, PR+) or hormone receptor-negative (ER-, PR-). When the hormones estrogen and progesterone attach to cancer cells receptors, they fuel the cancer growth.

Another specific test concerns the status of HER2 (human epidermal growth factor receptor 2), a growth-promoting protein. Breast cancer cells with higher than normal levels of HER2 are called HER2+. These cancers tend to grow and spread faster.

All these analyses allow physicians to select specific hormone therapy or HER2 target drugs.

During the microscope pathology analysis, cancer cells receive a histologic grade based on three features: gland formation, nuclear grade, and mitotic count.

- Grade 1 or well differentiated. The cells grow slowly and look like normal breast tissue.
- Grade 2 or moderately differentiate. The cells grow with a speed between grades 1 and 3.
- Grade 3 or poorly differentiate. The cancer cells look very different from normal cells and probably grow and spread faster.

Another classification assigned to the specimen is the TNM staging system:

- primary Tumor (T): include information about the original tumor (e.g. its size and how deeply and quickly it has growing).
- Node involvement (N): information about cancer spreading into nearby lymph nodes.
- distant Metastases (M): information about the cancer spread.

Based on the previously described analysis, the patient treatment combines different types of treatments including:

- Surgery, to remove cancer in the breast through lumpectomy or mastectomy.
- Radiation Therapy, to destroy cancer cells with high-energy x-rays or other particles.
- Systemic Therapy, to destroy cancer cells through the bloodstream with substances, such as chemotherapy, hormonal therapy, target therapy and immunotherapy.

If cancer returns after treatment of an early-stage disease, it is called recurrent cancer; it may come back in the same place as the original cancer (local recurrence), in the chest wall or lymph nodes under the arm or in the chest (regional recurrence) or other organs such as the bones, lungs, liver, and brain (distant or metastatic recurrence). Diagnosis and treatment methods previously described are repeated, also in this case.

1.3. Data Management Issues and Solutions

From the data management point of view, the main aspects that emerge from the previously described scenario are the heterogeneity of data and data sources, the importance of information contained in pathology reports (very often written in natural language) and the high importance of the sequence of events occurred in the diagnosis and treatment of the cancer.

Such challenges have been addressed with solutions devoted to data integration and exportability, natural language processing and process mining.

1.3.1. Data Integration

The Hospital Information System (HIS) of HPG23 integrates patients' data from multiple sources collected since 2007. This context shows at least two of the so-called Big Data "V" properties [5][6]: Volume, i.e. "large" amounts of data, and Variety, i.e. diverse formats.

Another relevant aspect to be considered is Veracity (the uncertain nature) of the data. Even if HPG23 data were collected accordingly quality and reuse principles, a series of controls needed to be performed.

Electronic Health Records (EHR) have progressively become a crucial component of clinical practice, and public health [7].

However, they can be also exploited as drivers for clinical research. In order to achieve this goal, it is essential to build suitable data warehouses and repositories, which requires to manage simultaneously multiple type of data, extract relevant information from them and aim to provide more tools to the research or clinical practice.

The i2b2 (Informatics for Integrating Biology & the Bedside) data warehouse [8] is one of the platforms that can fulfill these requests. The mission of i2b2 is to improve clinical and research practice with an infrastructure able to integrate both clinical records and research data [9]. It is an open source tool that permits data collection from multiple sources, combining them in a database, designed to easily describe very different types of data and to provide easy methods to make patient cohort identification.

The i2b2 platform is currently installed in more than 250 international locations. One of the most important features is the capacity of create multi institutional networks thanks to explicit i2b2 installation procedures, which allow preserving important aspects of privacy. This allows to query data in multiple health and research centers in the same network, using the results to boost research activities.

Examples of i2b2 implementations have demonstrated good results in data integration between clinical and administrative records. I2b2 allows identifying patient phenotypes using clinical variables and temporal references [10]. It supports translational research in oncology, integrating anatomical pathology unit database with biobank information [11].

The approach of creating the i2b2 platform as a secondary database, leaves the health centers free to continue using their existing clinical systems. This type of platform can also enable a “sidecar” use of the data warehouse, i.e. not only using it to support research but also as a tool for finding groups of similar patients during clinical care [12]. In this scenario, while patient data are collected as usual with EHR during the daily clinical practice, the i2b2 task is to aggregate copies of the data coming from EHR to its own data warehouse, dealing with interoperability, privacy and security.

1.3.2. Natural Language Processing (NLP) Techniques

During daily clinical practice a great amount of textual reports are generated, containing meaningful clinical knowledge. The automated extraction of this information with Natural Language Processing (NLP) methods allows filling structured databases and enabling statistics.

NLP methods have been successfully used to analyze and extract information from English clinical texts [13]. An interesting research area is therefore to provide suitable modifications or reimplementation of NLP pipelines for other languages and for other domains [14].

Looking at Information Extraction (IE) methods in English clinical narratives, it is possible to note that the IE task is performed with rules and lexicons [15], like Unified Medical Language System (UMLS [16]) or with machine learning approaches [17].

The application of the same approaches to Italian narratives is made difficult by the lack of annotated resources.

Algorithms that do not require annotated texts has been explored also for Italian. The use of Italian UMLS [18] resulted in smaller coverage than the English application and IE results were lowered by the lack of tools able to perform the step of lexical variant generation.

Dictionaries and standard NLP tools can be used to discover entities relations, too. For example, Alicante et al. [19] applied clustering methods to identified relations and automatically label sentences.

Domain ontologies can be successfully exploited to guide the IE task. This method has been applied also to non-English settings, combined with rule-based approach [20]. An ontology-based system has been applied to the recognition of anatomical sites, attributers and values in the text [21]. Authors created the ontology in strict collaboration with domain experts, iteratively modifying the system based on the annotations produced.

In this thesis we considered a pipeline described in [22], which uses ontology to perform IE task on a set of Italian medical reports of cardiology domain.

The exploitation of the NLP pipeline required a domain ontology specific of the breast cancer. The ontology was developed in collaboration with the HPG23 oncologists and produced using the Protégé tool [23].

The contents of the ontology specify the concepts to be extracted and relations between them; this enables to recognize main events in pathology reports and search relative attribute that characterize them.

1.3.3. CareFlow Mining (CFM) Methods

Usually, healthcare data are collected with time stamps. Data that span over time are called longitudinal data. The time windows of this longitudinal data can be very different, like the evolution of a chronic disease during the years or a monitoring of blood glucose in a day.

In many cases, the sequence of clinical events and their time relations reveal important aspects of clinical practice. The reconstruction of these clinical pathways is a field of data mining in healthcare, known as careflow mining (CFM) methods. The growing availability of EHR generated data pushes towards the creation of tools that take advantage of patient clinical histories to reconstruct pathways.

Careflow mining often uses techniques and tools developed in the process mining. [24][25].

CFM algorithms study patients' events of interest incorporated in event logs and try to extract meaningful real-life histories and provide useful insight of the care processes. They have been applied to chronic disease [26] and can elaborate events logs containing both clinical and administrative information [27].

Unlike clinical process modelling, which creates workflows based on guidelines, careflow mining illustrates the most frequent sequence of events from the real data.

Once the careflows have been discovered, they can be compared with theorical clinical guidelines, permitting to evaluate the adherence of clinical practice with the hospital protocols. Moreover, these methods can highlight the typical process carried out in the hospital to manage a specific disease case, summarizing past clinical knowledge.

Patterns of care of patients with the same disease can be different in clinical practice. This variability must be considered during careflows discovery and presentation.

Some solutions are possible to mitigate the heterogeneous behavior of clinical processes, helping to synthetize the results discovered [28], also thanks to the use of probabilistic topic models [29].

The creation of events logs is a critical part of process mining, they describe a sequence of events and contain all the necessary information about clinical activities of interest (e.g. resources, types of event).

Each event in an event log is related to a timestamp, indicating the time reference and an id indicating the event group case. In healthcare, events are usually assigned to their patients, using the patient id as case.

In careflow mining three steps can be recognized:

- Preprocessing: where the clinical process is studied in order to create the best event log that describes events of interest.
- Discovery: in which the main algorithm performs careflows search and extraction from the event logs.
- Postprocessing: where additional information contained in the event log can be used to perform analysis on the careflows. Moreover, the result presentation is an important part of the method, because it permits the interpretation of careflows by users.

The solution proposed in this thesis is a CFM algorithm developed in [30] and already tested in [31]. This method finds frequent careflows in clinical and administrative records, reducing variability of the results with parameters that hinder the overgrowth of the mined careflows. It also provides the display of results with clear and simple graphical outputs. The resulting careflows can be used to stratify patients and recognize their clinical phenotypes [32].

1.4. Thesis Outline

The structure of the thesis is organized as follows:

Chapter 2 Materials and Methods.

This chapter describes the different data that has been considered.

Moreover, an introduction to methods and tools used in the course of the PhD is provided: the i2b2 platform as data warehouse; Mirth Connect as an ETL (Extract, Transform, Load) tool; Protégé and an NLP (natural language processing) pipeline applied to the extraction of information from pathology reports; the CFM (CareFlow Mining) algorithm to provide additional tool to oncologic research.

Chapter 3 Implementation.

The chapter reports the way in which the tools previously described has been applied to the Hospital Papa Giovanni XXIII scenario.

The chapter explains the ETL model used to populate the i2b2 data warehouse, the NLP application to the case of breast cancer pathology reports and the presentation of an i2b2 plugin designed for CFM purposes.

Chapter 4 Results.

The chapter reports the results that the implementation has produced. Starting from a description of the current data present in the i2b2 data warehouse and the studies that has been obtained thanks to this solution.

A validation of the NLP task is presented, studying the performance and the lack of the actual pipeline.

Regarding the CFM, the application of the algorithm in a similar scenario has been proposed as proof of concept, showing its capability to generate and discover novel aspects of the data.

Finally, **Chapter 5** Conclusions.

The chapter summarizes the outcome of the thesis, future perspectives and further developments.

Chapter 2

Materials and Methods

This chapter describes the different data source of the HPG23 and introductions to the informatics solutions and tools used in the course of the PhD: the i2b2 platform as data warehouse; Mirth Connect for the ETL task; Protégé and the NLP pipeline applied to the extraction of information from pathology reports; the CFM algorithm to provide additional tools to oncology research.

2.1. Source Datasets

The sources of data coming from the hospital were managed with the collaboration of the Hospital Information System (HIS) technicians and third-party Dedalus staff in the Papa Giovanni XXIII Hospital (HPG23), in Bergamo. For each data system of interest, a database view has been extracted to allow ETL:

- General demographic information about patients
- Hospitalizations data from the Hospital Discharge Register (HDR)
- Outpatients and ambulatorial activities
- Laboratory test and analysis
- Drugs
- Anatomical pathology reports

The start year is 2007, the date when the hospital started to collect data in a consistent way from all these different sources.

Demographics

A single database view was developed from HIS and shared. The data contains:

- Demographic hospital id
- Zip code of residence
- Hometown
- Death date
- Birth date
- Nationality
- Job Type
- Race
- Sex
- Date of last update of these demographic information

In this view, each row corresponds to a patient. The data collection process sometimes has errors related to duplicated patients, and periodically HIS performs a task of merging patients ids when corresponding to the same real patient.

2.1.1. Hospital Discharge Register (HDR)

The Hospital Discharge Register is online from 2007, but in 2012 the software application to manage the hospitalizations inside the hospital has changed:

- “Monitor” application: from 2007 to 2012
- “Galileo” application: from 2012

There is also a tiny overlap in the first semester of 2012, where hospitalizations were registered in one application or in the other, slowly transitioning to the newest. (Figure 1).

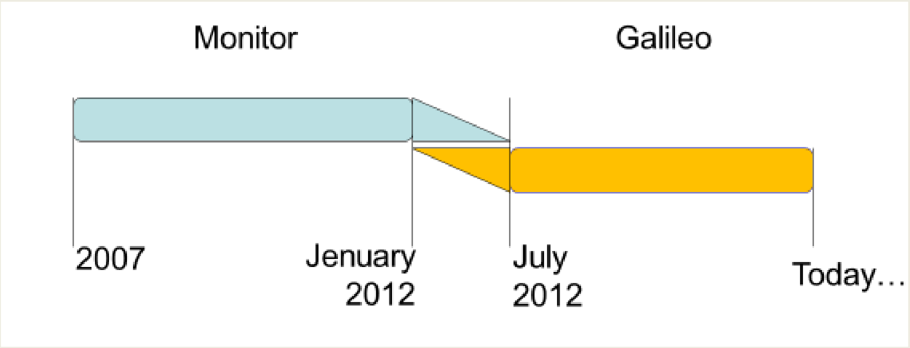


Figure 1 Hospitalizations applications transition.

Data were extracted from both software applications, and the two database views have very similar structure in terms of columns and information.

For each row of the views, a single hospitalization of a patient is exposed, with all the necessary columns to provide data about the admission date and ward, and the diagnosis and procedures the patients had undergone, using ICD9-CM code. Differences regard the tracking of hospital wards and the presence or not of the date of the diagnosis.

2.1.1.1. Monitor

The older HDR application was called "Monitor"; it contains data from 2007 to July 2012 and provides this data:

- Patient id
- Hospitalization id
- Initial ward id
- Initial ward name
- Discharge ward id
- Discharge ward name
- Hospitalization start date
- Hospitalization end date, date of dismission
- Code for discharge
- Hospital stay, in day
- Main Diagnosis ICD9 code
- Secondary Diagnosis 1 ICD9 code
- Secondary Diagnosis 2 ICD9 code
- Secondary Diagnosis 3 ICD9 code
- Secondary Diagnosis 4 ICD9 code
- Secondary Diagnosis 5 ICD9 code
- Main Diagnosis date
- Secondary Diagnosis 1 date
- Secondary Diagnosis 2 date
- Secondary Diagnosis 3 date
- Secondary Diagnosis 4 date
- Secondary Diagnosis 5 date
- Main Procedure ICD9 code
- Secondary Procedure 1 ICD9 code
- Secondary Procedure 2 ICD9 code
- Secondary Procedure 3 ICD9 code
- Secondary Procedure 4 ICD9 code
- Secondary Procedure 5 ICD9 code
- Main Procedure date
- Secondary Procedure 1 date
- Secondary Procedure 2 date

- Secondary Procedure 3 date
- Secondary Procedure 4 date
- Secondary Procedure 5 date

2.1.1.2. Galileo

The newer HDR application is called "Galileo" and contains data from January 2012 to nowadays and provide this data:

- Patient id
- Hospitalization id
- Actual ward id
- Actual ward name
- Initial ward id
- Initial ward name
- Previous ward id
- Hospitalization start date
- Hospitalization end date, date of discharge
- Code for discharge
- Main Diagnosis ICD9 code
- Secondary Diagnosis 1 ICD9 code
- Secondary Diagnosis 2 ICD9 code
- Secondary Diagnosis 3 ICD9 code
- Secondary Diagnosis 4 ICD9 code
- Secondary Diagnosis 5 ICD9 code
- Main Procedure ICD9 code
- Secondary Procedure 1 ICD9 code
- Secondary Procedure 2 ICD9 code
- Secondary Procedure 3 ICD9 code
- Secondary Procedure 4 ICD9 code
- Secondary Procedure 5 ICD9 code
- Main Procedure date
- Secondary Procedure 1 date
- Secondary Procedure 2 date
- Secondary Procedure 3 date
- Secondary Procedure 4 date
- Secondary Procedure 5 date

The Galileo application is currently used in the hospital and open hospitalizations are recorded, too; they are recognizable because they lack the day of discharge.

Moreover, the presence of the field for modality of discharge can provide the patient vital status at discharge.

2.1.2. Outpatients and ambulatorial activities

These records start from 2007 and keep track of the outpatient's visits. Usually more than one procedure is performed in a single outpatient visit. Since the database view describe each single procedure within a row, more rows are present for a single visit. The columns of the shared outpatient data are:

- Patient id
- Booking number
- Visit progress status
- Visit date
- Clinical Specialization id
- Clinical Specialization name
- Procedure Internal code
- Procedure Internal description
- Procedure SISS code
- Procedure SISS description

The services provided are saved using a double codification, the hospital internal code and the SISS code. The SISS (Sistema Informativo Socio Sanitario) is the regional health information system of the Lombardy region.

2.1.3. Laboratory Tests

These records start from 2007, all the laboratory test performed in the hospital are tracked. Laboratory test can also be requested during a patient hospitalization or ambulatory service, so two columns with this information can attribute a lab event to one of these bigger "container" events.

In each row of the view a single result of an analysis is exposed. The content is:

- Patient id
- Hospitalization id
- Booking number
- Lab request id
- Lab request date
- Analysis Sector internal code
- Analysis Sector internal name
- Analysis internal code
- Analysis internal name
- Result internal code
- Result internal name
- Result
- Result Unity of Measurement

- Minimum value of the Result to be considered in normal range
- Maximum value of the Result to be considered in normal range
- Result release date
- End of validity of this type of exam

"Lab request id" is a code that groups all the analysis requested for a patient in the same moment in the hospital.

Lab tests are internally organized in sectors, a hierarchy that divides and regroups similar analysis in the same sector. Each analysis routine can provide more than one result. All these concepts are tracked with internal code system.

Moreover, within the years analysis and result can be semantically modified, for example the unit of measurement can change (maintaining the same analysis and result code).

2.1.4. Drugs

Drugs records inside the hospital start from 2007 and are divided in two different views:

- Drugs administrated inside the hospital
- Drug delivery to patient, who will take the drug at home.

2.1.4.1. Drug administered inside the hospital

Drugs administration, as laboratory test, can be requested during a hospitalization or outpatient visit, so two columns with this information can attribute a drug event to one of these bigger "container" events.

Each row of this view describes a single administration of a medicine, and these columns are included:

- Patient id
- Event id
- Event type, if from outpatient visit or hospitalization main event
- Hospitalization id
- Booking number
- Cost center id
- Administration id
- Infusion administration id
- Administration route
- Event start date
- Event end date
- Therapy type, if singular administration or continue infusion o part of a schedule
- Drug internal code

- Drug name
- ATC code
- Prescription dosage
- Prescription dosage Unit of Measurement
- Equivalent converted dosage
- Equivalent converted dosage Unit of Measurement
- Equivalent milligrams of the administration
- Equivalent milliliters of the administration
- Equivalent International Unit of the administration
- Administration note
- Real administration start date
- Real administration end date
- Nominal administration duration
- Schedule, name of the specific pharmaceutical schedule this administration is integrated in
- Schedule diagnosis motivation
- Therapy stage, from initial to metastatic therapy line
- Actual cycle number of the schedule, number of times this entire schedule has been repeated
- Actual cycle start date
- Actual cycle note
- Nominal schedule administration day
- Actual schedule administration day
- Administration schedule sequence order of the day
- Prescription date
- Prescription notes
- Patient weight
- Patient height
- Body surface area
- Patient measurement date

As can be observed by the nature of the view created, great importance has been assigned to the pharmaceutical schedule which an administration can be part of. A pharmaceutical schedule indicates exactly which drugs, a patient must assume, specifying day number from the start of the schedule, drugs order within a day, quantity and duration expected.

2.1.4.2. Delivery to patient

Drug delivery to patients regards pharmaceutical prescriptions planned to be directly given to patients, who will take the drug later. Main differences with pharmaceutical administration are the lack of hospitalization link code and the absence of pharmaceutical schedule.

Each row in the view corresponds to a single package of drugs delivered to the patient:

- Patient id
- Event id
- Booking number
- Theorical delivery date
- Actual delivery date
- Prescribed drug internal code
- Prescribed drug name
- ATC code
- Prescribed drug total quantity
- Prescribed drug Unit of Measurement
- Single assumption dosage
- Frequency assumption of the dosage
- Days covered by the quantity prescribed
- Posology notes
- Delivered drug internal code
- Delivered drug name
- Delivered drug total quantity
- Delivered drug Unit of Measurement
- Delivery status

Here some columns are purposely used to describe pharmaceutical assumption modalities and posology. Also, the prescribed and delivered products can be different, but the drug type and ATC codes are the same.

2.1.5. Anatomical Pathology Reports

Anatomic pathology reports are electronically available since November 2008. A report describe analysis of one or more specimen of a patient. This view makes available hand-written text of the report divided in section for each column.

Each row, however, does not contains only report textual information, but also name and codes relative to the specimen analyzed.

Often, one pathology report contains more than one sample analysis. For this reason, a single report can be repeated on more rows.

Columns provided in the view describe:

- Patient id
- Report id
- Report type, if Histological, Autopsy or others
- Report date
- Pathologist
- Hospitalization id
- Specimen collection date
- Specimen arrival date
- Specimen description

- Specimen morphology SNOMED code
- Specimen topography SNOMED code
- Specimen procedure SNOMED code
- Anamnestic information section
- Clinical information section
- Macroscopic description section
- Microscopic description section
- Diagnosis section
- Preliminary (extemporaneous) diagnosis section; it is used when the specimen is collected and analyzed during a surgery procedure and report will guide the following part of the surgery
- Comment
- Addendum to the report, complements to the exams of the report

SNOMED codes are used not only for morphology, main descriptor of the tumor type and behavior, but also for providing information about topography and procedure, describing respectively the locus and the procedure for collecting the specimen. Unfortunately, these fields are largely misused (commonly empty or filled with generic and not informative terms of SNOMED terminology).

2.2. i2b2

The reuse of Electronic Health Records to support clinical research can be very valuable [9]. To maximize its potential, it is necessary to define methods and tools to provide all information to researcher in a consistent and easy-to-be accessed way.

This goal is not a trivial task, because it is not limited to query data about individual patients. These methods need to allow queries across multiple patients and are therefore dependent on the standards and formats used to store the data. Moreover, these queries must consider privacy concerns.

I2b2 is known to be a useful tool for the clinical researchers, with a solid data model at its foundations, providing a software platform capable of integrating clinical records and research data [9].

The main components of i2b2 are a data warehouse and a data management service.

2.2.1. i2b2 Data Warehouse

The i2b2 data warehouse is named CRC (Clinical Research Chart) and it is based on the “star schema” data model [11]

In this model, a central fact table is present, the “Observation Fact” table, and four connected dimension tables provide information and represent patient, care providers, visits and medical concepts Figure 2.

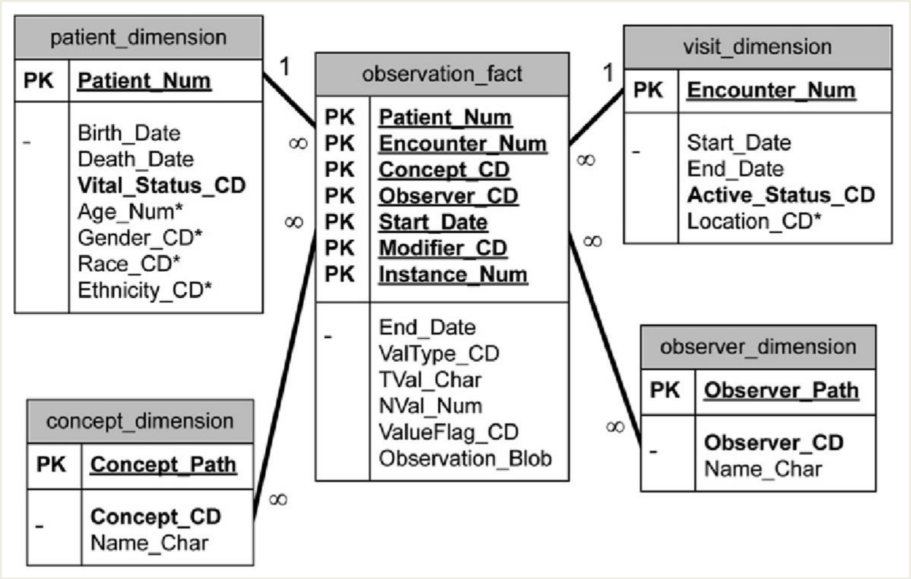


Figure 2 In this representation of the star schema, boxes are the database tables and rows are the columns of these tables [9].

With this model implemented, the central table contains all the information about patients care events with codes and ids, relying on surrounding dimension tables to contain the description of these codes.

One particular dimension table can contain information about how a group of codes are organized, forming a hierarchy structure. This is a very useful and important addition, because these hierarchies can be used to help the user create the queries, interacting with these "i2b2-ontology" [11] rather than directly with the data.

In the observation fact table, a row describes a single fact or event happened to a patient. In a single row are present info about the event:

- patient experiencing such event;
- encounter context, for example a hospitalization or a visit;
- concept object of the event itself (disease, surgery, lab test, etc...);
- start and stop date of the event;
- event observer, for example clinician, laboratory expert;
- modifier can give some additional information about the concept (typical example is with drug administration: the drug is the concept, but additional information can be the administration route of that drug).

The event, explicated by the concept, can be a code, a specific procedures or test, but can also represent genetic data [33]. The choice of describing a concept in an entire row instead of squeeze it in a column is known as entity-attribute-value (EAV) model [34]. This model used in combination with the star schema can improve the efficiency of query data using an index build on the concept column of the observation fact table.

2.2.2. i2b2 Data Management Services

The data management services are composed by server-side software modules (the "hive") and client-side applications (query tools).

2.2.2.1. i2b2 Hive

The components of server software are called "cells" and together they form the i2b2 "hive". Each cell is a web-service that communicates with the other cells through XML messages following REST standard over HTTP [35]. Main i2b2 cells, that compose core functionalities of the i2b2 hive, are: file repository, data repository, ontology management, identity management and project management (Figure 3).

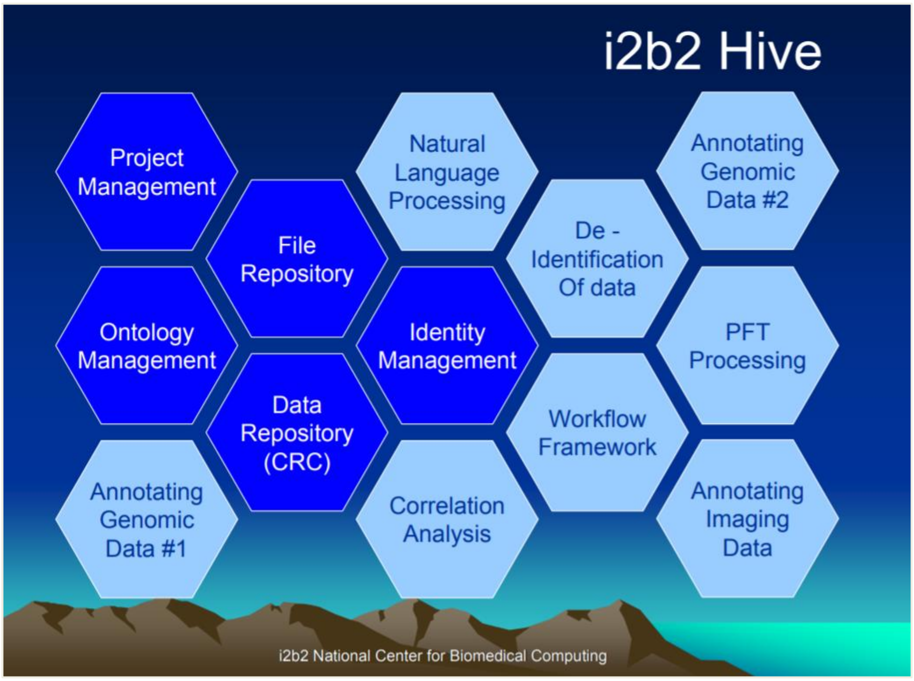


Figure 3 i2b2 hive structure [8]

It has been previously described how the CRC cell contains EHR data from hospital, typically loaded using an ETL task that transforms the row data of the HIS to i2b2 observations. In this cell, data are exposed to queries performed by users.

Every user accesses the system thanks to the project management cell, which manages its password, authorizations, preferences and accessible projects.

The ontology cell contains all the structured terms used as concepts in the observation fact table, allowing to create queries on patient events.

Another important cell is the identity management. Information in the identity management protects the privacy and anonymity of the patients by substituting the real patient id with an internal anonymized patient id that will be used by the other cells of the hive.

2.2.2.2. i2b2 client

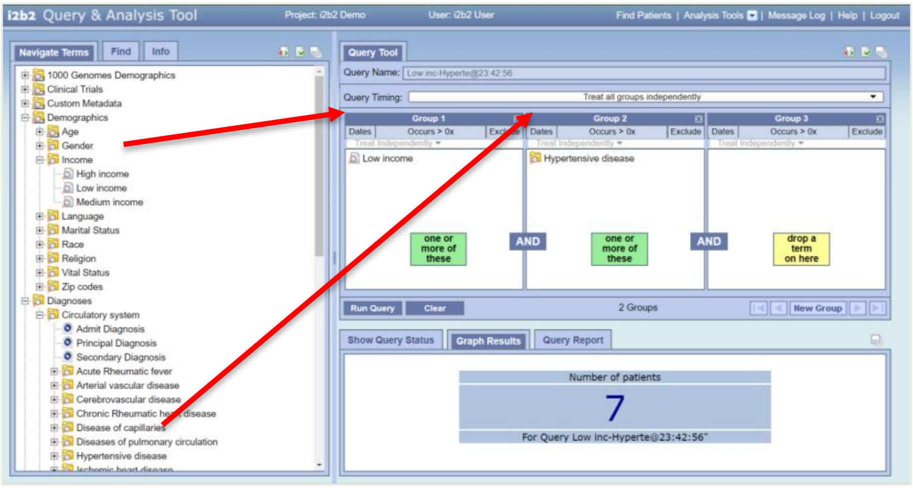
Two client-side applications are available, a web- browser access which leverages on JavaScript and a desktop Java application, the workbench, based on the Eclipse platform [36].

These tools allow the user to create query and investigate the i2b2 data using a drag and drop approach, enabled by the existence of ontology of concepts describing the event of the observation fact table.

More than one concept can be used in a single query and combined with simple Boolean AND/OR operators (Figure 4). The result of a query is a pool of patients that match the requested concepts. Each query result is stored, and the patient groups could be later used as input for additional modules, known as i2b2 plugins, to perform further analysis.

More complex query types can be constructed using this interface. A particular window permits to use the same concepts of the taxonomies to define sequence of events, controlling relative time associations between them. The result will show how many patients in the database has such requested events pattern.

As a measure of privacy, all patients' related data are anonymous. Furthermore, average user of the tool cannot have the actual pool of patient, neither their anonymous ids, but can only see aggregated results (i.e. the number of patients) in the queried result. Users with low permission will also see this number as obfuscated by the addition of a random number.



Query Name: Low inc-Hyperte@23:42:56

Query Timing: Treat all groups independently

Group 1: Low income

Group 2: Hypertensive disease

Number of patients

7

For Query Low Inc-Hyperte@23:42:56"

Figure 4 i2b2 web browser query tool [8]. Metadata items are dragged from the "Navigate Terms" panel on the left. The query is built in the Venn-diagram-like panels on the right. Concept dragged into the same panel are combined using OR logic, instead concepts in different panels use AND logic.

2.3. Mirth-Connect

Mirth Connect is part of a suite of software and tools that aim to address interoperability in healthcare domain. Mirth has been developed to simplify and organize the messaging task in healthcare, providing interfaces and translating automatically information provided in a data format into another one. It can act as a message broker or as an ETL (extract, transform, load) tool [37].

At its core there is the concept of "channel", a modular structure that is able to read information from various type of sources, manipulate and transfer the data to many destinations [38]. It has scalability in terms of volumes of data and due to its open source nature, it has low licensing costs [39] and platform independence.

This tool is currently owned by the Nextgen Connected Health [40], which changed its name into NextGen Connect Integration Engine and the latest version is 3.8.0.

Mirth is a standalone software, so no container is needed (Tomcat, Glassfish etc.). The only requirement is a Java installation with Oracle JRE/JDK (also OpedJDK is supported from Mirth 3.7).

Mirth Connect relies on a Database in order to work, memorize the configurations and storing messages. The supported Database types are PostgreSQL, MySQL, Oracle and SQL Server. It is also possible to install Mirth even without one of these databases, using an embedded Apache Derby database, even if this choice is suggested only for development and testing, not for production purposes.

2.3.1. Mirth General Design

Mirth capability to read data from a source, working on them, changing their formats and expose the results to multiple destination is enabled by the channel structure. The chunk of data that pass through this structure is referred as a message.

A channel comes along as an interface that order and automate the operations and act as aggregation of functionalities (Figure 5).

The sub-structure present in a channel is the connector. For each channel a single source connector is provided. This is where the connection with the data source is configured, incoming data are read, and a message is created.

Once the data has left the source connector, the channel proceeds with another connector type that manage the exit of information from the channel: the destination connectors can be more than one, matching the number of desired destinations and splitting the outgoing message of the source connector for each destination.

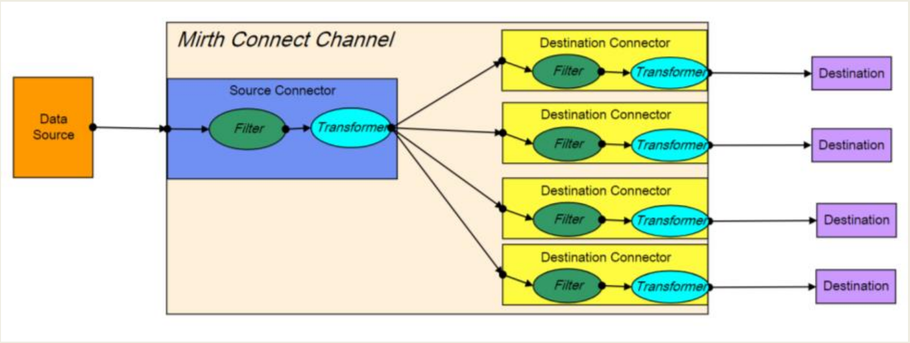


Figure 5 Mirth Channel layout [41].

Traveling inside a connector, data can be filtered out accordingly to a set of rules, requiring a feature to be present or a variable to match certain values. After that, one or more transformation step can change, shrink or enrich the content or the structure of the message. These filter and transformer are configurable individually for each present connector, both source and destination type.

The overall path of a message in a channel is described in Figure 6 , with the representation of other functionalities, like the attachment handler, the response system and the pre/post processor scripts.

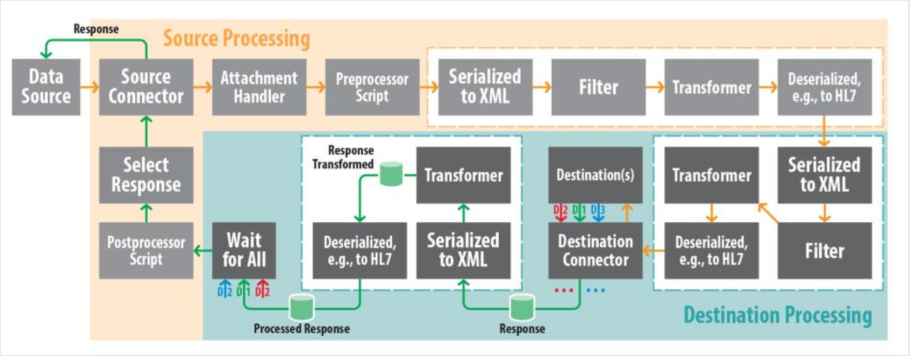


Figure 6 Mirth message path inside a channel [40]. At the beginning, the processes inside the source connector can be seen, followed by processes of the destination(s). Each blank box refers to a group of serialization-filter-transformer-deserialization steps. Other functionalities are present: attachment handler extracts and/or stores any attachment data at the very beginning of the channel; preprocessor script can influence the initial source message form or content before the serialization process; response from the external destinations can be accepted back and processed; postprocessor script can set up the overall response of the channel to the external data source.

The message is stored at many levels of each connector, so that each processing step corresponds to different state of the message:

- Raw: initial state of the message at the start of a connector;
- Processed Raw (only source): state of the message after the preprocessor script;
- Transformed: this is the serialized internal representation of the message, which exists only after the optional filter or transformer;
- Encoded: after the deserialization, this is the state of the message at the end of a connector. The encoded message from the source connector will become the raw message of the destination(s);
- Sent: the final message built at the end of each destination connector and sent to the outbound system;
- Response: Mirth Connect also developed a response system for both external source and destination:
- Each destination connector gets a message from the outbound system as a response to the previous sent message;
- After all the destination processing (Figure 6) a response message is sent back to the originating system;
- Response Transformed: The response from outbound system is treated like a message, with the only difference that it is not subject to a filter step. This transformed state corresponds to the serialized internal representation of the response, which exists only after the optional transformer;
- Processed Response: after the deserialization, this is the state of the response as it exits the destination connector.

In Mirth, connectors are very versatile and capable of reading/writing data using various protocols, interacting with file from a directory, Database connection, SOAP web service, TCP/IP socket, etc. Moreover, each connector is able of decoding/encoding useful standards and Data Types, for example allowing to move from a message in HL7 V2 to HL7 V3 standard.

This procedure of decoding messages from various formats is called serialization, and map the information and, more important, the structure of information to an internal data structure (XML or JSON).

This permits also to work using the interface, which internally relies on the structured nature of the serialized message, allowing dragging and dropping data transformation blocks.

Mirth Connect also allows direct JavaScript programming, with scripts that can totally drive the serialization/deserialization procedures and the filter and transformer steps.

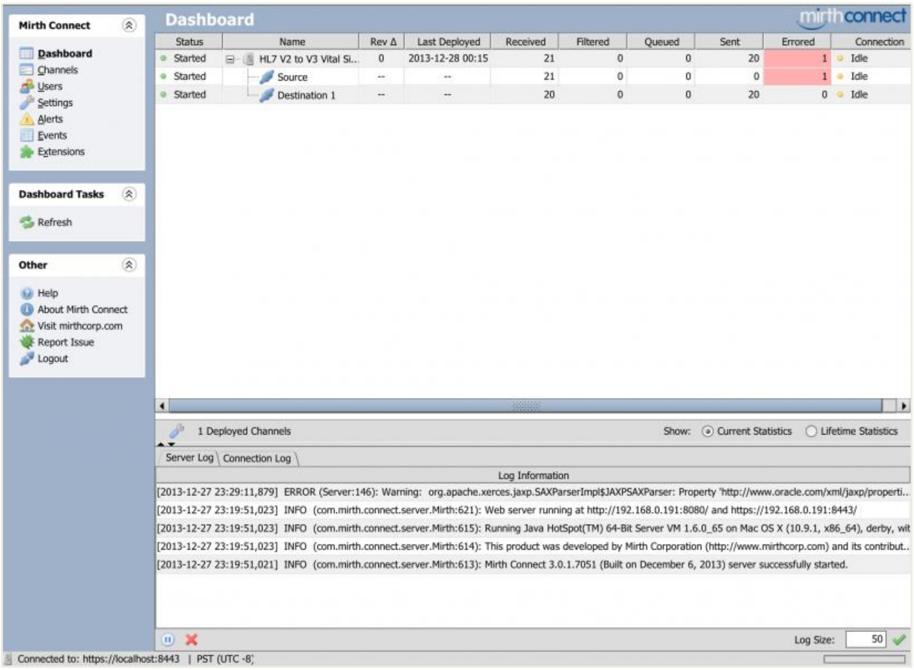
Moreover, each channel can have access to the “Code Template Libraries”, a space in mirth where JavaScript custom functions can be written and organized.

2.3.2. Mirth Interface

The user interface (GUI) is known as “Mirth Connect Administrator” and it can be used to manage the channels and all their configurations.

2.3.2.1. Dashboard

The dashboard view (Figure 7) handles deployed channels, showing the status, name and other information about messages of each channel. Here a channel can have a started, paused or stopped status.



Status	Name	Rev Δ	Last Deployed	Received	Filtered	Queued	Sent	Errored	Connection
Started	HL7 V2 to V3 Vital Si...	0	2013-12-28 00:15	21	0	0	20	1	Idle
Started	Source	--	--	21	0	0	0	1	Idle
Started	Destination 1	--	--	20	0	0	20	0	Idle

```
Log Information
[2013-12-27 23:29:11,879]  ERROR (Server:146): Warning:  org.apache.xerces.jaxp.SAXParserImpl$JAXPSAXParser: Property 'http://www.oracle.com/xml/jaxp/properti..
[2013-12-27 23:19:51,023]  INFO  (com.mirth.connect.server.Mirth:621): Web server running at http://192.168.0.191:8080/ and https://192.168.0.191:8443/
[2013-12-27 23:19:51,023]  INFO  (com.mirth.connect.server.Mirth:615): Running Java HotSpot(TM) 64-Bit Server VM 1.6.0_65 on Mac OS X (10.9.1, x86_64), derby, wit
[2013-12-27 23:19:51,023]  INFO  (com.mirth.connect.server.Mirth:614): This product was developed by Mirth Corporation (http://www.mirthcorp.com) and its contribut..
[2013-12-27 23:19:51,021]  INFO  (com.mirth.connect.server.Mirth:613): Mirth Connect 3.0.1.7051 (Built on December 6, 2013) server successfully started.
```

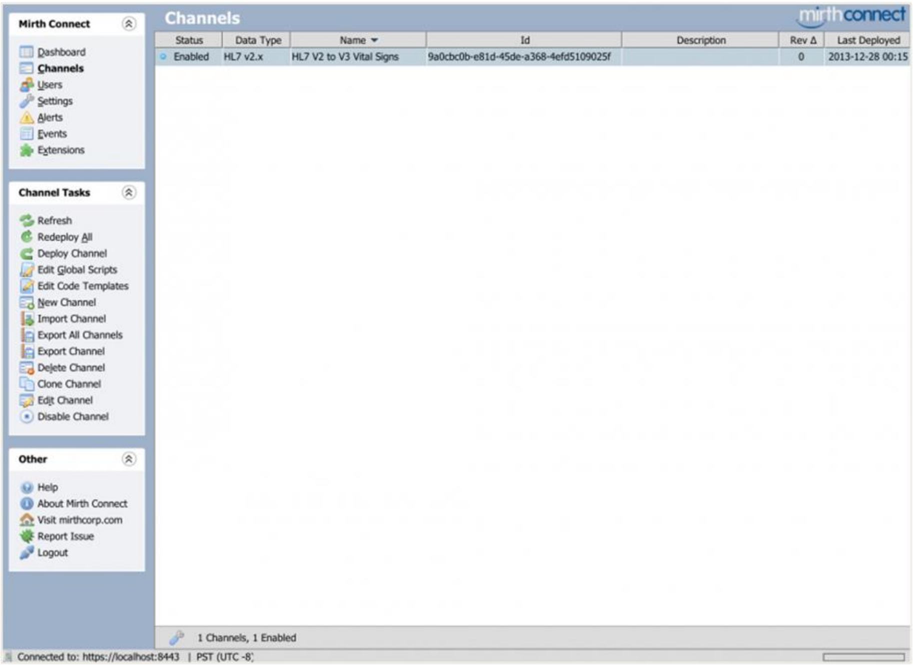
Figure 7 Mirth dashboard [37] The top section shows the channels that are currently deployed, their status, their name, any changes since the start of the server, last date of deployment, number of messages received, number filtered, number queued, number sent, any errors or alerts, and the status of the connection

A started channel works normally by processing inputs and outputs, while a stopped channel does not accept any incoming message. Pause is a halfway status between previous ones: a paused channel accepts incoming messages and puts them in a queue but does not manipulate them until it is started.

In the bottom section, logs information is displayed, reporting server and channel activities. The server log can also function as first step for debug possible errors. Connection log reports the transition of messages inside channels.

2.3.2.2. Channels

The “Channels” view page (Figure 8) is similar to the dashboard view previously presented Here, channels are created and edited or imported and exported. Once a channel status is enabled, it can be deployed, and start to be visible also in the dashboard view.



Status	Data Type	Name ▾	Id	Description	Rev Δ	Last Deployed
● Enabled	HL7 v2.x	HL7 V2 to V3 Vital Signs	9a0cbc0b-e81d-45de-a368-4efd5109025f		0	2013-12-28 00:15

1 Channels, 1 Enabled

Connected to: https://localhost:8443 | PST (UTC -8)

Figure 8 Mirth channels view [37].

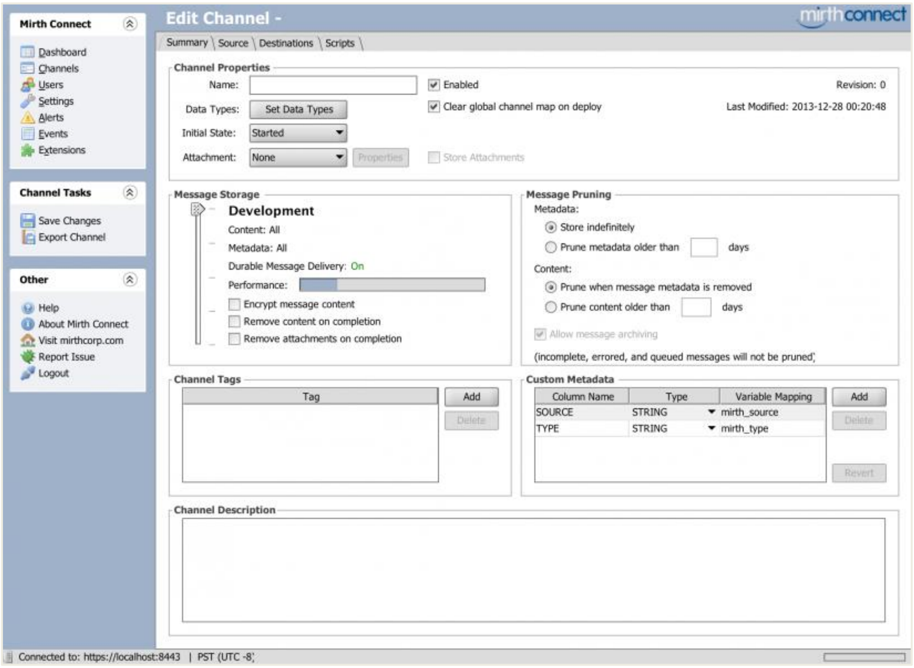
On the right are listed all the channels, both enabled and disabled. More details are provided for each channel respect the dashboard view like name, unique id, description and data type expected from the external source.

Most of the work inside Mirth takes place inside channels, along the global scripts of the server and the code templates (library of custom JavaScript functions), also reachable from this view.

To configure a channel in every aspect, the “Edit Channel” option is available. In the edit channel view (Figure 9) four horizontal tabs are provided. In each tabs many settings of the channel can be decided:

- “Summary Tab” includes general information about the channel;
- “Source Tab” specifies how the channel receives its data;
- “Destination Tab” specifies where to send the output message for each destination connector created;

- “Script Tab” allows to specifying JavaScript Pre/Post processor scripts (Figure 6) and scripts to run when the channel is deployed and undeployed.



Edit Channel -

Channel Properties

Name:

☑ Enabled

Revision: 0

Data Types:

☑ Clear global channel map on deploy

Last Modified: 2013-12-28 00:20:48

Initial State: Started

Attachment: None

☐ Store Attachments

Message Storage

Development

Content: All

Metadata: All

Durable Message Delivery: On

Performance:

☐ Encrypt message content

☐ Remove content on completion

☐ Remove attachments on completion

Message Pruning

Metadata:

● Store indefinitely

○ Prune metadata older than days

Content:

● Prune when message metadata is removed

○ Prune content older than days

☑ Allow message archiving

(incomplete, errored, and queued messages will not be pruned)

Channel Tags

Tag

Custom Metadata

Column Name	Type	Variable Mapping
SOURCE	STRING	mirth_source
TYPE	STRING	mirth_type

Channel Description

Connected to: https://localhost:8443 | PST (UTC -8)

Figure 9 Mirth edit channel view [37] begins with the summary tab.

An example usage of Mirth Connect about the creation of a channel to convert an HL7 V2.5 message to the equivalent in HL7 V3. In case of static functionality, a complete channel can be created just using the drug and drop function while in case of dynamic functionality few code lines in JavaScript are required.

2.3.2.3. Channel Summary

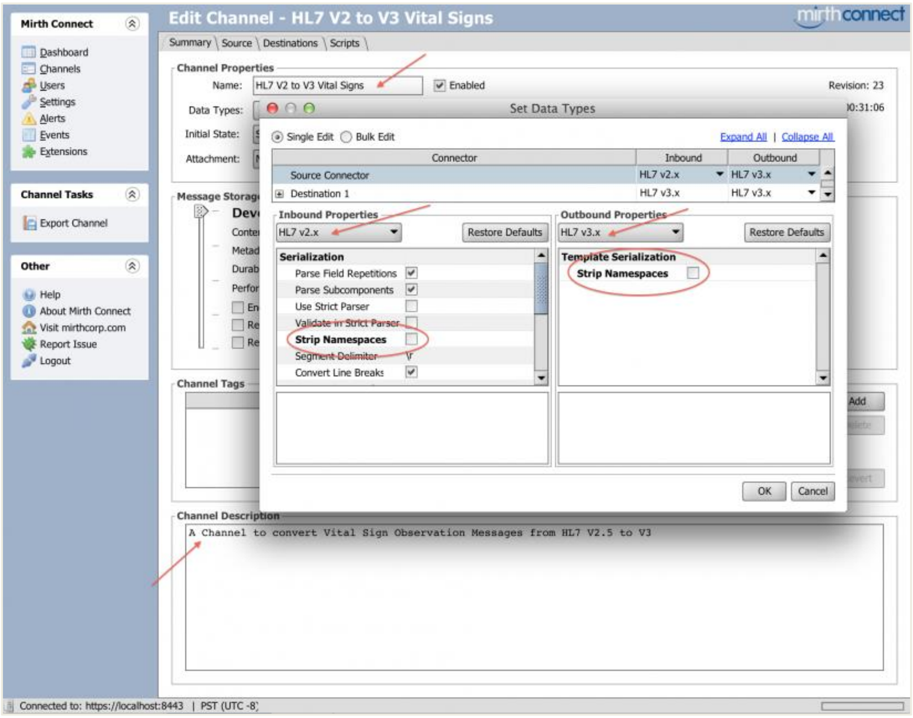
In the summary tab of a channel (Figure 9) general data of the channel can be modified, as channel name, tag and description.

The preferences about message storage can tune the various version of a message as it travels inside the processing steps of a channel (Figure 5) and some privacy setting about the message content.

Message pruning decides how long should the messages remain in the database before being purged.

Custom metadata can contain some useful variable inside the channel scope.

Last, but not least, channel properties contain the actions upon attachments, preferred channel status on deploying and the important “Data Type” button: it permits to set the data types of the messages as they travel inside the whole channel, from the source connector to the several destination connectors (Figure 10).



Name: HL7 V2 to V3 Vital Signs ☑ Enabled

Revision: 23

Data Types:

Initial State:

Attachment:

● Single Edit ○ Bulk Edit

Connector	Inbound	Outbound
Source Connector	HL7 v2.x	HL7 v3.x
Destination 1	HL7 v3.x	HL7 v3.x

Inbound Properties: HL7 v2.x

Serialization

Parse Field Repetitions	☑
Parse Subcomponents	☑
Use Strict Parser	□
Validate in Strict Parser	□
Strip Namespaces	□
Segment Delimiter	\r
Convert Line Breaks	☑

Outbound Properties: HL7 v3.x

Template Serialization

Strip Namespaces □

Channel Description

```
A Channel to convert Vital Sign Observation Messages from HL7 V2.5 to V3
```

Connected to: https://localhost:8443 | PST (UTC -8)

Figure 10 Mirth channel data types example [37].

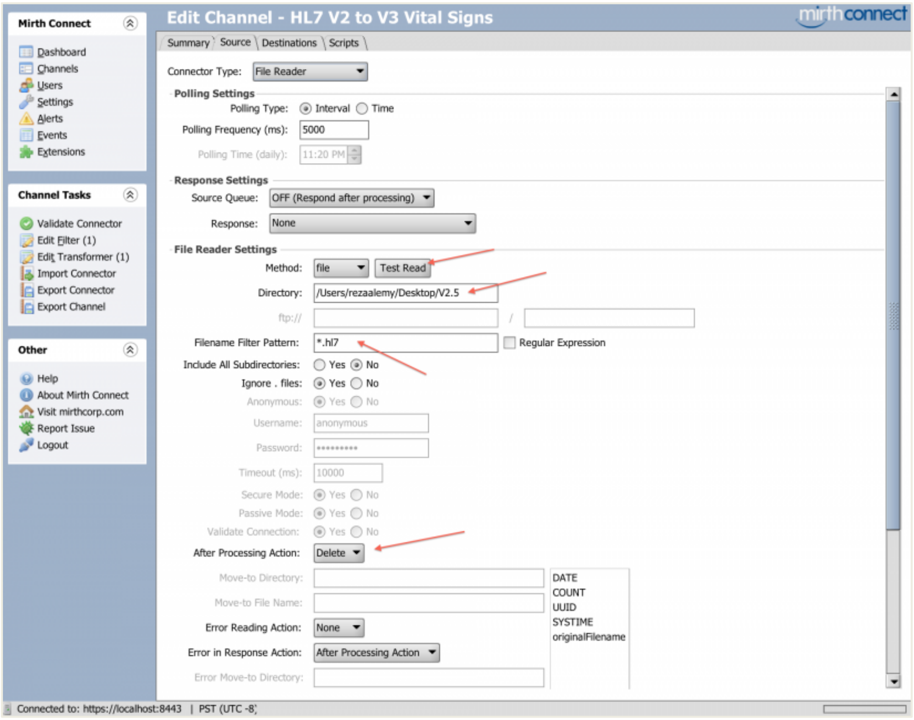
In the Channel "Convert Vital Sign HL7v2x to HL7v3", the data type for "Source Inbound" is HL7 V2.x. "Source Outbound" is set to HL7 V3, as well as the matching "Destination 1 Inbound" (the output of the source is the input of all the destinations). A Description is specified for the Channel. In all inbound and outbound properties, the "Strip Namespaces" checkbox is cleared.

Available data types in NextGen Connect Integration Engine are:

- Delimited Text Data
- Raw Data
- JSON Data
- XML Data
- DICOM Data
- NCPDP Data
- EDI / X12 Data
- HL7 v2.x Data
- HL7 v3.x Data

2.3.2.4. Channel Source

In Figure 11 the settings of input data in the channel source tab is shown. From the top of the tab the user needs to select a "Connector type" and its options.



Edit Channel - HL7 V2 to V3 Vital Signs

Connector Type: File Reader

Polling Settings
Polling Type: ● Interval ○ Time
Polling Frequency (ms): 5000
Polling Time (daily): 11:20 PM

Response Settings
Source Queue: OFF (Respond after processing)
Response: None

File Reader Settings
Method: file
Directory: /Users/rezaalemy/Desktop/V2.5
ftp:// /
Filename Filter Pattern: *.hl7 □ Regular Expression
Include All Subdirectories: ○ Yes ● No
Ignore . files: ● Yes ○ No
Anonymous: ● Yes ○ No
Username: anonymous
Password: *********
Timeout (ms): 10000
Secure Mode: ● Yes ○ No
Passive Mode: ● Yes ○ No
Validate Connection: ● Yes ○ No
After Processing Action: Delete
Move-to Directory:
Move-to File Name:
Error Reading Action: None
Error in Response Action: After Processing Action
Error Move-to Directory:

Connected to: https://localhost:8443 | PST (UTC -8)

Figure 11 Mirth channel source tab example [37].

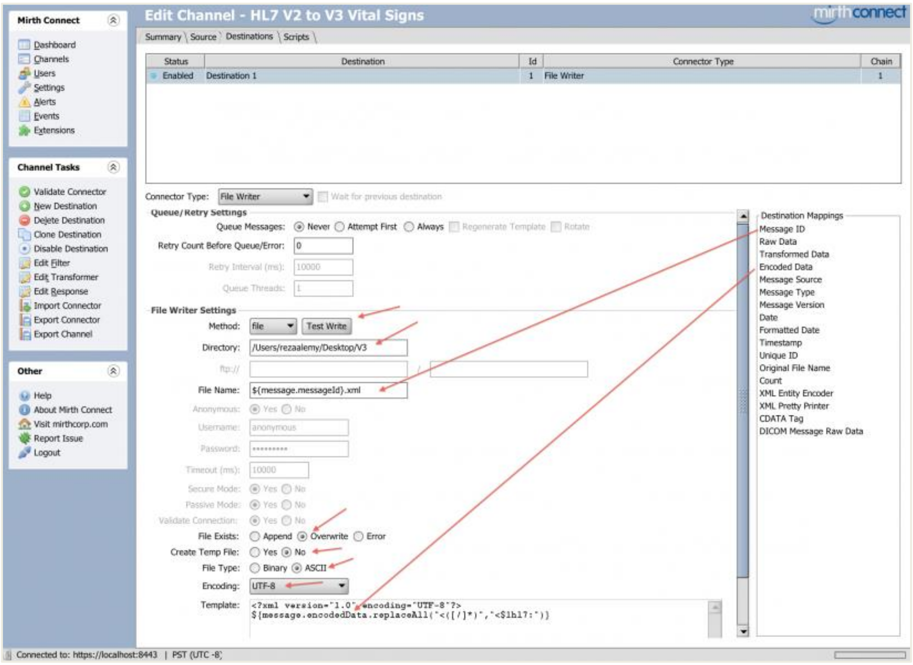
The channel in the example starts from the addition of HL7 V2.5 files in a directory named "/ Users/ rezaalemy/ Desktop/ V2.5". It follows the selection of the "File Reader" connector in the Source Tab, the selection of a file pattern and the check box to delete the file after reading and with the permission from the connector. It is strongly suggested to validate of the source connector.

Different connectors are available in Mirth:

- Channel Reader
- Database Reader
- File Reader
- JavaScript Reader
- DICOM Listener
- HTTP Listener
- JMS Listener
- TCP Listener
- Web Service Listener

2.3.2.5. Channel Destination

The output message is specified in the destination tab (Figure 12). In the top part of the tab, user can set the destination for a channel, the destination mapping, the connector type and its status enabled/disabled.



Edit Channel - HL7 V2 to V3 Vital Signs

Status	Destination	Id	Connector Type	Chain
Enabled	Destination 1	1	File Writer	1

Connector Type: File Writer ☐ Wait for previous destination

Queue/Retry Settings

Queue Messages: ● Never ○ Attempt First ○ Always ☐ Regenerate Template ☐ Rotate

Retry Count Before Queue/Error: 0

Retry Interval (ms): 10000

Queue Threads: 1

File Writer Settings

Method: file

Directory: /Users/rezaalemy/Desktop/V3

ftp://

File Name: ${message.messageId}.xml

Anonymous: ● Yes ○ No

Username: anonymous

Password: •••••••••

Timeout (ms): 10000

Secure Mode: ● Yes ○ No

Passive Mode: ● Yes ○ No

Validate Connection: ● Yes ○ No

File Exists: ○ Append ● Overwrite ○ Error

Create Temp File: ○ Yes ● No

File Type: ○ Binary ● ASCII

Encoding: UTF-8

Template:

```
<?xml version="1.0" encoding="UTF-8"?>
${message.encodedData.replaceAll("<([/]*)","<$1hl7:")}
```

Connected to: https://localhost:8443 | PST (UTC -8)

Figure 12 Mirth channel destinations tab example [37].

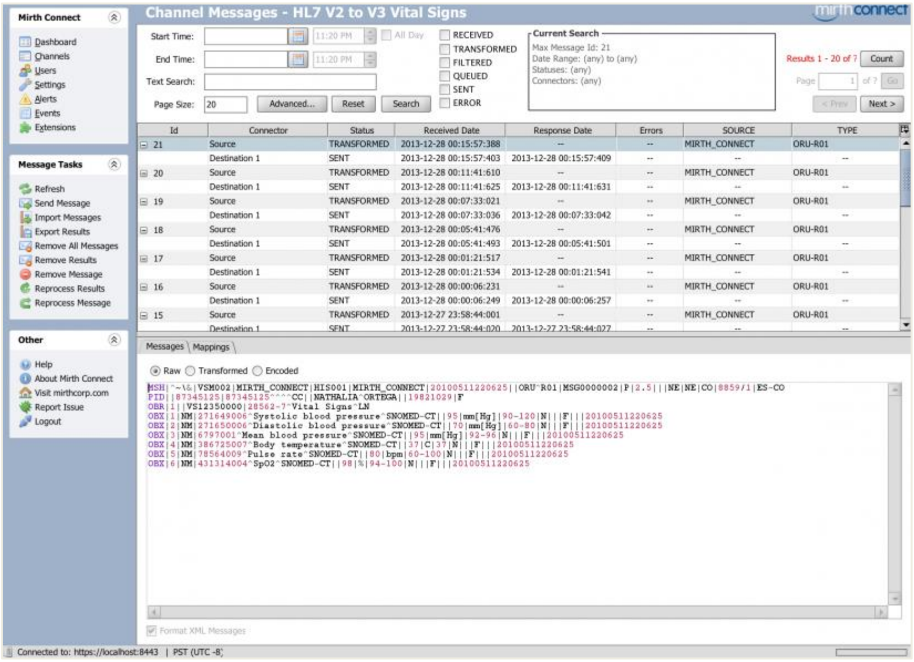
The file connector is created to point to the output directory “/ Users/ rezaalemy/ Desktop/ V3” The file name is composed by the message id variable (${message.id}) and the xml extension. Transformed data are obtained dragging Encoding data from the right column to the template box. The ‘Overwrite’ bottom could be selected to prevent the existence of two files with the same message id.

Available other connector types in this tab are:

- Channel Writer
- DICOM Sender
- Database Writer
- Document Writer
- File Writer
- HTTP Sender
- JMS Sender
- JavaScript Writer
- SMTP Sender
- TCP Sender
- Web Service Sender

2.3.2.6. Channel Monitoring

Statistics and information about the transformation, the encoding and the mapping of messages passing through Mirth Channels are shown in Dashboard. Figure 13 shows a received message. The Error tab shows errors if present. The Metadata shows the unique id assigned for the server, the correlation and the message.



Current Search
Max Message Id: 21
Date Range: (any) to (any)
Statuses: (any)
Connectors: (any)

Id	Connector	Status	Received Date	Response Date	Errors	SOURCE	TYPE
21	Source	TRANSFORMED	2013-12-28 00:15:57:388	--	--	MIRTH_CONNECT	ORU-R01
	Destination 1	SENT	2013-12-28 00:15:57:403	2013-12-28 00:15:57:409	--	--	--
20	Source	TRANSFORMED	2013-12-28 00:11:41:610	--	--	MIRTH_CONNECT	ORU-R01
	Destination 1	SENT	2013-12-28 00:11:41:625	2013-12-28 00:11:41:631	--	--	--
19	Source	TRANSFORMED	2013-12-28 00:07:33:021	--	--	MIRTH_CONNECT	ORU-R01
	Destination 1	SENT	2013-12-28 00:07:33:036	2013-12-28 00:07:33:042	--	--	--
18	Source	TRANSFORMED	2013-12-28 00:05:41:476	--	--	MIRTH_CONNECT	ORU-R01
	Destination 1	SENT	2013-12-28 00:05:41:493	2013-12-28 00:05:41:501	--	--	--
17	Source	TRANSFORMED	2013-12-28 00:01:21:517	--	--	MIRTH_CONNECT	ORU-R01
	Destination 1	SENT	2013-12-28 00:01:21:534	2013-12-28 00:01:21:541	--	--	--
16	Source	TRANSFORMED	2013-12-28 00:00:06:231	--	--	MIRTH_CONNECT	ORU-R01
	Destination 1	SENT	2013-12-28 00:00:06:249	2013-12-28 00:00:06:257	--	--	--
15	Source	TRANSFORMED	2013-12-27 23:58:44:001	--	--	MIRTH_CONNECT	ORU-R01
	Destination 1	SENT	2013-12-27 23:58:44:020	2013-12-27 23:58:44:027	--	--	--

```
MSH|^~\&|VSM002|MIRTH_CONNECT|HIS001|MIRTH_CONNECT|20100511220625||ORU^R01|MSG0000002|P|2.5|||NE|NE|CO|8859/1|ES-CO
PID||87345125|87345125^^^CC||NATHALIA^ORTEGA||19821029|F
OBR|1||VS12350000|28562-7^Vital Signs^LN
OBX|1|NM|271649006^Systolic blood pressure^SNOMED-CT||95|mm[Hg]|90-120|N|||F|||20100511220625
OBX|2|NM|271650006^Diastolic blood pressure^SNOMED-CT||70|mm[Hg]|60-80|N|||F|||20100511220625
OBX|3|NM|6797001^Mean blood pressure^SNOMED-CT||95|mm[Hg]|92-96|N|||F|||20100511220625
OBX|4|NM|386725007^Body temperature^SNOMED-CT||37|C|37|N|||F|||20100511220625
OBX|5|NM|78564009^Pulse rate^SNOMED-CT||80|bpm|60-100|N|||F|||20100511220625
OBX|6|NM|431314004^SpO2^SNOMED-CT||98|%|94-100|N|||F|||20100511220625
```

Figure 13 Mirth channel messages [37].

2.3.3. Mirth Usage

Mirth Connect is a tool for data interoperability in healthcare. The user just needs to understand the basis of the application and without deep programming knowledge to use the software. Furthermore, when necessary, many levels of complexity are allowed by the application very open and flexible, with JavaScript and java programming skills. Finally, for each kind of problem with Mirth the community is an invaluable source of information.

2.4. Protégé

Protégé software and documentation are available at [23]. It is an open source platform providing tools to construct ontologies and domain models describing concepts and relationships between them [42]. The software supports OWL (Web Ontology Language) the last development of the World Wide Web Consortium (W3C) in standard ontology language.

Currently the Protégé product lineup is differentiated in two cross-compatible offerings:

- Protégé Desktop: an ontology editing environment that supports OWL 2 standard, with a useful interface and features of visualization, refactoring and reasoners for inconsistencies;
- Web Protégé: a web development environment for ontologies, thought to enhance collaboration with features of shared work, change tracking and revision history.

2.4.1. Protégé entities

Different entities interact in an OWL ontology. The domain objects in which we are interested are named Individuals (Figure 14). Given two individuals, the relationship between them is expressed as a Property. In the example shown in Figure 15, Matthew is the brother of Gemma and Matthew lives in England. A property can have an inverse property; for example, Gemma is the sister of Matthew.

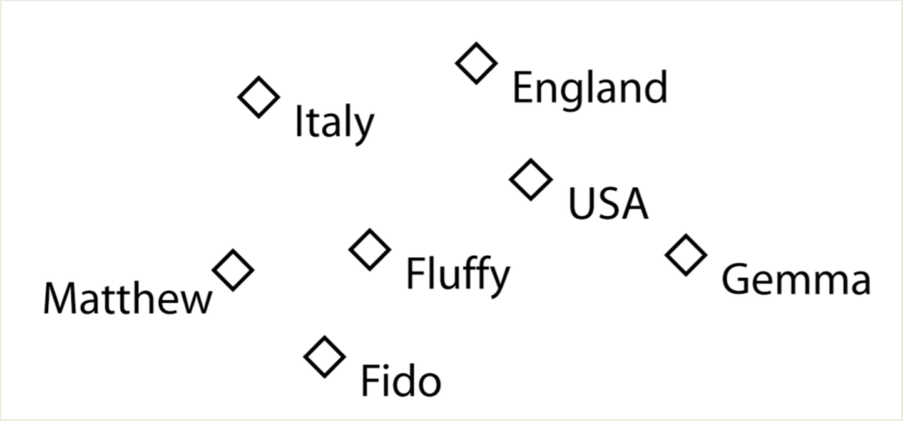


Figure 14 Example of representation of individuals in Protégé [42].

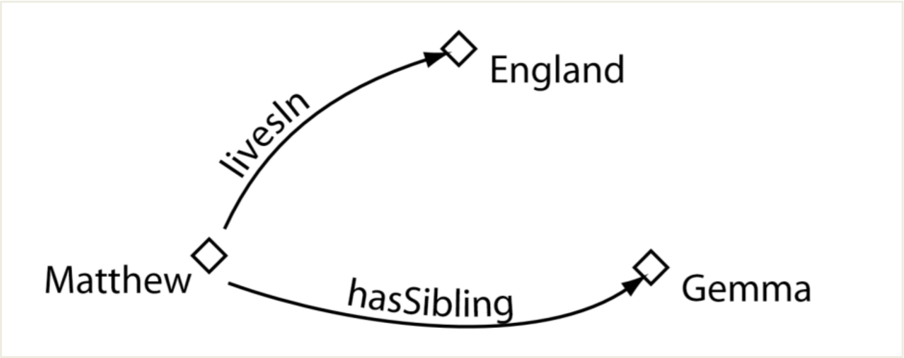


Figure 15 Example of representation of properties between couple of individuals [42].

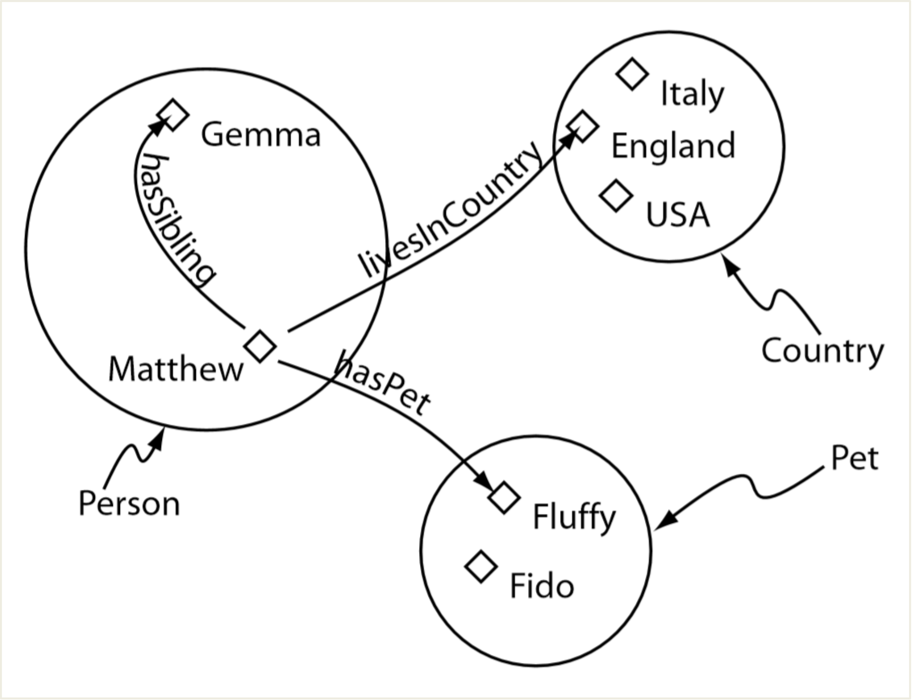


Figure 16 Example of representation of OWL classes and their individuals [42].

Each individual in the ontology belongs to an OWL class. For example, the class Person, represented by a circle in the Figure 16, contains both Matthew and Gemma, who are persons in this specific domain. Classes may be organized in a taxonomy including a superclass and subclasses. Subclasses specialize their super classes.

2.4.2. Ontology Class Hierarchy

The creation of the classes and their hierarchy is the main building block in an OWL ontology. Protégé allow the classes editing in the 'Classes Tab'. An empty ontology can be filled starting from a root class named Thing, so all the classes that the user create are subclasses of Thing. Figure 17 shows the hierarchy of the class of the example proposed. All the classes in the hierarchy are disjoint, so that an individual can be an instance of only one of these three classes. This characteristic may be set in the bottom part of the 'Class Description' view.

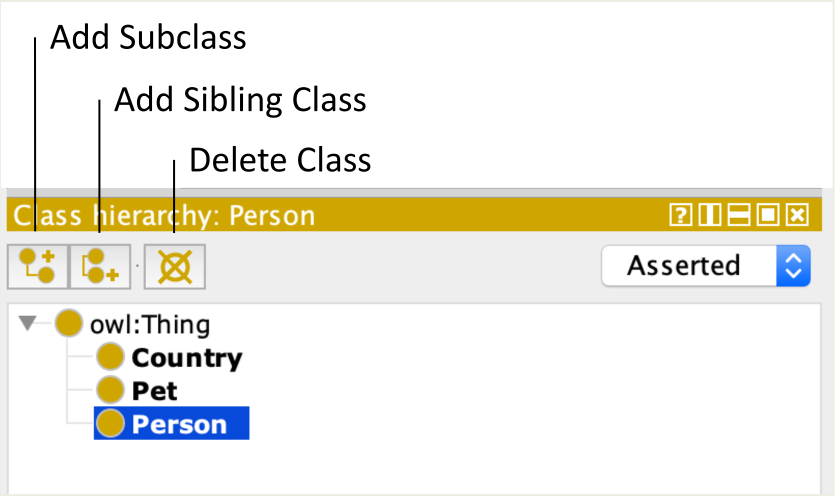


Figure 17 Example of Class hierarchy in Protégé [42].

2.4.3. Ontology Properties

As previously mentioned, relationships in an ontology are represented by OWL Properties. Three types of properties can be distinguished (Figure 18):

- Object Properties representing a relationship between two individuals;
- Data Properties link a Class with a Datatype;
- Annotation Properties adds metadata to classes, individuals and object/datatype properties.

For each property is possible to identify a domain class and a range class. In particular, properties link individuals from the domain to individuals from the range. In the example (Figure 16) Matthew 'hasPet' Fluffy so the domain is Person class and the range is Pet class.

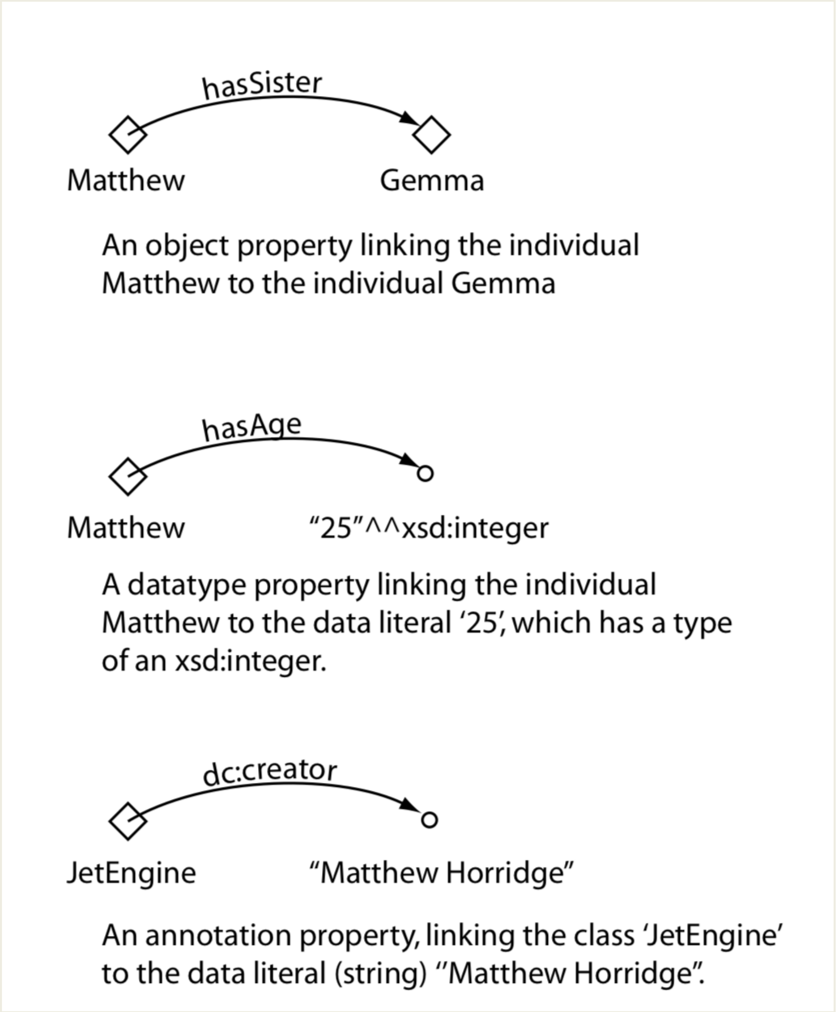

hasSister
Matthew
Gemma
An object property linking the individual Matthew to the individual Gemma
hasAge
Matthew
"25"^^xsd:integer
A datatype property linking the individual Matthew to the data literal '25', which has a type of an xsd:integer.
dc:creator
JetEngine
"Matthew Horridge"
An annotation property, linking the class 'JetEngine' to the data literal (string) "Matthew Horridge".

Figure 18 Examples of the three types of properties in an OWL ontology [42].

All the ontology objects can be annotated with metadata in the annotation properties. This information includes auditing or editorial data such as example, comments, author, creation date etc.

This kind of property is quite flexible, since OWL does not have any constraints on its usage method.

OWL has pre-defined annotation properties:

- owl: versionInfo (range: string);
- rdfs: label (range: string). This property is usually used to add alternative human readable names or multi-lingual names to an ontology element;
- rdfs: comment (range: string);
- rdfs: seeAlso (range: URI identifier to the related resource);
- rdfs: isDefinedBy (range: URI identifier to the reference of the ontology that defines one or more ontology elements).

2.5. NLP Pipeline

The NLP pipeline exploited in this work has been presented in [22] and it was developed for information extraction (IE) of relevant concepts from Italian medical reports of patients with inherited arrhythmias.

Clinical texts considered in the original cardiology application referred to a specific visit occurred in a specific visit date. Data within the report was organized in sections including an anamnestic fitting, the family history, information on performed tests, and at the end a conclusion and the possible drug prescriptions.

The IE pipeline (Figure 19) use different annotators and the UIMA [43] frameworks.

Different steps were implemented: TextPro tool [44] was used to preprocess the report texts in order to implement the sentence splitting, the tokenization, the lemmatization and the speech tagging.

The preprocessed data were analyzed with the first UIMA annotator to identify text sections by using a configuration file including all possible names for standard sections (e.g., "family history"). Two annotators were then used: one to identify events and one to select attributes of interest using the ontology.

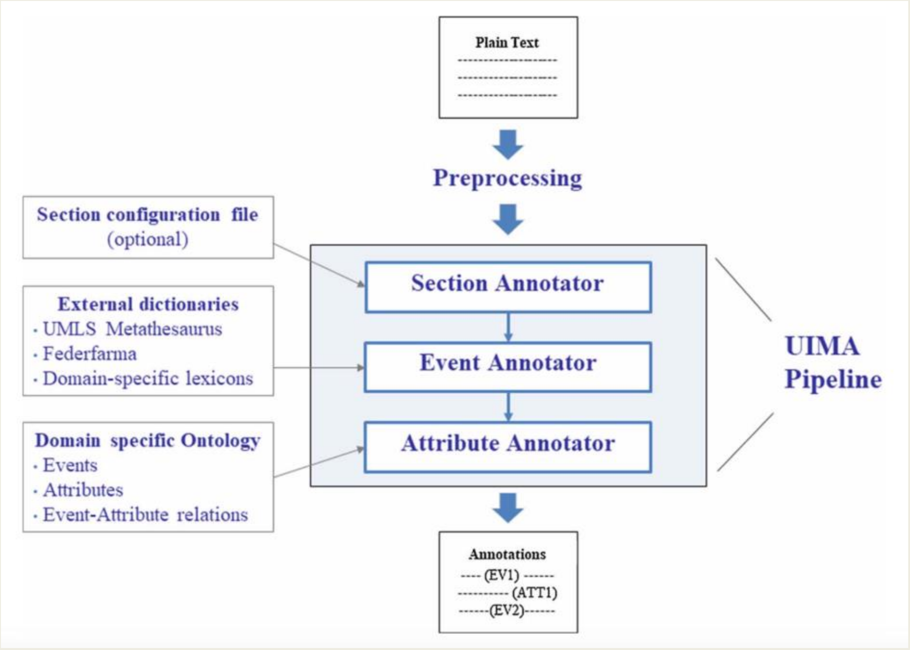


Figure 19 UIMA pipeline [22]

2.5.1. The Event Annotator

The Event annotator uses external dictionaries terms to identify events in the text. In the original work the main dictionaries were:

- the Italian version of UMLS (Unified Medical Language System);
- the FederFarma Italian dictionary of drugs [45];
- custom dictionaries for domain-specific procedures and events of interest;
- a dictionary of common acronyms.

CTAKES UMLS Lookup Annotator [46] was used to identify problems concepts such as diagnostic procedures and treatments. TextPro was reused to normalize plural forms and finally ConText algorithm [47] was used to contextualize selected events

2.5.2. The Attribute Annotator

Thanks to the use of a domain ontology, the Attribute annotator extracts attributes and associated values for the selected events. Event and Attribute instances were loaded in Event and Attribute Classes of the ontology, and linked together through ontology relations, so the same attribute (e.g., “AverageHeartRate”, “Rhythm”) can be connected to multiple events (e.g., “Holter test”, “ECG test”).

The ontology was developed in Protégé [23] and exported as XML file including events, attributes and their relationships.

All the concepts in the ontology are related to a regular expression (Figure 20), a set of properties and a value that can be numeric (including the unit of measurement) or categorical.

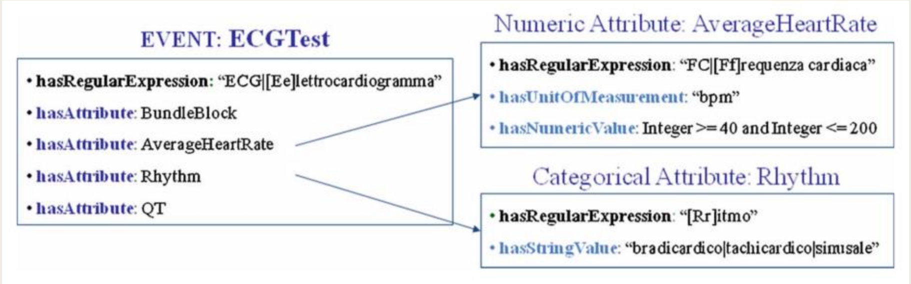

EVENT: ECGTest
• hasRegularExpression: “ECG|[Ee]lettrocardiogramma”
• hasAttribute: BundleBlock
• hasAttribute: AverageHeartRate
• hasAttribute: Rhythm
• hasAttribute: QT
Numeric Attribute: AverageHeartRate
• hasRegularExpression: “FC|[Ff]requenza cardiaca”
• hasUnitOfMeasurement: “bpm”
• hasNumericValue: Integer >= 40 and Integer <= 200
Categorical Attribute: Rhythm
• hasRegularExpression: “[Rr]itmo”
• hasStringValue: “bradicardico|tachicardico|simısale”

Figure 20 Example of the properties for the ECG event and two of its attributes [22].

The Attribute annotator takes the resulting XML file from Protégé as input to match each event found by the dictionary lookup carried on by the Event Annotator to the corresponding concept in the ontology. Their relation is used to identify which are the attributes to be searched and how to search them in text around the events, for example, defining specific lookup windows for each event. In particular, paragraphs (for tests events) and sentences (for drug prescriptions events) are lookup windows for their attribute in the analysis.

In the final step, regular expression set as attribute are evaluated to extract the value for each attribute.

2.6. CFM Algorithm

The algorithm used has been described in [30] and implemented in [31]. It is capable to mine patients' relevant careflows considering the temporal aspect of the available process data.

This algorithm has been developed to overcome some flaws of typical process mining and temporal data mining algorithms applied to clinical routine data. In process mining "Spaghetti-like" models are caused by the presence of very large numbers of different patterns of the patients in routine clinical practice, which is related to the great variability of care pathways due to single patient's management peculiarities [48].

The proposed pipeline tries to reduce this inherent variability and provide clearer event process model. The main goal of the CFM algorithm is to identify groups of patients ("sub-cohort") that share similar careflow pattern. After all the relevant careflows are extracted, results are visualized using Direct Acyclic Graphs (DAG), where rectangles represents events and arcs between them describe their temporal relation (Figure 21).

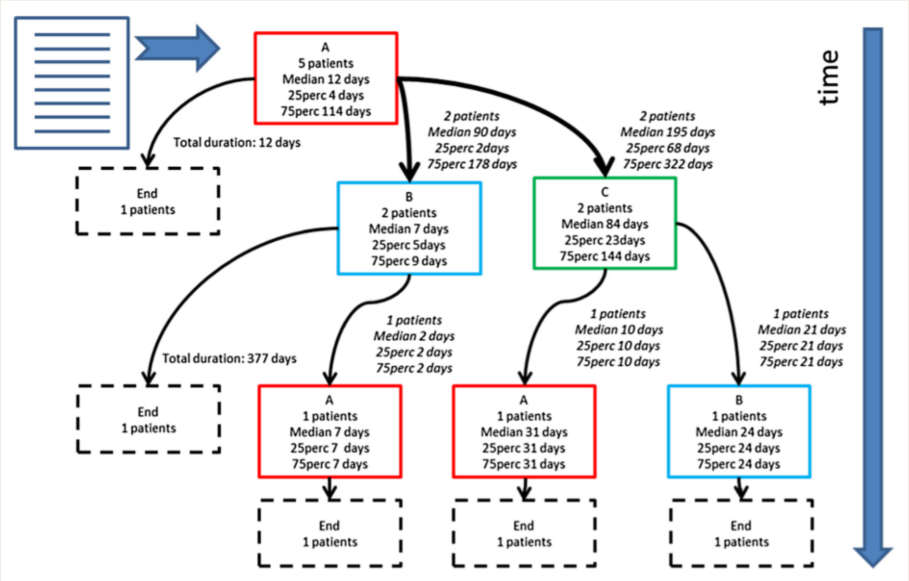


Figure 21 Example of resulting event model [30]. The name and duration of the event is specified in a rectangle, together with the number of patients experiencing that event. Numbers on the arcs represent the number of patients transitioning to the following event and the median duration of the transition. Patients who exit the flow after a specific event are represented as leaf nodes, labeled with "End". Individual careflow represents individual sub-cohort (for example careflow events "A, B, A, End" define the group of patients that have experienced that pattern).

The algorithm provides an explicit temporal characterization of events and event transitions, and allows the identification of patient sub-cohorts. It works in four steps:

- data pre-processing that collect the data and creation of the input event log;
- discovery phase of the careflow events;
- temporal enrichment of events;
- optional clinical enrichment.

2.6.1. Creating the event log

The CFM algorithm works on a file with a list of ordered events, where each row includes the following information:

- ID: the patient subject to the event;
- EVENT: name of the event;
- DATE_INI: event start date;
- DATE_END: event end date;

Each row of the event log is considered as different and independent event, therefore it is important to consider the impact on the results of consecutive repeated events [49]. The clinical scope and analysis goals drive the decision to merge or not consecutive events with the same name. The merging of identical consecutive events results in a unique event with the start date of the first event merged and the end date of the last event merged.

In general, the entire result of the CFM algorithm is heavily determined by the event log description of the clinical events of the patients. The scope of the analysis must be defined in advance and, when possible, following an assessment with clinicians.

2.6.2. Discovery Phase

Within the discovery phase the careflows are mined. This step relies on the definition of frequency of a sequence of events, namely the support [50]. The support of a sequence is the proportion of patients that experience a particular sequence of events against the total population.

A user-defined threshold is used to retain only sequences that have a frequency greater than a fixed support threshold. In this way, only the most frequent patterns are extracted.

The search process is based on another user-defined parameter that sets the maximum length of a sequence to be included in the careflow search.

At the end of each careflow, a final event, called the exit event, is added. A patient careflow ends in an exit box for three reasons:

- The patient has proceeded through all his/her events available in the data log.
- The patient has other events, but the careflow that would result from them has a support lower than “minimum support” threshold.
- The patient has other events, but the careflow that would result from them has a length that is higher than the “maximum length” parameter.

This discovery step of the algorithm requires a careful assessment of the minimum support and maximum length parameter. These two parameters affect the generalization and precision of the CFM models: low support and high maximum length can lead to overfitting and a difficult interpretation of the results, losing power to summarize care pathways. On the other hand, doing the opposite can retain only a general description of the initial events of most patients, loosing details and becoming underfitted.

2.6.3. Temporal Enrichment

The resulting DAG (Figure 21) describes also the temporal information of the present elements. Each event is represented as a box, and its temporal description is summarized by the median and percentiles of patients’ durations that are included in that event of the sequence. In the same way, the median and percentiles durations of the transition of patients between two events are reported on an arc between two consecutive events.

The duration of an event and the time between two consecutive events can give useful insights to better characterize sub-cohort of patients. For example, in the oncology domain, the time between a surgery and the following chemotherapy, and the duration of the treatment can provide useful information about the status of patients.

At the end of every careflow, the total history time of the patients of the careflow is reported, computed as the median of times between first and last displayed events of the careflow.

2.6.4. Clinical Enrichment

A final, and optional, step can be used to add other type of information to the CFM results. It is called clinical enrichment, and it aims to better describe the patient sub-cohorts created with the careflows, reporting values of clinically significant variables.

For example, laboratory test results summaries, if available during the study, can be added to the DAG after a certain event, showing the difference of values of the test in the different patient sub-cohorts (Figure 22). To this end, clinical values of the sequence events are needed.

This last step of the enrichment phase can add clinical meaning to the evaluation of the patient careflows, allowing comparison of sub-cohorts and thus statistically validating the difference in patient health status in different sub-cohorts. Moreover, clinicians can have a tool to check the clinical characterization and consistency of the careflows.

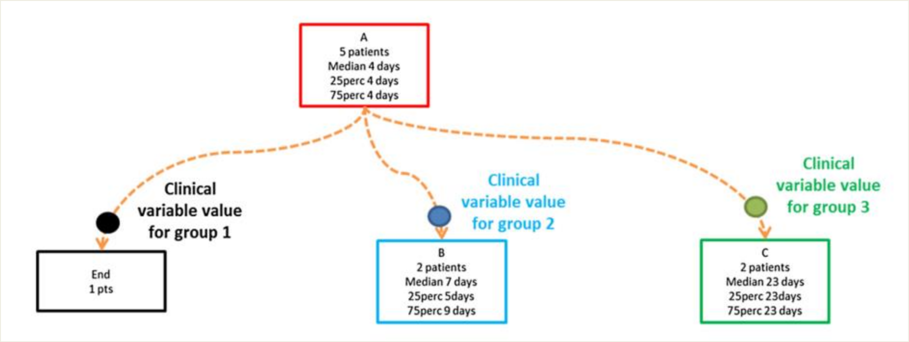


Figure 22 Example of clinical enrichment of careflows [30], performed by considering the last measurement of the clinical variable before the second event of the careflow

Chapter 3

Implementation

This chapter explains how the previously described tools has been applied to the Hospital Papa Giovanni XXIII (HPG23) scenario.

It describes the ETL (Extract, Transform and Load) model used to populate the i2b2 data warehouse, the NLP (Natural Language Processing) application to breast cancer pathology reports, and the design of an i2b2 plugin for CFM (CareFlow Mining).

3.1. Populating i2b2 Data Warehouse

A crucial part of the work carried on in this thesis was the development of the ETL processing of the data. Starting from raw data sources, described in chapter 2.1, data were analyzed and manipulated using the Mirth tool (2.3), version 3.5.

The ETL task is programmed to make periodical updates, matching the actual HIS (Hospital Information System) data to the i2b2 content. After some discussion with hospital clinicians and technicians of the HIS, it was decided to run the whole operation weekly, on Saturday night. Figure 23 summarizes the principal information involved in the ETL task.

The i2b2 data structure is designed to handle multiple visits of the patients during their care histories and to store several observations for each encounter. Patients and encounters data fill corresponding dimensions tables.

After that, patient and encounter codes can be used for the creation of observations, where information about clinical events are stored using concepts.

It is important to populate also the concept dimension table that explain the meaning of the concept codes used in observation table.

Finally, taxonomies for these concepts are developed, in order to allow an easy exploration with the i2b2 query tool (2.2.2.2).

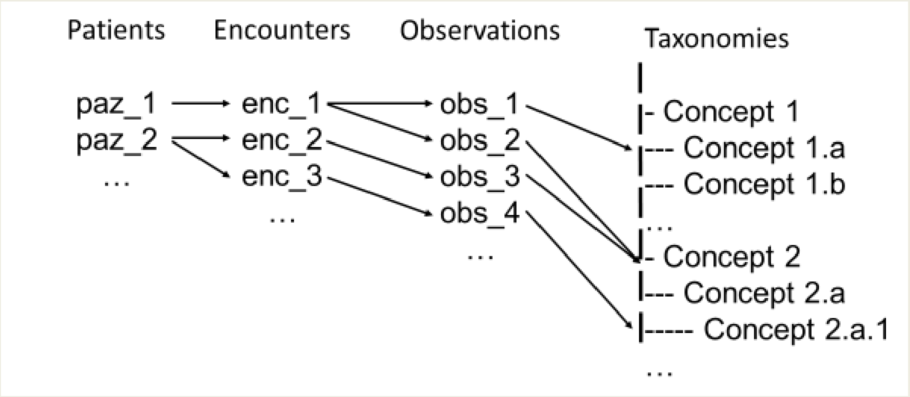


Figure 23 relevant entities created by ETL task and loaded into i2b2 data warehouse.

Typically, taxonomies are based on standard coding systems, ICD9 for diagnosis and procedures, LOINC for laboratory values and ATC for drugs.

There are also ad hoc taxonomies describing internal codification system of the hospital, like internal codes of laboratory tests, or to describe other general concepts (patient age, gender and visits occurrences). For example, demographic data will not only be inserted as information in the Patient Dimension, but also as standalone observation, so that they can be queried (Figure 24).

In the same way, additional ad-hoc taxonomy describes the “contact” of a patient with an encounter type (Figure 25). For example, every time a patient generates a hospitalization encounter, a “contact” concept is inserted in its own observation.

Since some events type (i.e. events from laboratory source) can be absorbed into “container” events (i.e. events from HDR or outpatients sources), they do not have their own encounter, but they still create a contact observation. In this manner, this type of information can be used in the query tool.

Patients, encounters and observations are generated from each available data source. Although, all data sources are different, and it required specific data manipulation and i2b2 uploading procedures.

Patient and encounter ids are pseudo-anonymized, as required by the i2b2 framework. Mapping from the original ids to the anonymized ids is maintained in separate tables. Only with special permissions this anonymization can be reversed.

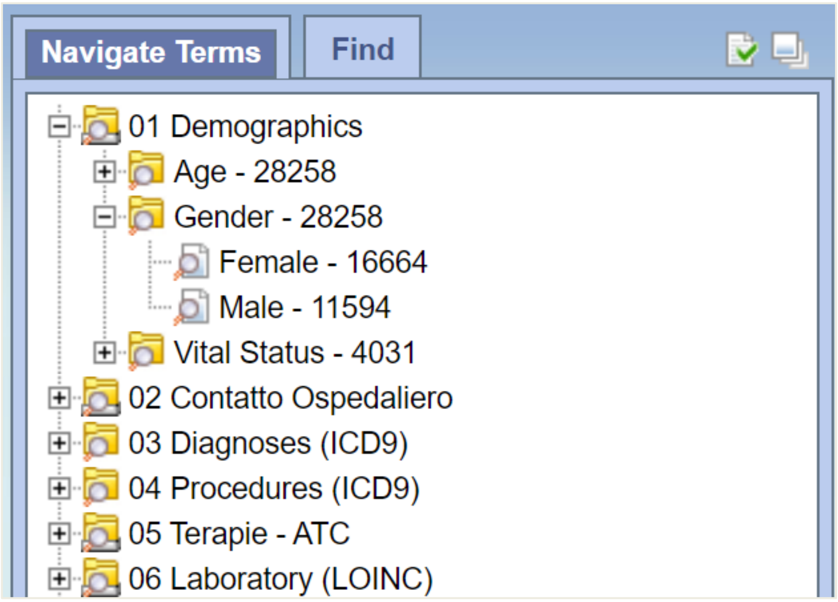


Figure 24 I2b2 web query tool, with demographics taxonomy expanded. Thanks to the creation of these observations, these concepts can be dragged and dropped, like other clinical concept and used for patient identification.

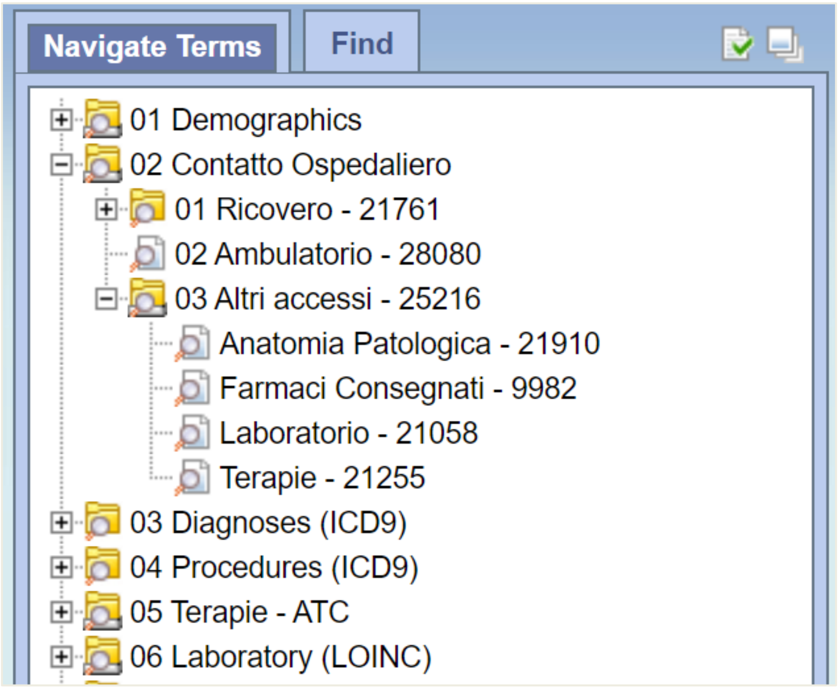


Figure 25 I2b2 web query tool, with contacts taxonomy expanded. Thanks to the creation of these observations, these concepts can be dragged and dropped like other clinical concepts and used for patient identification.

In a cascade process, all source database views are read, and information is stored in i2b2 tables following a weekly schedule, i.e. the data of the last week are imported. Some preliminary tasks are usually done in preparation of the data import, such as updating of the patient dimension, and related observations, in relationships with the corrections made in the Demographic view of the HIS in the last week.

Also, in this preliminary part of the process, the duplicated patients inside the Demographic view reported by the HIS technicians are considered. A table that match the old patient codes to the new ones is created and is used to update i2b2 encounters and observations that are referring to the old patient codes.

3.1.1. Mirth Channel Architecture

Here the Mirth channels architecture that is the basis of the ETL tasks is briefly reported.

Database connectors (2.3.2.4) are currently used to start a connection with the database views, but most of the work is done using custom JavaScript scripts and functions. Instead of messages, data are handled with variables and data structures inside each channel. JavaScript permits to open connection directly from the scripts, to create dynamic queries with strings management, to run them on the views connections and to supervise the results, to call external java libraries and finally to change another channel status.

The weekly firing of the ETL pipeline is triggered by a first channel, which also provides the preliminary tasks on the Demographics view.

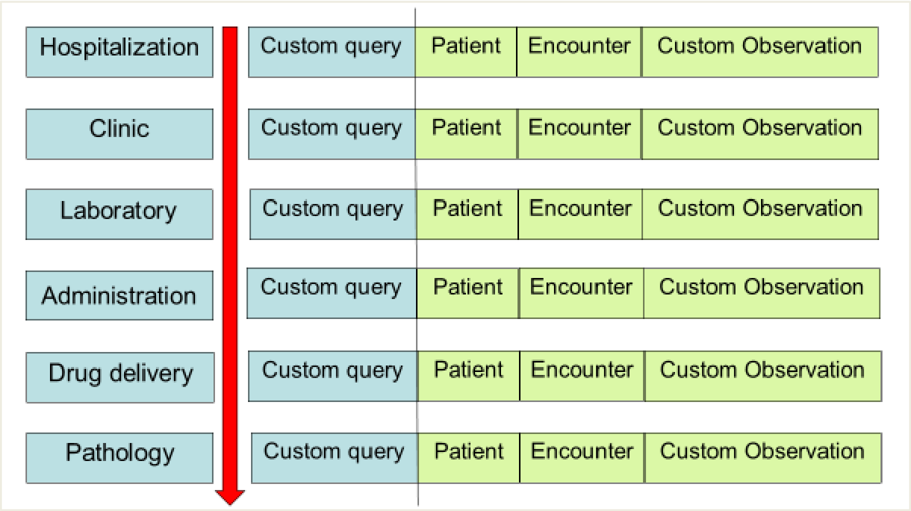
The channels are activated subsequentially. A single channel is devoted to each database source. These channels are all similar in their structure, and common operations are inserted in functions. Figure 26 shows a representation of the final ETL model.

At the beginning of a channel, an obligated source-dependent operation is the oracle query that will select data from the view and store them in a JavaScript structure, the result set. All the queries have the fundamental task to prepare the necessary information for i2b2 tables. The result set will have columns reserved to the patient, some to the encounter and variable number of columns that will create the observations.

The idea is to have different queries dealing with different sources, and to uniform the output of these queries. Thanks to this strategy, the i2b2 patient and encounter data is performed by a unique function that can be called from all the different channels.

The remaining data from the result set are used to construct the i2b2 observations of the data source, another source-dependent operation.

This approach also allows minimizing the number of query operations on a data source, the real time bottleneck of the overall process. For each database view a single query retrieve all necessary information, using the same result set to fill the i2b2 data table.



Hospitalization	Custom query	Patient	Encounter	Custom Observation
Clinic	Custom query	Patient	Encounter	Custom Observation
Laboratory	Custom query	Patient	Encounter	Custom Observation
Administration	Custom query	Patient	Encounter	Custom Observation
Drug delivery	Custom query	Patient	Encounter	Custom Observation
Pathology	Custom query	Patient	Encounter	Custom Observation

Figure 26 ETL model visualization of the mirth channels. The database views are shown on the left side of the red arrow; on the right side the mirth channels. Database operation on the sources are in blue color, while entities creation and insertion in data warehouse is shown in green.

3.1.2. Hospitalizations

Both Monitor and Galileo data are managed in the same way, given the similarity of the two data sources.

A peculiar aspect of hospitalization data is that information inserted in i2b2 in previous week could be obsolete. This happens because a hospitalization can last longer than a week, even months in some cases.

Each week, all the current hospitalizations are reloaded into i2b2, deleting all information previously inserted. This ensure that i2b2 data are up to date.

This data process is performed also to data related to hospitalizations that are closed in the current week, so that old data are deleted, and the data related to a finalized discharge are loaded into i2b2.

The different i2b2 entities (patients, encounters, observations) are created as follows:

- i2b2 Patients
    - New patients are inserted in the Patient Dimension table using the field “Patient id”, and based on Demographic data
- i2b2 Encounters

    - New encounters are inserted in the Visit Dimension table using the field "Hospitalization id" with date of field "Hospitalization start date"
- i2b2 Observations
    - Age observation is created for each patient, using the fields "Birth date" and "Death date" of the Demographic view;
    - Gender observation is created for each patient, using the field "Sex" of the Demographic view;
    - Contact of hospitalization is created as open or closed (only closed for MONITOR hospitalizations) accordingly to the presence of the field "Hospitalization end date". The reported date is taken from the field "Hospitalization start date";
    - ICD9 diagnosis observations using fields "Main Diagnosis ICD9 code", as well as secondary diagnosis, with date "Hospitalization start date";
    - ICD9 procedures observations using fields "Main Procedure ICD9 code" and secondary procedures, with the dates of fields "Main Procedure date" and secondary procedure dates, respectively.

3.1.3. Clinical Outpatient Services

Clinical outpatients services view gives the information of what procedures an outpatient had done. Using the SISS code of the procedures, one of the preprocessing works on this channel is to translate it into ICD9 code.

The different i2b2 entities (patients, encounters, observations) are created as follows:

- i2b2 Patients
    - New patients are inserted in the Patient Dimension table using the field "Patient id", and based on the Demographic data
- i2b2 Encounters
    - New encounters are inserted in the Visit Dimension table using the field "Booking number" with date reported in the field "Visit date"
- i2b2 Observations
    - Age observation is created for each patient, using the fields "Birth date" and "Death date" of the Demographic view
    - Gender observation is created for each patient, using the field "Sex" from Demographic view

- Contact of outpatient visit is created with date of field "Visit date"
- ICD9 procedures observations are created using field "Procedure SISS code" with the dates of field "Visit date". This SISS code is mapped with a regional reimbursement code [51] that is similar to ICD9 code.

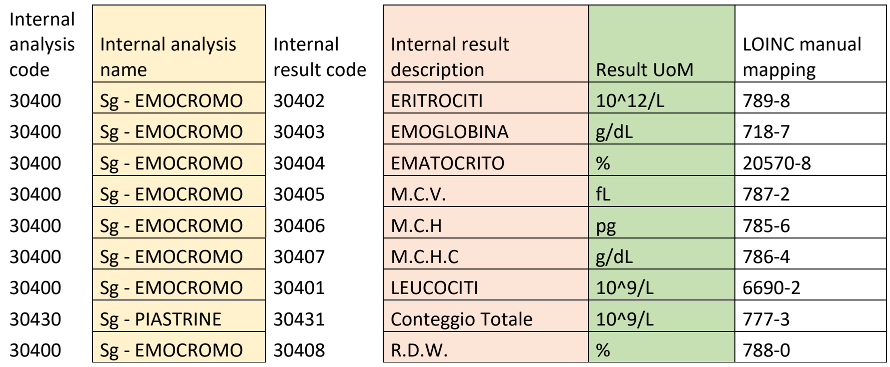
3.1.4. Laboratory

In this view, laboratory results are presented only using the internal coding system. A manual mapping, with the help of laboratory technicians, has been performed for the most used analysis results (less than 100 mappings cover around the 85% of the analysis made).

The LOINC mapping was performed using the LOINC official site [52] and the hospital internal application that describe also the modality and other information of analysis execution. One example of this mapping can be observed in Table 1.

The LOINC mapping inside whole ETL process is heavily automatized. A single table in the database contain all necessary information to transform an internal laboratory code to a LOINC code. The table can be updated incrementally. For this reason, the ETL procedure automatically finds correct mapping in the table and is capable of retrospectively change laboratory analysis already imported in i2b2 to LOINC code in case the mapping was not present in a first upload.

Table 1 Each internal analysis and result code corresponds to a specific LOINC code



Internal analysis code	Internal analysis name	Internal result code	Internal result description	Result UoM	LOINC manual mapping
30400	Sg - EMOCROMO	30402	ERITROCITI	10^12/L	789-8
30400	Sg - EMOCROMO	30403	EMOGLOBINA	g/dL	718-7
30400	Sg - EMOCROMO	30404	EMATOCRITO	%	20570-8
30400	Sg - EMOCROMO	30405	M.C.V.	fL	787-2
30400	Sg - EMOCROMO	30406	M.C.H	pg	785-6
30400	Sg - EMOCROMO	30407	M.C.H.C	g/dL	786-4
30400	Sg - EMOCROMO	30401	LEUCOCITI	10^9/L	6690-2
30430	Sg - PIASTRINE	30431	Conteggio Totale	10^9/L	777-3
30400	Sg - EMOCROMO	30408	R.D.W.	%	788-0

The different i2b2 entities (patients, encounters, observations) are created as follows:

- i2b2 Patients
  - New patients are inserted in the Patient Dimension table using the field "Patient id" and based on the Demographic data.

- i2b2 Encounters
  - New encounters are inserted in the Visit Dimension table using the field “Lab request id” with corresponding date reported in the field “Lab request date”.
  - When possible the fields “Hospitalization id” and “Booking number” present in the lab view are used. Instead of creating a new encounter, laboratory observations are associated with the hospitalization or outpatient visit where the request has been created.
- i2b2 Observations
  - Age observation is created for each patient, using the fields “Birth date” and “Death date” of the Demographic view
  - Gender observation is created for each patient, using the field “Sex” of the Demographic view
  - Contacts observations of lab request with date of field “Lab request date”
  - LOINC observations using the mapping at Table 1, with date “Lab request date”
  - Internal lab codes observations, for analysis not mapped to the LOINC code.

3.1.5. Drugs

The two drugs administrations and deliveries views (2.1.4) are quite different: both can occur in outpatient visits, but only the in-hospital administrations may occur also during hospitalizations. Moreover, information about drug schedule and administration route are not present in the “delivery to patient” view. But some important columns are shared between the two, so a similar ETL task was performed.

The different i2b2 entities (patients, encounters, observations) are created as follows:

- i2b2 Patients
  - New patients are inserted in the Patient Dimension table using the field “Patient id”, and by accessing to Demographic data.
- i2b2 Encounters
  - New therapy encounters are inserted in the Visit Dimension table using the field “Event id” with the date of the field “Event start date” (for drug administration) or “Actual delivery date” (for drug delivery).
  - When possible the fields “Hospitalization id” (only from direct administration view) and “Booking number” present in the drug views are used. Instead of creating a new encounter, drug observations are associated with the

hospitalization or outpatient visit in which the prescription is given to the patient.

- i2b2 Observations
    - Age observation is created for each patient, using the fields “Birth date” and “Death date” from Demographic view.
    - Gender observation is created for each patient, using the field “Sex” from Demographic view.
    - Contacts observations of administration or delivery of drugs are created with date corresponding to the field “Event start date” (for drug administration) or “Actual delivery date” (for drug delivery).
    - Drug observations with the available field “Drug internal code” (for drug administration) or “Prescribed drug internal code” (for drug delivery). The start dates of these drug observations are collected from the field “Real administration start date” (for drug administration) or “Actual delivery date” (for drug delivery).

As described above, observation concepts are built using internal drug code. All internal codes utilized have an associated ATC code, so the available i2b2 observation data can be queried with an ATC taxonomy.

Each i2b2 observation contains also the information of dosage and unit of measurement of prescriptions:

- For drug administration, data are taken from the fields of “Prescription dosage” and “Prescription dosage Unit of Measurement”.
- For drug delivery, data are taken from the fields “Prescribed drug total quantity” and “Prescribed drug Unit of Measurement”.

Moreover, some drug administrations are collected with a continuous time stamp. For these observations the end date is set using the field “Real administration end date”.

Drugs observations regarding administration inside the hospital have also schedule therapy information and administration route, from the fields “Schedule” and “Administration route”. These two data are tightly connected with the drug observation that already contains information of dosage.

This is the typical case that can be exploited to explain the use of modifier inside an observation. In 2.2.1 it was reported that modifier can give some additional information about the concept. In our case the drug observation concept is the drug code, and additional information are dosage, route and schedule.

The overall result, in terms of i2b2 observation table, is the creation of three different observations, linked by belonging to the same event (Table 2).

Table 2 A concise representation of created observations

Patient	Encounter	Concept	Modifier	Value type	Value	UoM
“1”	“234”	drug code 1		Number	2	ml
“1”	“234”	drug code 1	“Schedule”	String	Schedule name	
“1”	“234”	drug code 1	“Route”	String	Route type	

3.1.6. Pathology reports

Within the pathology reports view, most of the contents are free text.

i2b2 can store large textual information in its database structure. A dedicated column of the observation fact table, the “observation blob”, permits to insert text observations.

i2b2 optimizes the exploration of these type of textual data, using a dedicated database index and allowing the user to make simple textual searches from the web query tool.

As a preprocessing step, the data structure of the database view is slightly modified. In the original view there is a single row per specimen, so a single report is represented with several rows. The data has been manipulated in order to generate a single row per report, merging different rows content, such as specimen description.

- i2b2 Patients
  - New patients are inserted in the Patient Dimension table using the field “Patient id”, and by accessing to Demographic data.
- i2b2 Encounters
  - New encounters are inserted in the Visit Dimension table using the field “Report id” with relative date of field “Report date”.
  - When possible, the field “Hospitalization id” present in the pathology view is used. Instead of creating a new encounter, report observations are associated with the hospitalization when the pathology analysis has been performed.
- i2b2 Observations
  - Age observation is created for each patient, using the fields “Birth date” and “Death date” from the Demographic view.
  - Gender observation is created for each patient, using the field “Sex” from the Demographic view.

- Contacts observations of pathology reports with date of field “Report date”.
- Observation with textual information one for each section provided:
    - Specimen descriptions
    - Specimen morphology SNOMED codes
    - Specimen topography SNOMED codes
    - Specimen procedure SNOMED codes
    - Anamnestic information section
    - Clinical information section
    - Macroscopic description section
    - Microscopic description section
    - Diagnosis section
    - Preliminary diagnosis section
    - Comment section
    - Addendum

Since the provided view divide the reports in “sections”, the observations are created accordingly, storing a textual section in a dedicated observation.

Again, a single report is stored as a group of different but related observations, and the modifier column is used to identify which section is memorized in the observation blob column.

In this way, a textual search from the web query tool can be focused only on a specific section of a report.

3.2. NLP pipeline

In a previous work, our group developed a pipeline in the cardiology field that, starting from Italian reports, identifies domain events (e.g., diagnostic procedures) and their related attributes (e.g., test results) [22].

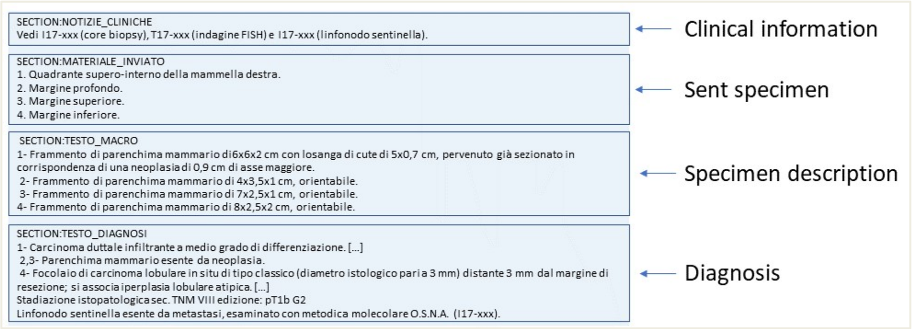
This pipeline is adaptable to other clinical domains with a proper ontology modification. In this thesis the pipeline has been adapted for the analysis of oncological reports related to breast cancer.

We converted the new domain ontology in a customized i2b2 ontology in order to implement the information extraction with the aim to store data directly into i2b2.

Since reports are currently stored as unstructured text, it was fundamental the application of an information extraction (IE) technique.

The pathology reports were generated in the hospital using an electronic form and, as shown in Figure 27, they have some predefined sections, in particular:

- "clinical information" regarding references to previous tests;
- "sent specimen", including the lists all the analyzed specimens, each of them being targeted with a specimen number that make it trackable;
- "specimen description", containing details about this specimen;
- "diagnosis", which reports the diagnostic conclusions. Different diagnoses are allowed in this section, each one related to a specific specimen.


SECTION:NOTIZIE_CLINICHE
Vedi I17-xxx (core biopsy), T17-xxx (indagine FISH) e I17-xxx (linfonodo sentinella).

SECTION:MATERIALE_INVIATO
1. Quadrante supero-interno della mammella destra.
2. Margine profondo.
3. Margine superiore.
4. Margine inferiore.

SECTION:TESTO_MACRO
1- Frammento di parenchima mammario di6x6x2 cm con losanga di cute di 5x0,7 cm, pervenuto già sezionato in corrispondenza di una neoplasia di 0,9 cm di asse maggiore.
2- Frammento di parenchima mammario di 4x3,5x1 cm, orientabile.
3- Frammento di parenchima mammario di 7x2,5x1 cm, orientabile.
4- Frammento di parenchima mammario di 8x2,5x2 cm, orientabile.

SECTION:TESTO_DIAGNOSI
1- Carcinoma duttale infiltrante a medio grado di differenziazione. [...]
2,3- Parenchima mammario esente da neoplasia.
4- Focolaio di carcinoma lobulare in situ di tipo classico (diametro istologico pari a 3 mm) distante 3 mm dal margine di resezione; si associa iperplasia lobulare atipica. [...]
Stadiazione istopatologica sec. TNM VIII edizione: pT1b G2
Linfonodo sentinella esente da metastasi, esaminato con metodica molecolare O.S.N.A. (I17-xxx).

Figure 27 Example of a pathology report. [53]

3.2.1. Ontology Development and Information Extraction

We use an ontology-driven IE pipeline for the extraction of events and their attributes exploiting the relations defined in the manually created ontology. Concepts within the report text are extracted using regular expressions related to each Event and Attribute classes in the ontology. Useful concepts and their possible variants were recognized in clinical notes using statistics approach like N-grams (Table 3). This approach is based on the detection of the most frequent sequences of n words to highlight the most widely used concepts and how they are written in the text.

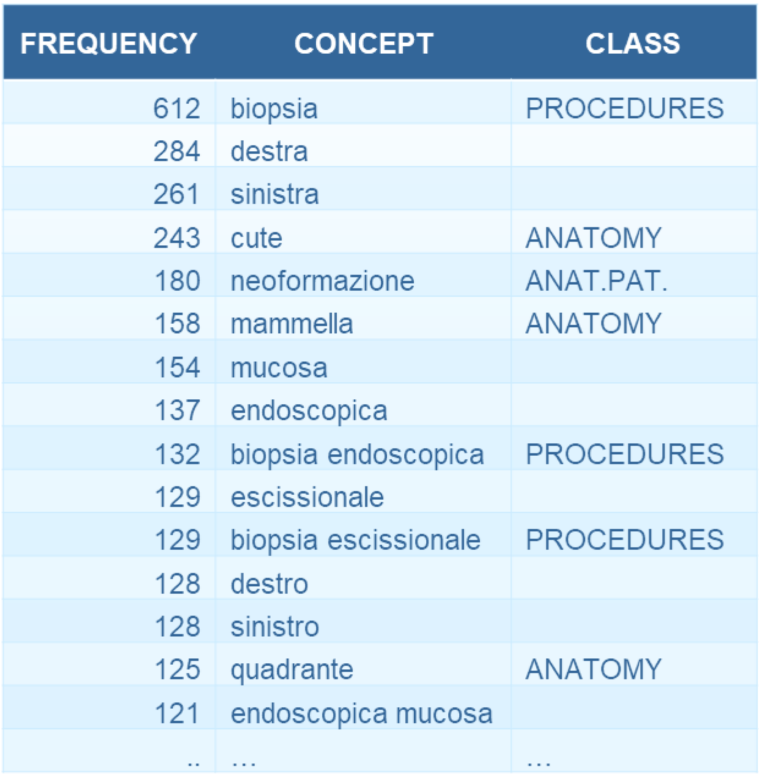
The NLP-ontology was finally optimized for the oncology domain thanks to the collaboration with physicians that help in the selection of relevant concepts, from a set of 20 reports randomly selected to be manually reviewed.

As a standard ontology we used Bio Portal PATH LEX [54] [55]; this ontology allowed the definitions of many useful concepts (Figure 28).

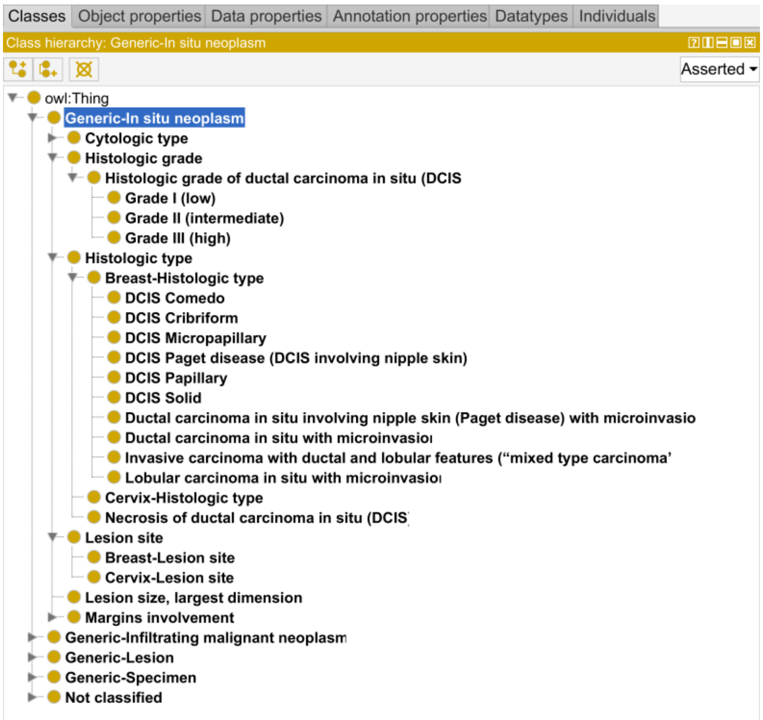
After this preliminary analysis the most relevant entities are the following:

- Specimen, such as core biopsies or organ portions described in pathology reports.
- Diagnosis (even in a negated form) in conclusions.
- Histopathological stage included in the case of breast cancer diagnosis.
- Prognostic factors such as the expression of estrogen and progesterone receptors; they are often included in reports.

Table 3 N-gram example table



FREQUENCY	CONCEPT	CLASS
612	biopsia	PROCEDURES
284	destra	
261	sinistra	
243	cute	ANATOMY
180	neoformazione	ANAT.PAT.
158	mammella	ANATOMY
154	mucosa	
137	endoscopica	
132	biopsia endoscopica	PROCEDURES
129	escissionale	
129	biopsia escissionale	PROCEDURES
128	destro	
128	sinistro	
125	quadrante	ANATOMY
121	endoscopica mucosa	
..	…	…



- owl:Thing
  - Generic-In situ neoplasm
    - Cytologic type
    - Histologic grade
      - Histologic grade of ductal carcinoma in situ (DCIS
        - Grade I (low)
        - Grade II (intermediate)
        - Grade III (high)
    - Histologic type
      - Breast-Histologic type
        - DCIS Comedo
        - DCIS Cribriform
        - DCIS Micropapillary
        - DCIS Paget disease (DCIS involving nipple skin)
        - DCIS Papillary
        - DCIS Solid
        - Ductal carcinoma in situ involving nipple skin (Paget disease) with microinvasio
        - Ductal carcinoma in situ with microinvasio
        - Invasive carcinoma with ductal and lobular features (“mixed type carcinoma’
        - Lobular carcinoma in situ with microinvasio
      - Cervix-Histologic type
      - Necrosis of ductal carcinoma in situ (DCIS
    - Lesion site
      - Breast-Lesion site
      - Cervix-Lesion site
    - Lesion size, largest dimension
    - Margins involvement
  - Generic-Infiltrating malignant neoplasm
  - Generic-Lesion
  - Generic-Specimen
  - Not classified

Figure 28 one of the parts of the PATH LEX ontology available in Bio portal [54]

For each of these four entities we used the event-attribute ontology structure to determine a set of attributes of interest. For example, specimens can be characterized by their size and are linked with a specimen number, while prognostic factors can be linked to a test result.

We used an iterative approach to refine the ontology and the process of information extraction.

Therefore, the ontology has been built in several steps. The first version was manually curated considering the knowledge and the reports of the clinical domain. Then, this version and the IE systems were iteratively refined with the contribution of domain experts.

The final ontology (Figure 29) implemented in Protégé framework includes 17 events and 21 attributes both arranged into four main classes: Specimen, Diagnosis, Histopathological Stage, and Prognostic Factors.

Based on our model of the breast cancer oncology domain, the two possible diagnosis are benign and malign tumor (cancer or metastasis) while specimen includes, for example, biopsies and chirurgical resections, ranging from single nodules to the entire breast or to lymph nodes.

The resulting ontology was used by the NLP pipeline to process reports and extract relevant concepts through the usage of regular expressions. Then, identified concepts and relations are saved in an output file in XML format.

As already mentioned, the resulting ontology was used by the NLP pipeline to process reports and extract relevant concepts through the usage of regular expressions (Figure 30 and Figure 31). Then, identified concepts and relations are saved in an output file in XML format.

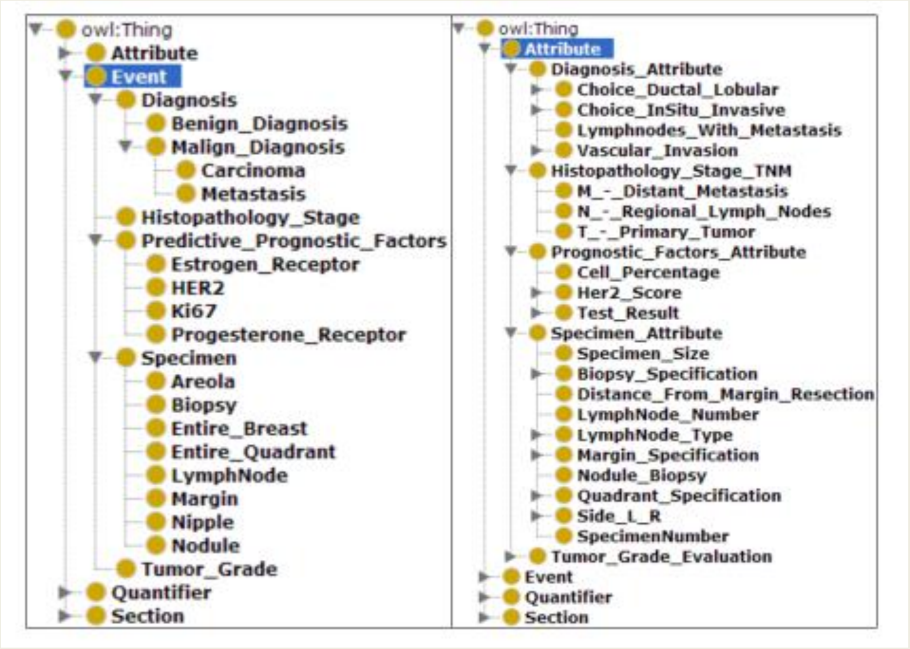
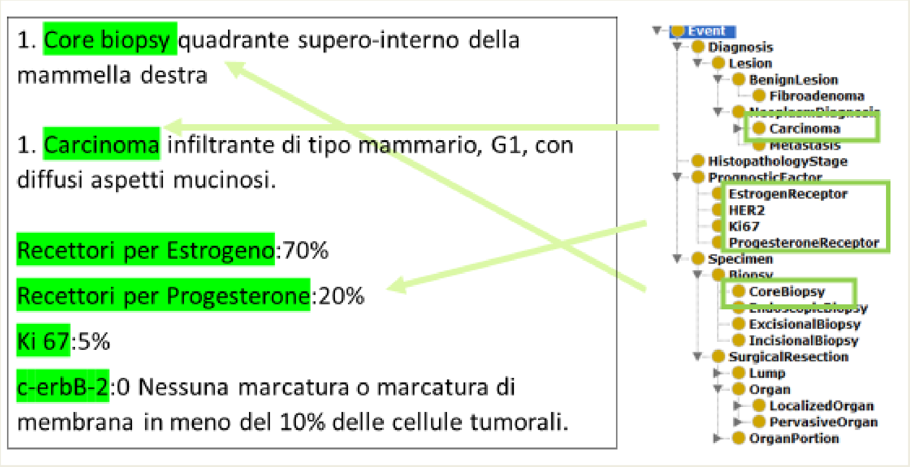


Figure 29 Breast domain ontology class structure: events (left) and attributes (right).


1. Core biopsy quadrante supero-interno della mammella destra

1. Carcinoma infiltrante di tipo mammario, G1, con diffusi aspetti mucinosi.

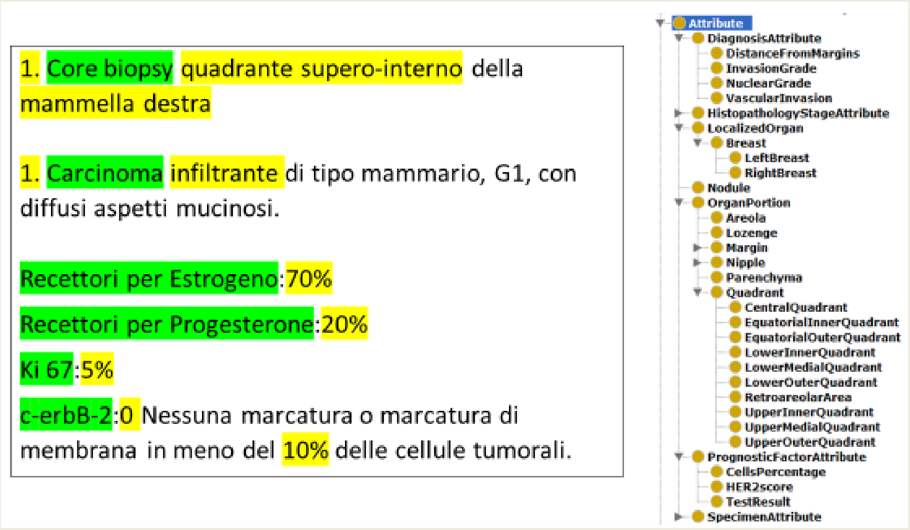
Recettori per Estrogeno:70%

Recettori per Progesterone:20%

Ki 67:5%

c-erbB-2:0 Nessuna marcatura o marcatura di membrana in meno del 10% delle cellule tumorali.

Figure 30 Events recognition in the text with regular expressions, specified for each ontology event



1. Core biopsy quadrante supero-interno della mammella destra

1. Carcinoma infiltrante di tipo mammario, G1, con diffusi aspetti mucinosi.

Recettori per Estrogeno:70%

Recettori per Progesterone:20%

Ki 67:5%

c-erbB-2:0 Nessuna marcatura o marcatura di membrana in meno del 10% delle cellule tumorali.

Figure 31 Attribute recognition in the text with regular expression. The right attribute to search is indicated, again, in the ontology, where object properties connect an event will all its attributes. For each attribute the correct regular expression is stored in a dedicated annotation property. Another Annotation property specify the lookup window.

3.2.2. Porting data into i2b2 taxonomy and observations

We manually derived the i2b2-Ontology starting from the defined NLP-Ontology. After the insertion of the structured data in the i2b2 data warehouse, the i2b2 taxonomy allows the user (usually physicians and researchers) to query the database.

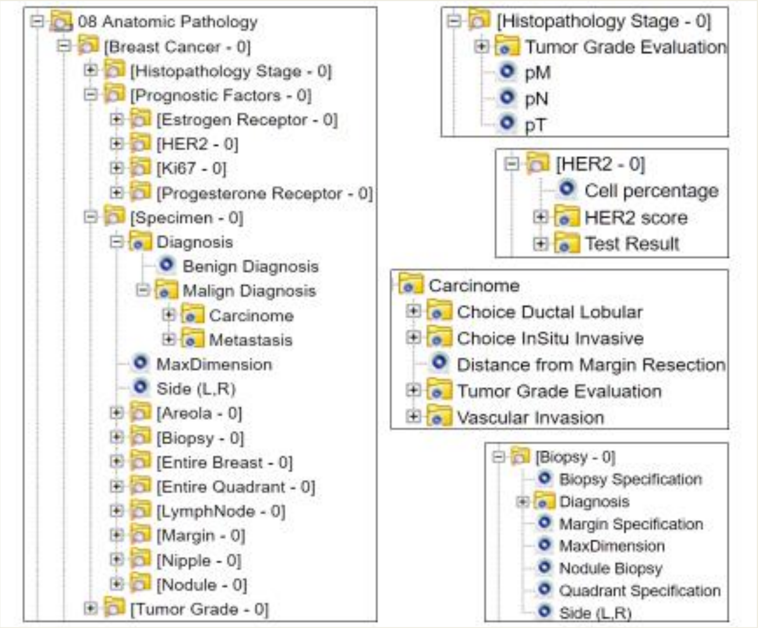


Figure 32 Final i2b2 taxonomy.

Events and Attributes will become in the i2b2 taxonomy concepts and modifiers, respectively.

An exception is done for Specimen and Diagnosis. Although these two Events are independently extracted by the NLP pipeline, a Diagnosis become modifier of its Specimens in i2b2.

After the i2b2 taxonomy is created, the output of NLP extraction can be used to populate the observation fact table, using concept and modifiers defined by the taxonomy.

The NLP pipeline output has an XML format. For each report, a file containing all extracted events attributes is created:

```
EVENTS:
<events.types.ClinicalEvent
_id="14907"
begin="268" end="286"
semanticType="SPECIMEN"
CE_id="15"
polarity="AFFIRMED"
eventName="Quadrante centrale"/>
…

ATTRIBUTES:
<values.types.AttributeValue
_id="15453"
begin="265" end="267"
attribute="SpecimenNumber" value="1."
AV_id="3" />
<values.types.AttributeValue
_id="15439"
begin="293" end="308"
attribute="LocalizedOrgan" value="mammella
destra" AV_id="2" />
…

EVENT-ATTRIBUTE LINKS:
<values.types.EvAtt_Link
idEvent="15"
idAttributeValue="2" />
<values.types.EvAtt_Link
idEvent="15"
idAttributeValue="3" />
```

These results could be analyzed by a custom script and manually inserted into i2b2 data warehouse as observations.

All the observation imported in i2b2 are associated with the date of the report from which it derives and with the visit id reported in the anatomical pathology view.

This kind of workflow is very time consuming, with manual steps involved. This has been done for a specific project (4.2.1). A general pipeline procedure (below) is being developed to avoid all manual steps between ontology creation and i2b2 observations.

3.2.3. General NLP procedure

Once the four steps were implemented, we started the development of a single software resource able to properly configure all the data extraction, transformation and loading processes (Figure 33); this unique resource is the Ontology authored in OWL format.

In detail, the comprehensive architecture includes:

1. the ontology creation and curation
2. the generation of the i2b2 taxonomy
3. the extraction of the information from the clinical narratives
4. the data normalization and population of i2b2

With this project, still under implementation, we try to solve the previously required manual insertion of the taxonomy entries (step 2) and a heavy postprocessing work to translate the NLP XML output to structured i2b2 observations (step 4).

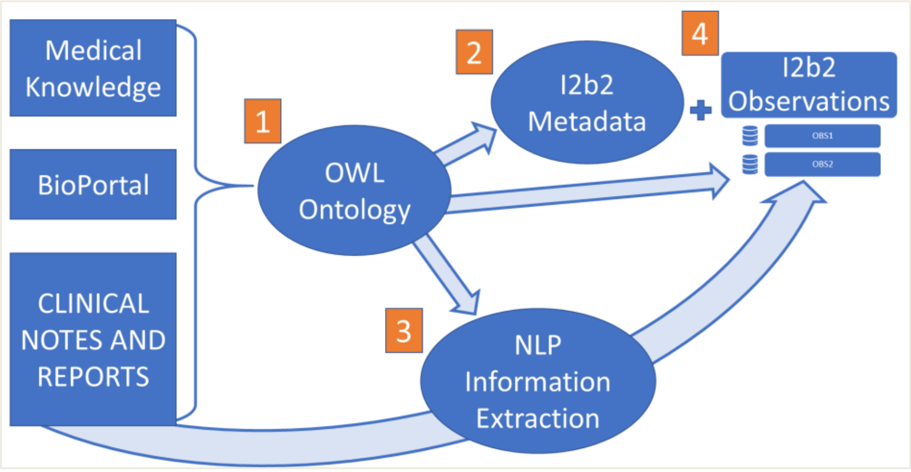


Figure 33 General NLP workflow [56]. Now after the first step of ontology development, all other steps are guided directly from properties of the OWL ontology.

The main component of this workflow is the OWL Ontology. In detail, it guides the extraction of the information from the reports, it defines the i2b2 metadata structure and helps storing the extracted data as i2b2 observations

Therefore, the Ontology entities has been enriched with Object Properties and Annotation Properties to help taxonomy creation and i2b2 data injection.

For the construction of the taxonomy in i2b2, Events and Attributes are saved as i2b2 concepts and modifiers, respectively, and two different Events can be linked together with a special relation.

Object Properties used in the taxonomy structure creation are:

- *subClassOf*: Event/Attribute class specified with this property contains the parent class and is used to maintain the original hierarchy both for Events and Attributes;
- *isModifier*: this property is used by an Attribute class to specify its target Event, so in the taxonomy the attribute will automatically become the modifier in the right place;
- *willBeModifierOf* forces an Event to be saved in the taxonomy as a modifier of anther Event, specified as the Property value.

The other Annotation Properties used to complete necessary i2b2 metadata of the concept, are the following:

- C_BASECODE, belonging to a standard or custom coding system, it is the i2b2 concept code;
- C_FULLNAME_token, containing the token of the hierarchical path that leads to the specific node through the i2b2 taxonomy. If a Class Entity does not have this Property it will be ignored during the construction of the i2b2 metadata (not all classes in ontology will become i2b2 taxonomy entries, for example the specimen number);
- C_METADATAXML, adding extra information and functionalities to a i2b2 concept or modifier inside the query tool.

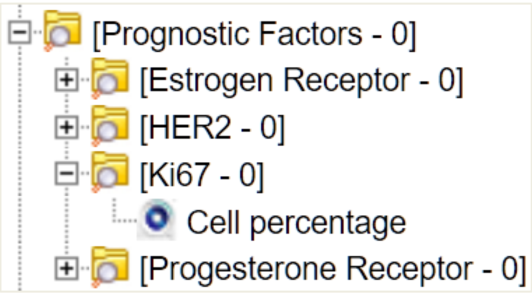
The two Object Properties *subClassOf* and *isModifier* permit to rebuild the ontology structure in i2b2 taxonomy.

In Protégé ontology, the hierarchy structure of the Event classes is represented with the *subClassOf* property. For example, the Ki-67 exam can be recognized as a subclass of the predictive and prognostic factors class:

```
"owl:Thing/Event/Predictive_Prognostic_Factors/Ki67"
subClassOf
"owl:Thing/Event/Predictive_Prognostic_Factors"
```

Event-Attribute links, instead, are represented with the *isModifier* property, which allows to the cell percentage result to be provided for the Ki-67 concept into the i2b2 taxonomy(Figure 34):

```
"owl:Thing/Attribute/Prognos-
tic_Factors_Attribute/Cell_Percentage"
isModifier
"owl:Thing/Event/Predic-tive_Prognostic_Factors/Ki67"
```



[Prognostic Factors - 0]
[Estrogen Receptor - 0]
[HER2 - 0]
[Ki67 - 0]
Cell percentage
[Progesterone Receptor - 0]

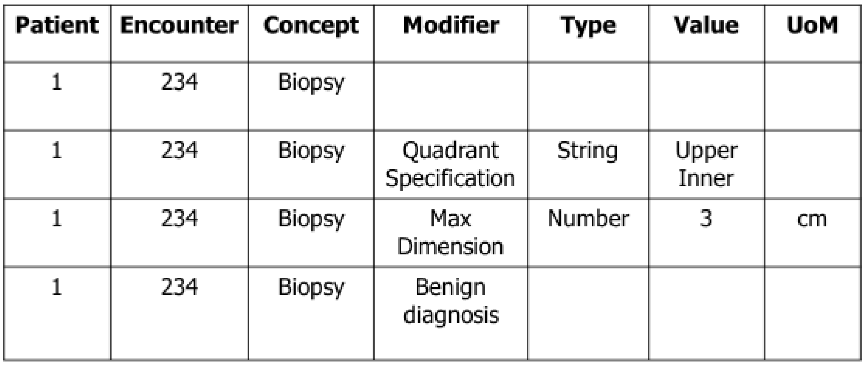
Figure 34 correct positioning of i2b2 taxonomy of ki-67 concept and cell percentage modifier.

Event-Event relations are handled, too. Thanks to the *willBeModifierOf* Object Property, the NLP extraction pipeline consider as Events Diagnosis and Specimen, while Diagnosis are saved as modifier of Specimen in i2b2.

For example, in i2b2 the “Carcinoma” Class will become a modifier while the “Biopsy” Class will remain the concept. This setting allows the selection of all the patients who were diagnosed with a lobular carcinoma with a core biopsy procedure.

Finally, an automatic ETL procedure can run automatically to load the NLP XML output Events into the observation fact table of the i2b2 data warehouse. Similarly, an Attribute of an Event is saved as child observation with the modifier as well as with the concept of the Event (Table 4).

Table 4 Example of observations created after a biopsy specimen of 3 cm from the upper inner quadrant of breast has been found to contain benign tumor cells



Patient	Encounter	Concept	Modifier	Type	Value	UoM
1	234	Biopsy				
1	234	Biopsy	Quadrant Specification	String	Upper Inner	
1	234	Biopsy	Max Dimension	Number	3	cm
1	234	Biopsy	Benign diagnosis			

3.3. An i2b2 plugin for CareFlow Mining

An important area of clinical informatics is the study and the development of tools able to extract and analyze patients' careflows from the digital data.
The application of CFM algorithm was considered by HPG23 oncologists an important add-on to support clinical oncology research. A plug-in for the i2b2 client browser seemed the most valuable way to offer the careflows discovery to researchers and clinicians.

i2b2 data warehouse has been already exploited to run careflow mining algorithms able to process heterogeneous longitudinal data [30].
The method mines specific events available from a log file and build most frequent pathways (discovery phase), identifying different patients' careflows. Afterwards, two steps of temporal and clinical enrichment are done, in which the method enlightens the importance of temporal relations between events and points relationships between the patient's clinical conditions and the extracted careflow.

The goal of the i2b2 plugin is to select from available taxonomies concepts to use in the "discovery phase" of the CFM algorithm (Figure 35).

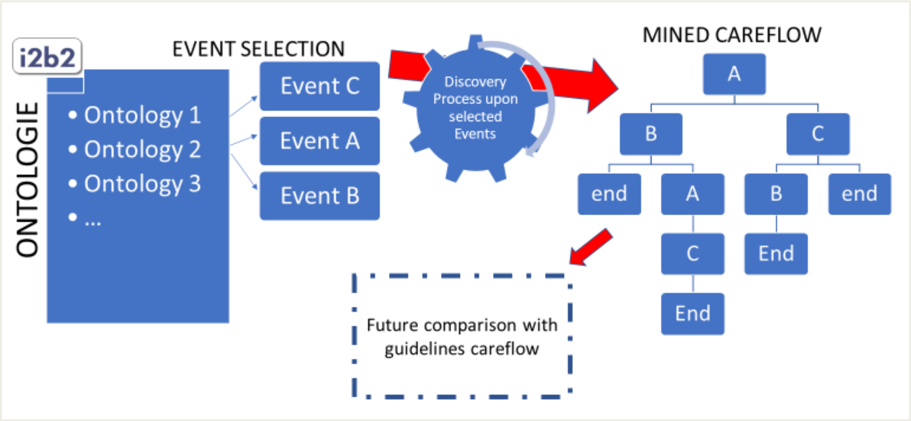


Figure 35 The idea behind the i2b2 plugin implementing CFM algorithm.

3.3.1. Plugin Mockup

The i2b2 plugin is not completed, lacking in its web-side component. Here a mockup and some use cases are presented, which had been discussed with clinicians.

The CFM plugin can be selected from the home page of the i2b2 query tool (2.2.2.2). In general, plugins implement functionalities directly into the i2b2 framework, in order to take advantage of standardized data in the data warehouse and, and the same time, to provide users with an easy access to that functionalities.

After the plugin is selected from the "Analysis tool" button, an empty page is shown (Figure 36). This page starts the preprocessing part and the event log creation (2.6.1). Here empty slots for starting population and concept of interest are provided.

The first step is to select a patient group and drag it into the Patient Set panel (Step 1 of Figure 37). In the example patients with diagnosis of infiltrating breast cancer are the starting population

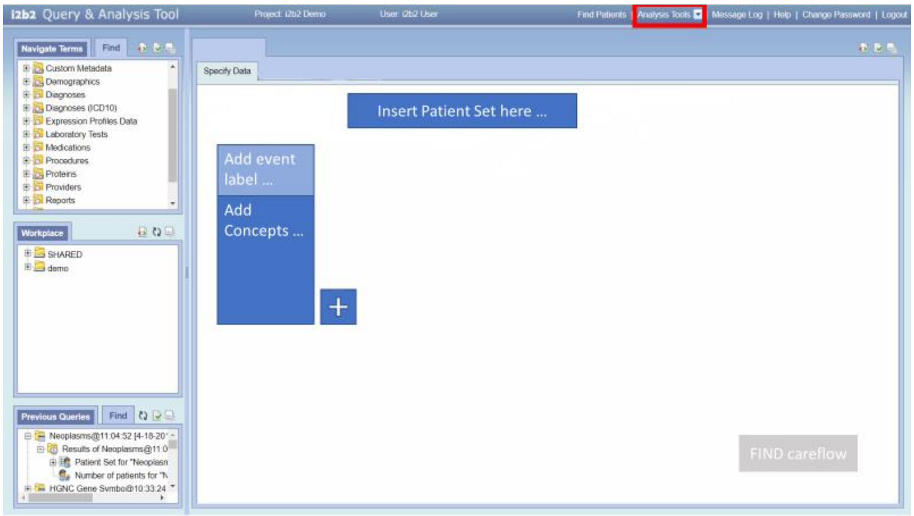


Figure 36 Start page of the plugin

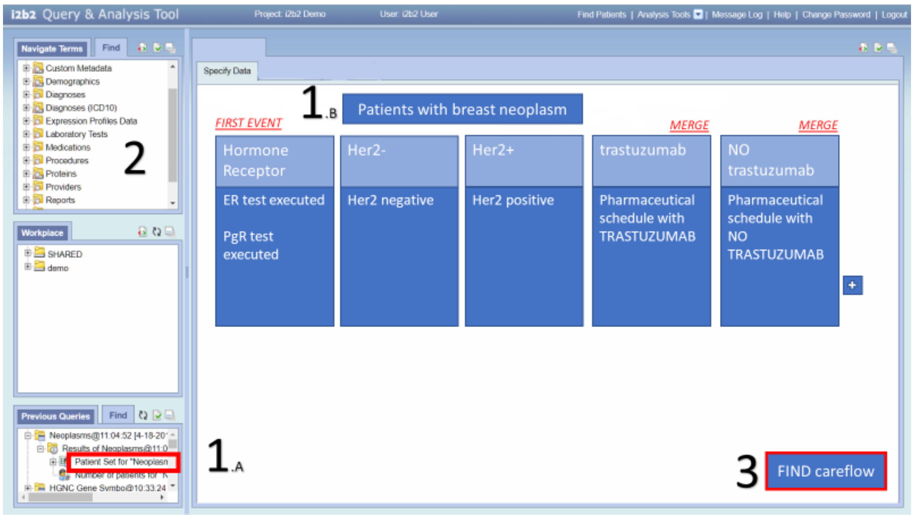


Figure 37 The plugin interface permits to implement all the preprocessing steps to create the events log from i2b2 observations.

The second step consists into drag and drop concepts from the available i2b2 taxonomies on the left of the page.

For each blue panel in the center, the user can create groups of concepts and define a labeled event. Concepts labelling of events must be carefully performed, since it drives the fundamental step of constructing an events log for the CFM.

In this simple example, five events labels are created:

- Hormone Receptor: regroups the presence of hormone receptors (Progesterone or Estrogen) assessments in the patient histories;
- HER2 -: summarizes the presence of observations of negative result of a HER2 test;
- HER2 +: like HER2 -, but with positive results;
- "Trastuzumab": the presence of drugs administrations, which has as modifier a schedule containing Trastuzumab;
- "NO Trastuzumab": like "Trastuzumab" event, but the drug schedule here does not contain Trastuzumab.

Other specifications can be selected to create the log file. For example, one event can be selected to have the "First Event" tag. All the observations of a patient history before the first appearance of this event would be ignored, so it is guaranteed that the first row of a patient in the events log would be this tagged event.

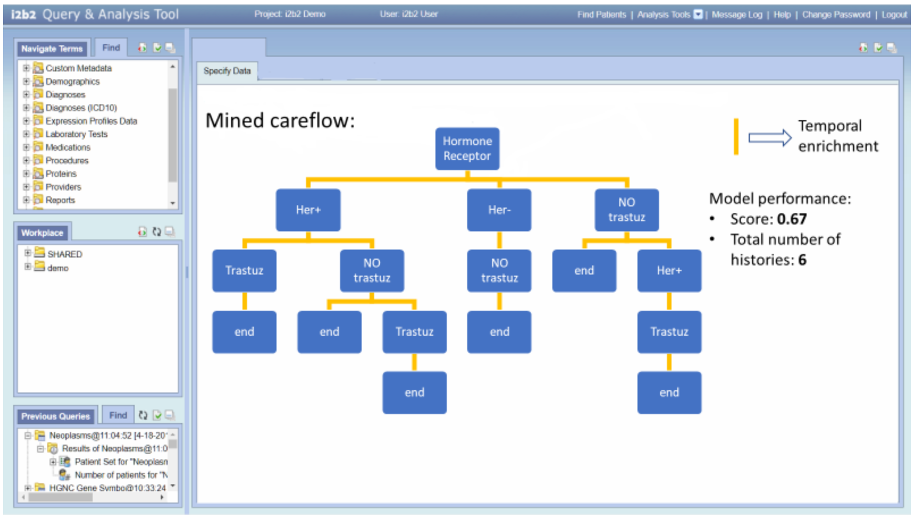
The "Merge" tag is needed when events are created with concepts that will likely have a lot of observations, in this case drug administrations ("Trastuzumab" and "NO Trastuzumab" events). The merge procedure unifies consecutive events. This would allow having a unique event that corresponds to an entire period of drug treatment.

After all the events and specifications are ready, the button of "FIND careflow" can be pushed, and a last window with insertion request of CFM parameters (minimum support and maximum history length) will appear.

After the computation, the result will be reported in the plugin page, as shown in Figure 38. The careflow model is drawn from the results of the underlying CFM algorithm, with relevant sequence of events, accordingly to our i2b2 events and CFM parameters choices.

The performances of the proposed result are listed, giving the user the ability to be aware of the number of histories produced and the data representation performance of the entire model.

With this application, the careflows of the patients' group of interest can be investigated, helping the user to quickly decide if the configuration of a study or research has enough support in terms of cases. Patients that belong to different branches will be further analyzed offline, by studying, for example, clinical outcomes, such as disease-free survival or time to recurrence.



Model performance:
- Score: **0.67**
- Total number of histories: **6**

Figure 38 The careflow is associated with a performance score that is indicative of how much the underlying evens are represented in the careflow. Within this example, a patient that have an event history of “Hormone Receptor -> NO trastuzumab -> trastuzumab” has a representation in the careflow model only for his two firsts events. The score for this patient would be 0.67, while the overall model score will be the mean of each patient score. To change the model score and number of histories developed, the CFM parameter should be changed accordingly

3.3.2. Plugin use case

A possible use case of the plug-in is the evaluation of CFM in terms of adherence to the hospital guidelines.

In Figure 39 it is shown a subpart of AIOM 2016 breast cancer guideline [57].

From this case, one of the suggestions, after a first hormone receptor assessment, is to follow a therapy line with trastuzumab in case of HER2 positive result.

The events built in the previous example of Figure 37 reconstruct a possible check of this guideline indication: the starting population is drawn from patient with infiltrating breast cancer diagnosis, and the events of HER2 +/- and trastuzumab treatment schedule match the AIOM guideline steps.

Figure 40 depicts the resulting careflow model.

The HER2 status splits the population in 2 sub-groups right after the first event, with positive and negative patients. The third line seems to show a weak trastuzumab segregation between the two groups, which after a month begin treatments with or without trastuzumab.

Nevertheless, looking at the durations of these treatment events, more information is needed to differentiate between the two cases: HER+ patients maintain the trastuzumab treatment three times more than the time of the HER2- patients.

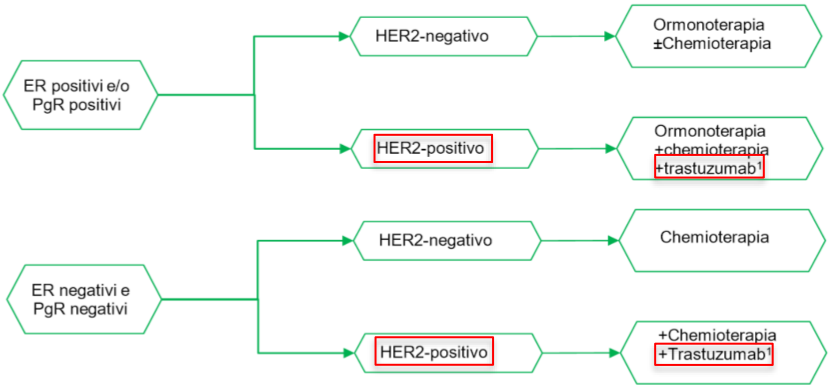

ER positivi e/o PgR positivi
HER2-negativo
Ormonoterapia ±Chemioterapia
HER2-positivo
Ormonoterapia +chemioterapia +trastuzumab1
ER negativi e PgR negativi
HER2-negativo
Chemioterapia
HER2-positivo
+Chemioterapia +Trastuzumab1

Figure 39 Adjuvant treatment choice in case of infiltrating breast cancer, based on predictive factors (hormones receptors and HER2 status)

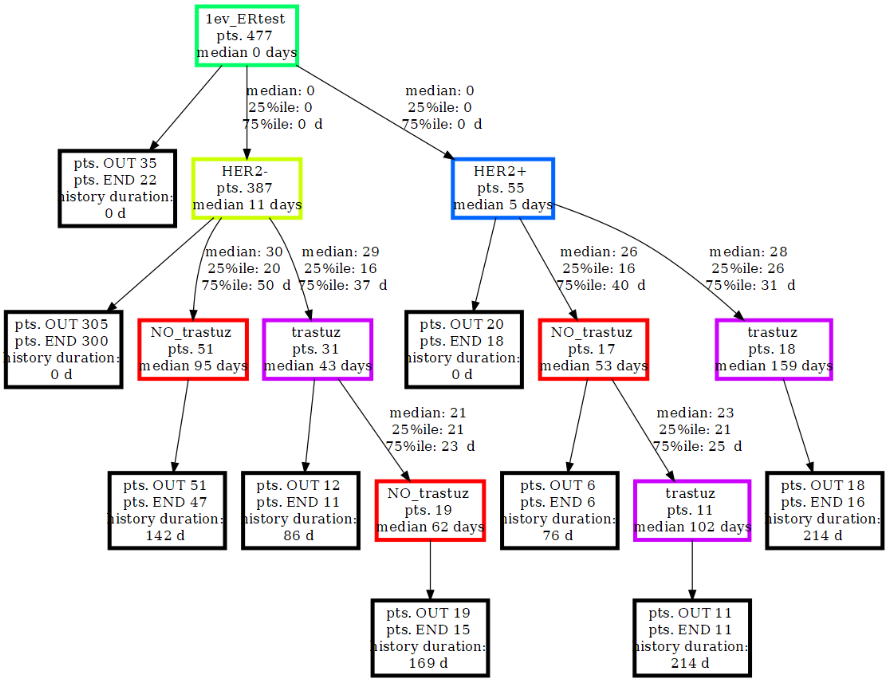

1ev_ERtest pts. 477 median 0 days
pts. OUT 35 pts. END 22 history duration: 0 d
median: 0 25%ile: 0 75%ile: 0 d
HER2- pts. 387 median 11 days
median: 0 25%ile: 0 75%ile: 0 d
HER2+ pts. 55 median 5 days
pts. OUT 305 pts. END 300 history duration: 0 d
median: 30 25%ile: 20 75%ile: 50 d
NO_trastuz pts. 51 median 95 days
median: 29 25%ile: 16 75%ile: 37 d
trastuz pts. 31 median 43 days
pts. OUT 20 pts. END 18 history duration: 0 d
median: 26 25%ile: 16 75%ile: 40 d
NO_trastuz pts. 17 median 53 days
median: 28 25%ile: 26 75%ile: 31 d
trastuz pts. 18 median 159 days
pts. OUT 51 pts. END 47 history duration: 142 d
pts. OUT 12 pts. END 11 history duration: 86 d
median: 21 25%ile: 21 75%ile: 23 d
NO_trastuz pts. 19 median 62 days
pts. OUT 6 pts. END 6 history duration: 76 d
median: 23 25%ile: 21 75%ile: 25 d
trastuz pts. 11 median 102 days
pts. OUT 18 pts. END 16 history duration: 214 d
pts. OUT 19 pts. END 15 history duration: 169 d
pts. OUT 11 pts. END 11 history duration: 214 d

Figure 40 CFM Acyclic Diagram result. Parameters of the run: minimum patient support of 10; maximum history length of 5. Model performance: score of 0.93; 9 number of histories

Moreover, on the fourth event line, it is possible to observe that HER2+ patients who had a brief NO trastuzumab schedule have later initiated a trastuzumab treatment type.

Possible non correspondences to clinical guidelines can be further analyzed offline, using the sub-groups created by the CFM plugin, searching for cases of co-morbidities, disease severity, toxicity or other causes that could have hindered, for example, the immediate trastuzumab therapy line to HER2+ patient.

Chapter 4

Results

This chapter describes implementation’s outcomes:

Data currently stored in the i2b2 data warehouse are presented, providing examples of the studies that have been carried on thanks to these results.

A validation of the Natural Language Processing (NLP) Information Extraction (IE) task is presented, studying the performance of the actual pipeline.

The application of the CFM algorithm in a similar scenario has been proposed as proof of concept to generate and discover novel aspects of the data in the oncologic research.

4.1. i2b2 Deployment

After the implementation work of the ETL procedure in Mirth (3.1), the result is a weekly process that load data from various sources into i2b2 data warehouse.

Data were collected from 2007 to September 2019, for all patient across the hospital.

The i2b2 entities account for:

- Total Patients: 937.956
- Total Encounters: 10.641.975
- Total Observations: 141.576.455

Some aggregated counts of patients and encounters over the years are reported below:

- Patients first appearance in i2b2 database (Figure 41)
- Unique patients treated in the hospital (Figure 42)
- Encounters performed in the hospitals (Figure 43)

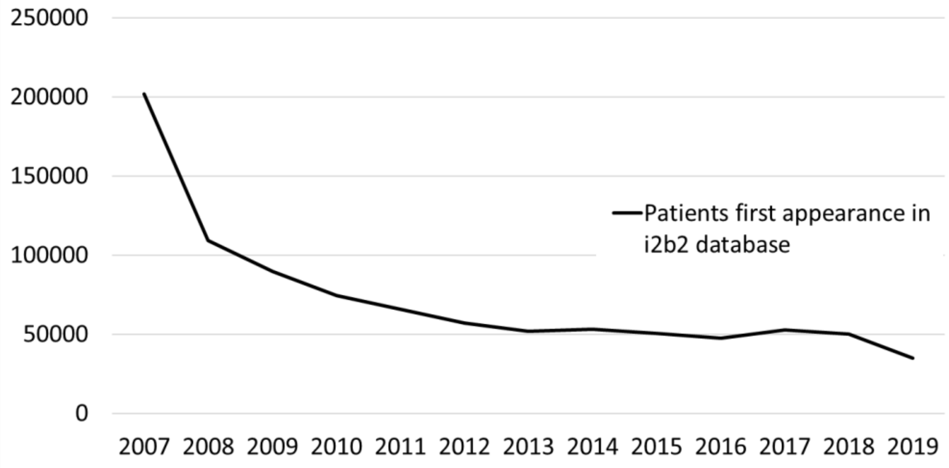


Figure 41 Patients first appearance in i2b2 database. At the first year recorded in i2b2, 2007, the number of new patients were equals to treated patients of Figure 42. Year by year, patients considered new has stabilized and the increase population of patients involved in healthcare inside the hospitals is 50’000 (ca) units each year

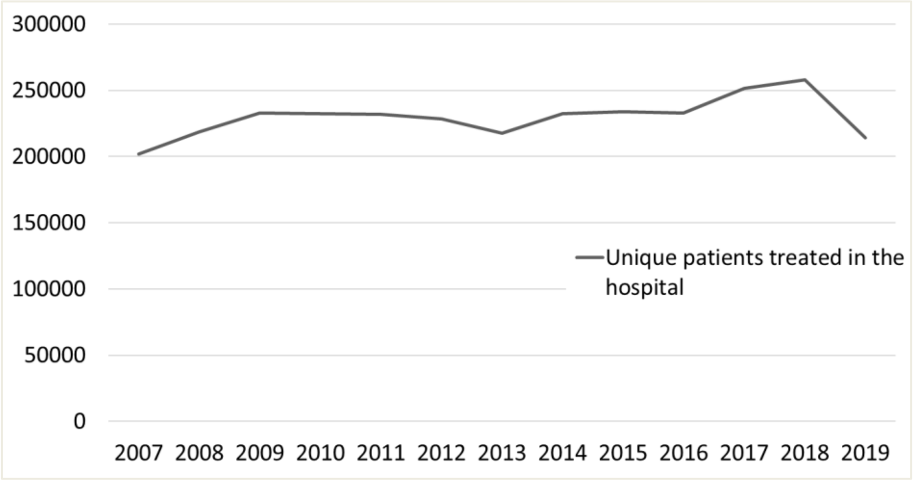


Figure 42 Unique patients treated in the hospital. From 2007 to 2018, the number of distinct patients treated in the hospital had an 20% (ca) increase.

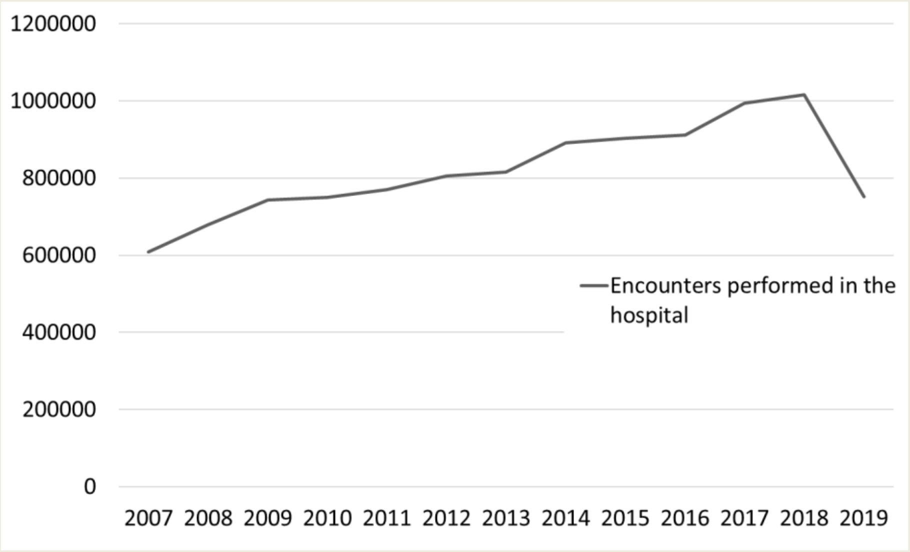


Figure 43 Encounters performed in the hospitals. Encounters, in general, in the hospital has increased by 67% (ca) from 2007 to 2018. Even if we do not consider the first 2007-2008 period where there was no pathology reports recordings, still it is possible to note a 37% (ca) increase from 2009 to 2018, unique patient treated (Figure 42) increased of 11% (ca) in the same time window.

Observations are stratified by data-source as described in Figure 44. Most observations derive from laboratory analysis and drug administrations. Also, outpatient services produce a fair amount of observations.

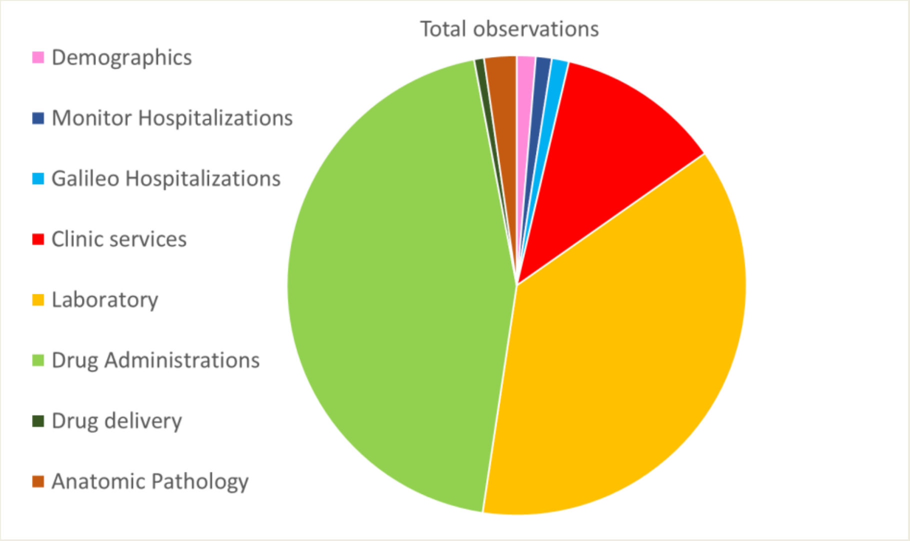


Figure 44 Observation proportions by data source.

In the following pages, counts and proportions are presented for each data source, separately.

Observation of “Contacts” (3.1) are present for each data source and reveal the number of encounters performed.

Demographics

Demographics observations contains data about gender, age and vital status, present in proportions shown in Figure 45. Vital status information is very rare in the Demographics view, because describe the death date of patients, and few of them have actual death date field filled.

- Total observations: 1.885.418
    - Gender 937.932
    - Age 937.932
    - Vital Status 15.928

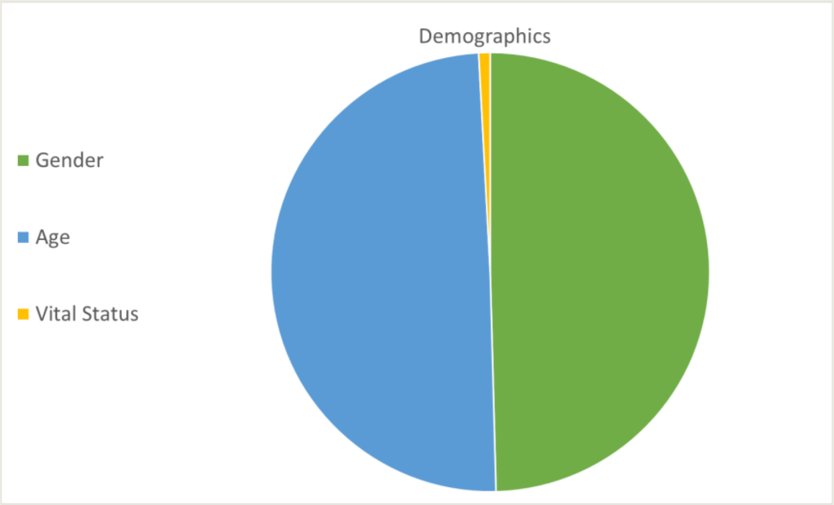


Figure 45 Observation proportions from Demographic data source

Monitor Hospitalizations

Hospitalizations from 2007 to July 2012 produced contacts and ICD9 observations. It is clear from Figure 46 that the proportion of contacts observations is quite high, since per each hospitalization, only one or two ICD9 observations are created. Diagnosis and Procedures are equally represented.

- Total observations 1.585.471
    - Contacts 563.794
    - ICD9 Diagnosis 503.633
    - ICD9 Procedures: 518.044

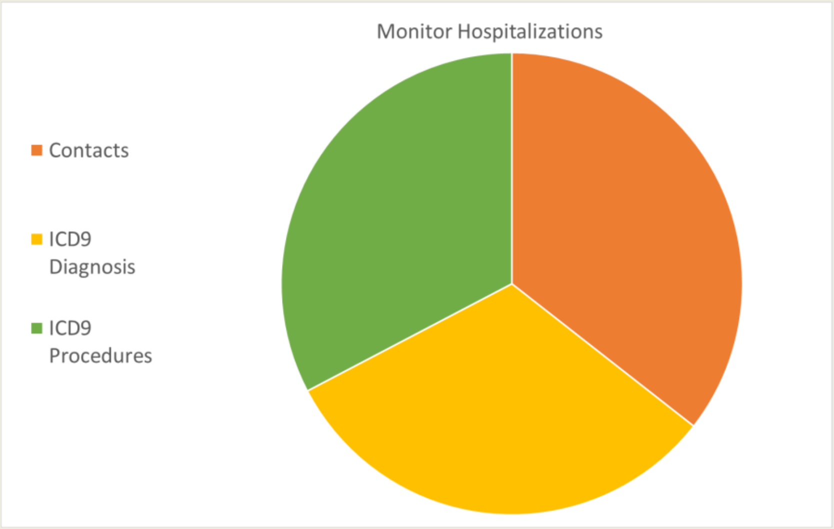


Figure 46 Observation proportions from Monitor hospitalizations data source

Galileo Hospitalizations

Hospitalizations from 2012 to September 2019 produced contacts and ICD9 observations. Like Monitor observations, the proportion of contacts observations (see in Figure 47) is quite high, and diagnosis and procedures are equally present.

A tiny amount of vital status observations is measured, thanks to the information of discharge modality, in case of patient's death at discharge.

- Total observations 1.699.622
  - Vital Status at discharge 6.374
  - Contacts 642.016
  - ICD9 Diagnosis 495.174
  - ICD9 Procedure: 556.058

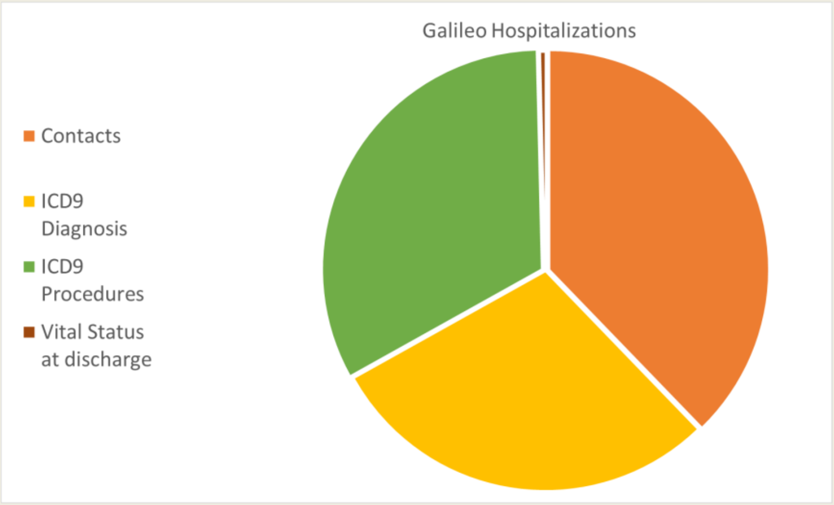


Figure 47 Observation proportions from Galileo hospitalizations data source

Clinical outpatients’ services

Outpatients contacts and ICD9 procedures observations are in same number, showing that few procedures are done in the same encounter. In Figure 48 shows that contact observations are larger than actual ICD9 procedures. This is caused by the mapping of the SISS procedures of the source data into the ICD9 codification system (3.1.3).

However, these type of contact observations often allows finding procedures missed by the mapping procedures. They contain, in a dedicated observation blob field, all the original procedures performed in a clinical encounter. Procedures saved in this textual format inside contact observations are represented using the hospital internal clinical services code.

- Total observations 16.436.379
    - Contacts 8.736.239
    - ICD9 Procedures 7.700.140

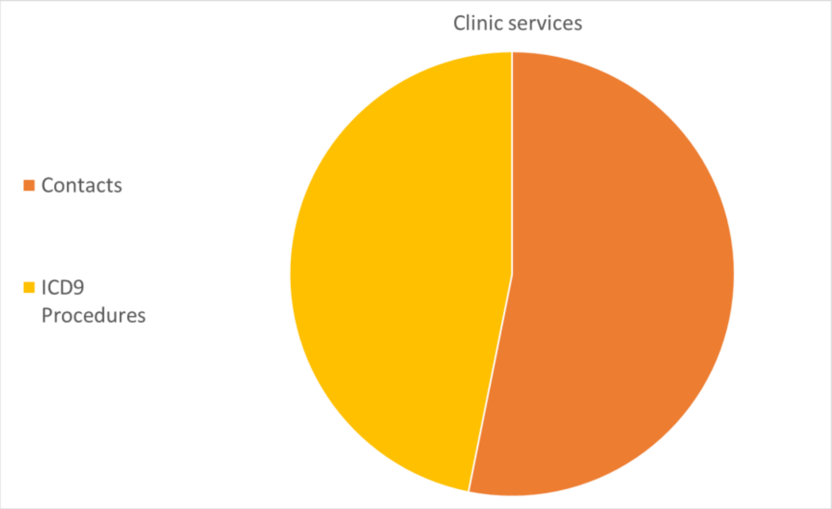


Figure 48 Observation proportions from Clinics data source

Laboratory

Observation proportions in Figure 49 show two interesting aspects.

First, many analysis observations are presents for each lab encounter, so that contact observations are a small proportion of the total laboratory observations.

Second, as anticipated in section (3.1.4), the LOINC mapping of less than 100 different laboratory analysis results generates a LOINC observation proportion that cover 85% (ca) of total laboratory observations.

- Total observations 52.543.286
    - Contacts 1.599.083
    - Internal code 7.801.098
    - LOINC 43.143.105

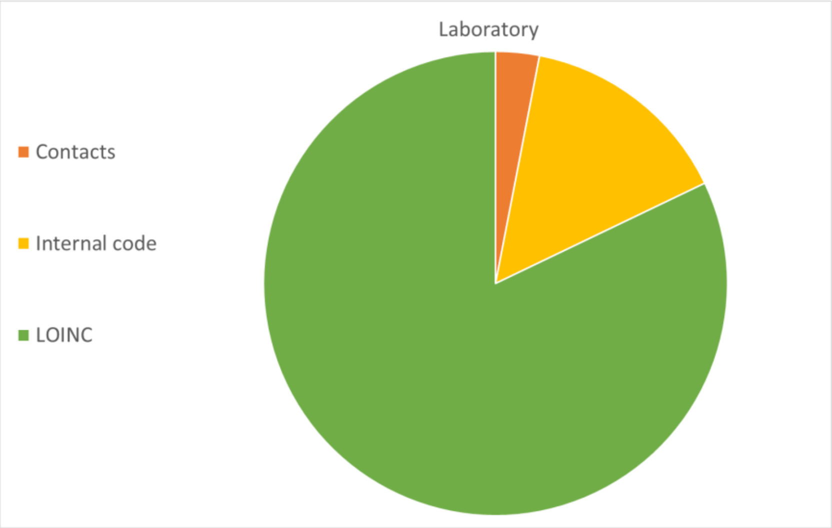


Figure 49 Observation proportions from laboratory analysis data source

Drug Administrations

Drug administrations (Figure 50) show results similar to the laboratory analysis ones, with many administration observations in each single contact.

More details are shown in Figure 51, where observation types are listed. Here is clear that most part of administrations in the hospital is saved with information about dosage and administration route. Only a minority of administrations can be referred to a schedule type.

- Total observations 63.209.076
    - Contacts 1.342.774
    - ATC administrations 61.866.302
        - Dose info 30.189.717
        - Schedule type 1.486.868
        - Administration route 30.189.717

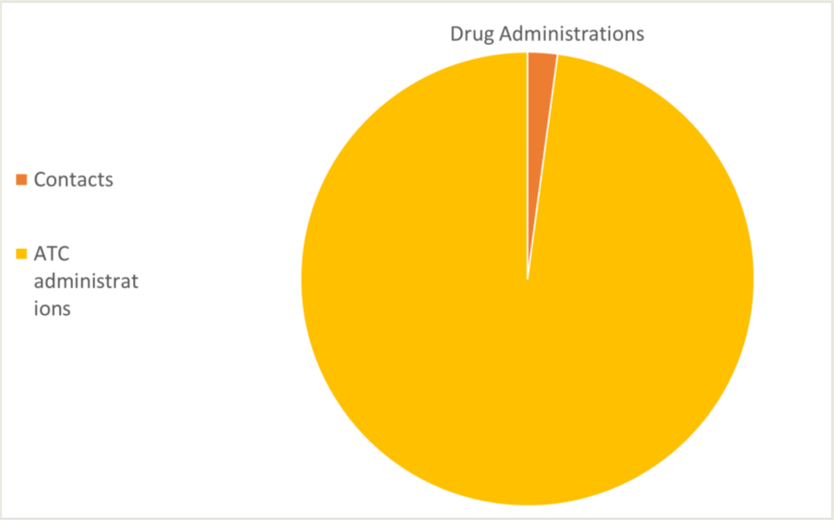


Figure 50 Observation proportions from in hospital drug administration data source

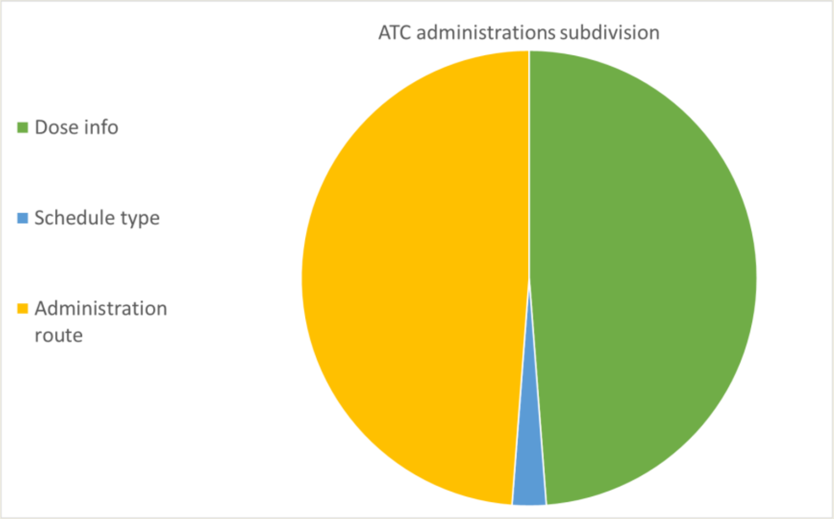


Figure 51 Observation proportions from ATC administration subgroup.

Drug Deliveries

This data source type, even if related to drugs, has different proportions on contact and actual delivery observations (Figure 52). This is caused by the very nature of this data source.

In clinical practice, drug delivery to patient consists in directly giving to the patient the medication, which is supposed to be taken at home. All the necessary drugs contained in the package will be assumed at home in more than one times, but in the system only one observation of actual delivery is created.

Instead, each individual drug administration performed in the hospital is individually reported. Moreover, as we have just seen, a single administration event can be represented into i2b2 up to three different observations, in order to contain associated drug administration information, such as schedule, dosage and route.

- Total observations 988.432
    - Contacts 377.157
    - ATC Deliveries 611.275

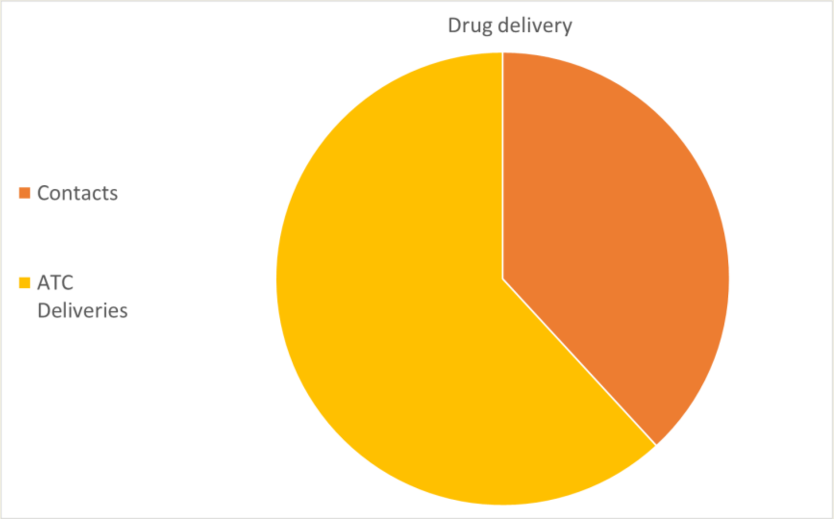


Figure 52 Observation proportions from drug delivery data source

Anatomical Pathology

Pathology observations are related to textual information from the reports. Since each report generates more observations, one for each present section in the report itself, the proportions shown in Figure 53 describe a low presence of contacts observations.

Each contact represents a single report: in fact, numbers of contacts are equal to number of mandatory observations, i.e. Specimen and SNOMED morphology.

- Total observations 3.228.771
    - Contacts 540.761
    - Reports text 2.688.010
        - Specimen description 540.761
        - Specimen SNOMED morphology 540.761
        - Specimen SNOMED topography 436.837
        - Specimen SNOMED procedure 32.054
        - Anamnestic information 56.692
        - Clinical information 183.436
        - Macroscopic description 337.516

Microscopic description	56.692
Diagnosis	540.744
Preliminary diagnosis	9.309
Comment	7.593
Addendum	2.307

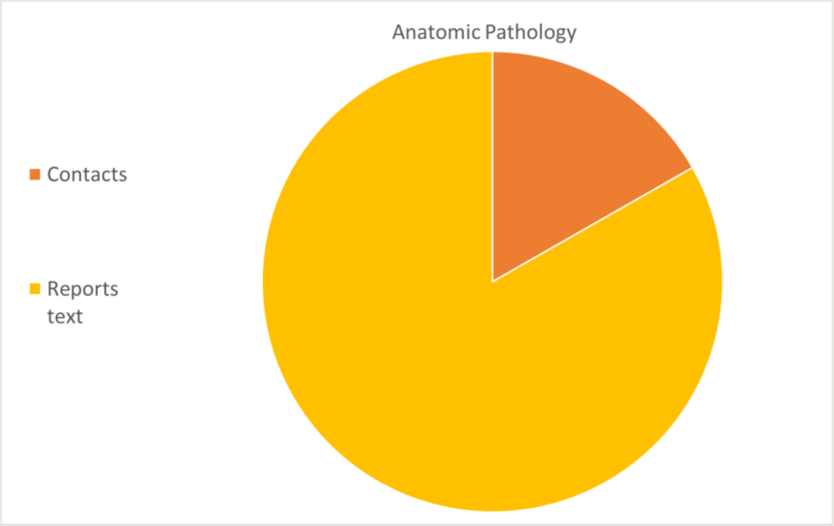


Figure 53 Observation proportions from anatomical pathology data source

Figure 54 shows the details of each type of report observation, corresponding to the sections of the reports. This can display which are the most used sections in the reports.

The SNOMED procedure, anamnesis, microscopic description, preliminary diagnosis and comment sections are the less used.

The specimen description is fundamental to the existence of the report itself, so it is obviously mandatory. Also, clinical information is quite used, and describe previous diagnostic exams or important reports.

Macroscopic description and diagnosis sections are the main containers of information inside the reports. The diagnosis section contains not only final diagnosis decided in the report, but also useful Prognostic Predictive factors results.

SNOMED section for morphology and topography have great presence in the reports (SNOMED morphology is mandatory at the time of report creation by pathologist). The import of SNOMED codes has been considered, but the final decision of omit them is due to their lack of reliability, confirmed by clinicians. Instead of giving a weak and potentially dangerous variable to i2b2 user, the choice to not use these codes was made.

As future perspective, it could be interesting to study if some stratification of these SNOMED variable, maybe by year or by pathologist, could contain reliable group of information, and try to import into i2b2 only this subgroup of data.

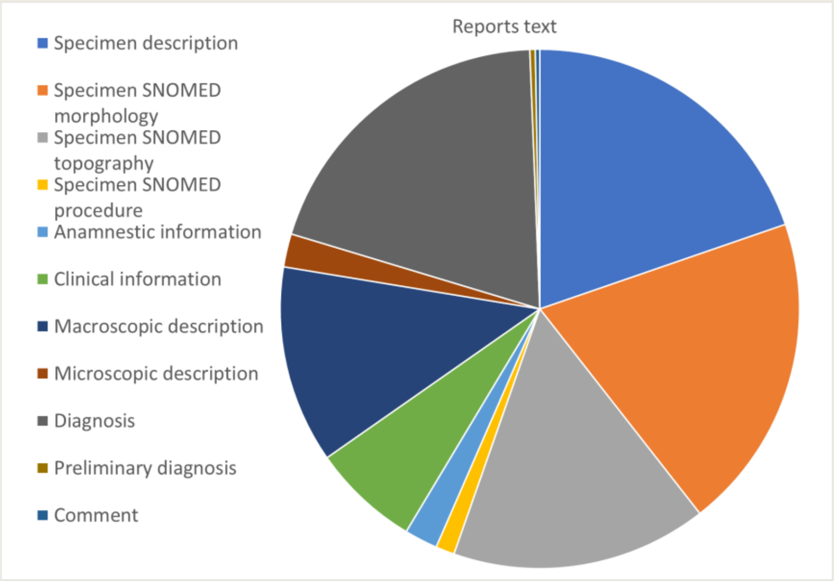


Figure 54 Observation proportions of the report sections.

4.2. i2b2 Oncology Project

The oncology ward was involved in all previous database aspects creation, but they needed a proper i2b2 project for patients' privacy reasons, focused on oncologic patient only. The structure of i2b2 platform is applicable also for this purpose.

In Figure 55, the i2b2 "horizontal" project includes all hospital patients and contains data of previously described data sources. Within this structure, many other "vertical" projects can be created, using only a sub-group of the original patients and adding independent data sources.

In our case the vertical project is the i2b2 Oncology Project. Oncology patients (provided by HIS) from horizontal project are imported, with all their encounters and observations. The new vertical project can rely on the same taxonomy structures of the horizontal one.

Actual i2b2 entities (and relative proportion to the horizontal entities) in oncology project drown by the horizonal project are:

• Patients:	28.301	of 937.956	(3% ca)
• Encounters:	1.527.653	of 10.641.975	(14% ca)
• Observations:	23.625.083	of 141.576.455	(17% ca)

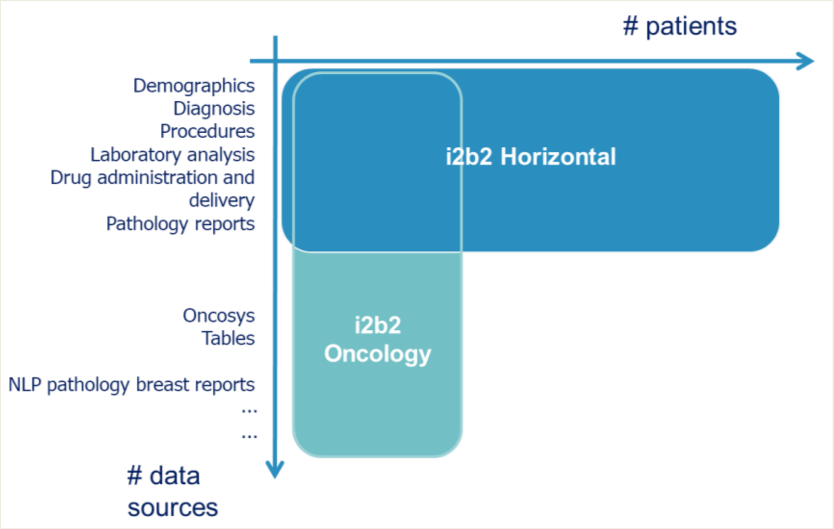


Figure 55 Visualization of the horizontal project, in dark blue, and the vertical oncology project, in light blue

Entities regarding oncologic patients seem to be more complex. Even if the proportion of patients relative to the horizontal project is only 3%, these patients account for more encounters and observations in the horizontal project. This can be caused by the continuity of care, exams and procedures that this kind of patients undergo in the hospital compared to other patients present in the horizontal project.

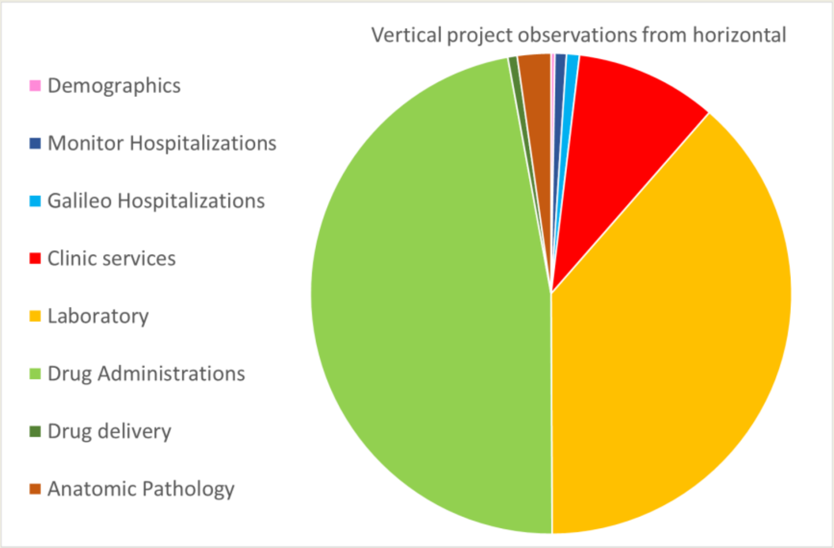


Figure 56 Observation proportions by data source. These observations are a fraction of those present in the horizontal project, imported in the vertical project because linked to Oncology patients.

The observations taken from horizontal i2b2 project are stratified by data-source as described in Figure 56. No major variations are evident respect horizontal project proportions in Figure 44.

Moreover, additional data had been added from the internal oncology ward Informative Systems, Oncosys. This allows the clinicians to enrich their query with concepts that are present only in their data systems. Oncosys internal tables have been imported in the i2b2 oncologic project with methods similar to the horizontal data sources.

Even if the detailed description of these data has been skipped, the resulting entities added to the oncologic project are:

- Total Encounters: 1.218.853
- Total Observations: 2.607.834

Some additional NLP extractions and other observations have been manually added to Oncologic Project in order to fulfill the requirements of a specific study, named HER2-BC-FROM. This study is conducted under the supervision of FROM (Fondazione per la Ricerca Ospedale Maggiore Bergamo) [58] and it is conducted on patients with breast cancer (BC) with HER2 positive result in the last 10 years.
This study does not require the addition of any further data source to the project, but some observations has been specifically created:

- Total observations 162.363
  - Treatment type 131932

    - Early Breast Cancer 70239
    - Metastatic Breast Cancer 61693
  - NLP prognostic factors 30431
    - Estrogen test 7291
    - HER2 test 16059
    - Ki-67 test 7081

"Treatment type" observations were created using administration schedule observations. With indications of oncologists, all the drug schedules used in the oncology ward are classified as early or metastatic stage of the cancer.

"NLP prognostic factors" observations have been imported manually into i2b2 vertical project after the implementation (3.2.1) and validation (4.3) of the Oncology and information extraction works. For this study 8.849 reports were found containing breast cancer evidences and prognostic factors extractable info.

Overall, in the vertical Oncology i2b2 project, data brought from different sources or created for different purposes are coexisting. In Figure 57, observations proportions are shown. Most of the observations belong to horizontal data extraction of oncologic patients.

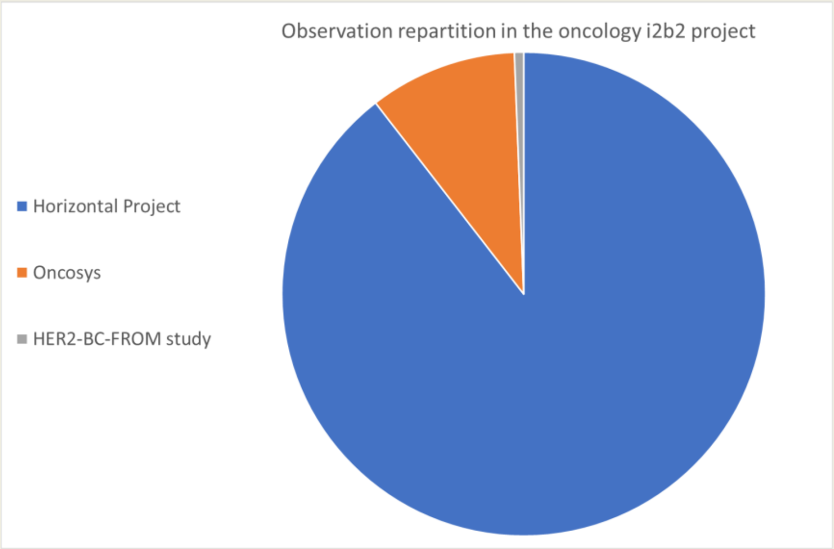


Figure 57 View of all the observation presents in the oncology i2b2 project.

4.2.1. HER2-BC-FROM study

The main goal of the study is to retrospectively describe the HER2 positive breast cancer cases, from September 2007 to September 2017, using the implemented i2b2 data-warehousing system.

During disease progression, patients experience numerous events such as hospital admissions and discharges, laboratory tests, and follow-up visits. These events occur in a temporal sequence.

i2b2 has added a fundamental value to the study, since it has been designed to simplify the process of using existing, deidentified, clinical data for preliminary research cohort discovery, thanks to its data model; with its natural purpose of identify patient stratifications, the i2b2 tool fits the study approach of coupling mining methods with temporal characterization of the transitions between breast cancer related events.

The structured information in i2b2, and the possibility to easily analyze the data using the i2b2 query tool, are methods of relevant importance in diseases studies. In oncology, and particularly in BC, different medical approaches are requested, and significant progress can be generated in the search for effective treatment able to maximize the patients' clinical outcome.

Thanks to the efforts of this study, two poster works has been accepted to AIOM 2018 [59] and 2019 [60] congress.

In the first study [61], 4.239 breast cancer patients from 2007 to 2017 were identified. The results show how i2b2 lets the users investigate the data, doing a stratification of the patients using HER2, treatment type, hormone receptor assessments and positive lymph nodes variables (from TNM code, using N classification)

The second study is focused on 531 HER2 positive patients characterized with early breast cancer treatments.

Kaplan-Meier estimates were used to evaluate cumulative incidence of relapse and survival, stratifying for the tumor burden (from TNM code, using T and N classifications) and tumor biology (tumor grade, estrogen receptor results).

A final Multivariable Cox proportional-hazard models were applied to estimate Hazard Ratios of relapse and death, adjusting for potential confounders. This analysis confirmed positivity of lymph nodes, negative estrogen receptors and age>65 as independent risk factors of BC relapse and death, in line with the literature reported evidences [62,63].

4.3. Evaluation NLP pipeline

The validation of the Information Extraction (IE) [53] has been conducted randomly extracting a corpus of 221 anatomical pathology reports belonging to patients with breast cancer, provided by the Papa Giovanni XXIII Hospital (HPG23).

We used a set of 20 randomly selected reports out of 221 to design the ontology structure and entities. These reports were manually reviewed and discussed with physicians to recognize relevant concepts and their variants to be included in the ontology.

As previously described in the NLP pipeline implementation (3.2.1) the most relevant entities extracted are Specimen, Diagnosis, Histopathological stage and Prognostic factor. Events and their attributes related to these entities were searched within the text, and, thanks to the presence of the specimen number attribute, it was also possible to create the Event-Event relations between the specimen and the diagnosis events.

In order to make a preliminary evaluation of the performance of the pipeline we randomly selected 34 documents out of 201 and we automatically extracted the items (both Events and Attributes) from each report (system items) (Figure 58). A total of 476 system items were identified, corresponding to an average of 14 items per report.

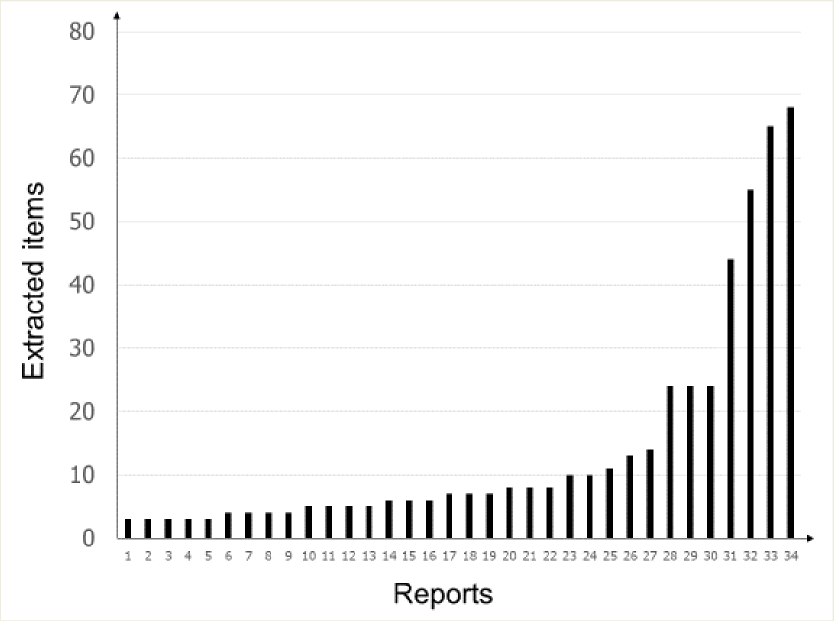
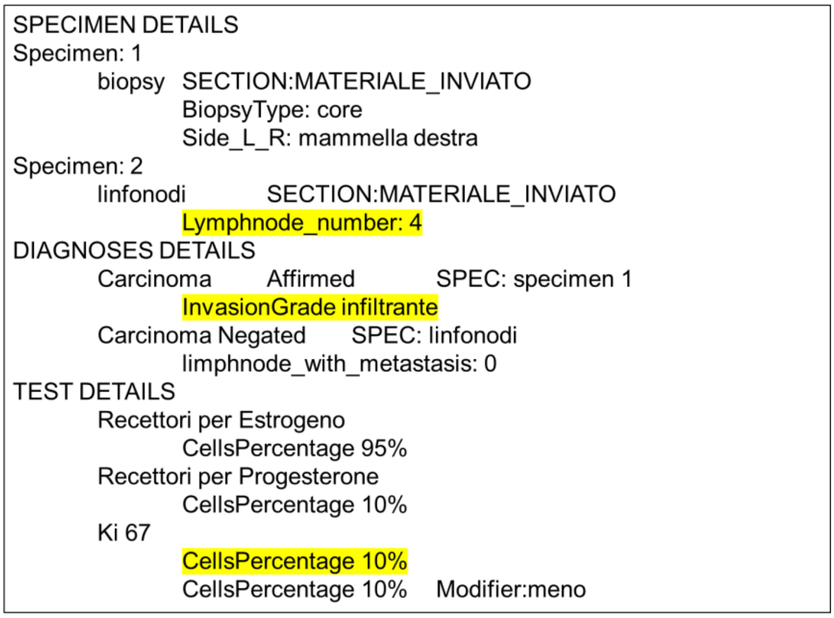


Figure 58 Distribution of the number of extracted items from 34 pathology reports.

The output of the IE process, adapted for validation task, is a text file including both the original report and the system items (Figure 59).



```
SPECIMEN DETAILS
Specimen: 1
        biopsy  SECTION:MATERIALE_INVIATO
                BiopsyType: core
                Side_L_R: mammella destra
Specimen: 2
        linfonodi       SECTION:MATERIALE_INVIATO
                Lymphnode_number: 4
DIAGNOSES DETAILS
        Carcinoma       Affirmed        SPEC: specimen 1
                InvasionGrade infiltrante
        Carcinoma Negated       SPEC: linfonodi
                limphnode_with_metastasis: 0
TEST DETAILS
        Recettori per Estrogeno
                CellsPercentage 95%
        Recettori per Progesterone
                CellsPercentage 10%
        Ki 67
                CellsPercentage 10%
                CellsPercentage 10%    Modifier:meno
```

Figure 59 An example of section of the documents created for validation purposes. The extracted items are listed aside of the original document (not showed). Clinicians were trained to make revision of this type of output.

Domain experts manually reviewed the output file in order to identify three types of errors:

- Missing items, i.e., relevant concepts not considered and thus not included in the ontology.
- False negatives (FN), i.e., information that should have been extracted but not detected by the pipeline.
- False positives (FP), i.e., extraction errors, such as incorrect specimen-number associations or attributes linked to the wrong event.

These last two errors can be very deleterious for the accuracy of the method because a FN event can lead to further wrong attribute associations.

Table 5 shows both the total and the deduplicated number of items found for each type of error. In fact, the same error can occur in several reports; for example, this is the case of the prognostic factor "c-erbB2".

The discrepancy between row and distinct counts in the table highlighted the necessity to improve the ontology in order to consider more concepts. In particular, at least the 38 relevant items that were not previously considered in the existing ontology can be included.

Table 5 Evaluation results error types

	Raw count	Distinct count
Missing items	57	38
FN items	15	11
FP items	26	21

Once the error types were identified, we evaluated the performance both IE system itself and the Ontology-IE overall system (Table 6).

For the IE-system performance, we used the raw counts of 15 and 26, for FN and FP respectively, while for the Ontology-IE overall system we used both missing items and FN items (15+57) as total FN row count.

The Ontology-IE performances were considered enough accurate (after the fundamental addition of "c-erbB2" alias to the HER2 event) to be used in the HER2-BC-FROM project (4.2.1), and the pipeline was used to extract prognostic factors by medical reports from 2007 to 2017.

A more complete NLP validation will be developed once completed the general pipeline described in chapter 3.2.3, including also validation of the created i2b2 observations.

Table 6 Performance measures of Precision, Recall and F1.

	Formula	IE system itself	Ontology-IE overall system
Precision	$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$	94,5%	94,5%
Recall	$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$	96,8%	86,2%
F1	$2\frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precisio} + \mathrm{Recall}}$	95,6%	90,0%

4.4. CFM application

During the PhD course, we had the opportunity to implement and test a pipeline based on Topic Modeling (TM) [64,65], and the Careflow Mining (CFM) algorithm described in chapter 2.6. The study was focused on the electronic patient phenotyping in the hospital IRCCS ICS Maugeri [66] of Pavia (ICSM).

The aim of the work [67] was to start from administrative and clinical data of breast cancer patients, to cluster them in terms of patterns of care (careflows), and to highlight potential relationships with clinically significant outcomes, such as the disease-free survival and the overall survival.

We selected Electronic Health Records (EHR) of a cohort of 3000 patients; patients' data has been collected from the first surgical procedure after a breast cancer diagnosis.

This kind of patients are exposed to a series of hospitalizations and Short Procedure Unit (SPU) visits, which deliver the therapy, perform additional surgical interventions or must deal with possible complications of the disease or treatment.

The two data sources used are ICD9-CM procedures of the Hospital Information Systems (HIS) and a registry of clinical and molecular data, collected by the Oncology ward service, including chemotherapy treatments.

Each hospitalization has been paired with a document, containing all the ICD9 procedures occurred inside the hospitalization itself.

Then, all the documents have been processed with Topic Modeling (TM). We used this method to assign a label to each hospitalization, summarizing the procedures described in each document.

After the TM step, the event log is created. In the log file, each event is a hospitalization: the event is named with the topic assigned to the considered hospitalization, and the start and end timestamps of the event correspond to hospitalization admission and discharge.

Therefore, the CFM algorithm has been applied to the event logs derived from the TM step. A grid-search approach was used to select the parameters of the CFM. We range from 2 to 50 the minimum support of patients and from 3 to 10 the maximum history length evaluated for each patient.

As is possible to see on heatmaps in Figure 60, true match rate indicator must be maximized while the number of mined careflows indicator must be minimized. In both indicators, red values should be avoided, so, discarding these values, optimal values for both parameters range from 5 to 10. To maintain less than 100 number of careflows we selected 10 as minimum number of patient and 10 for maximum history length given that same results are obtained from 5 to 10 values.

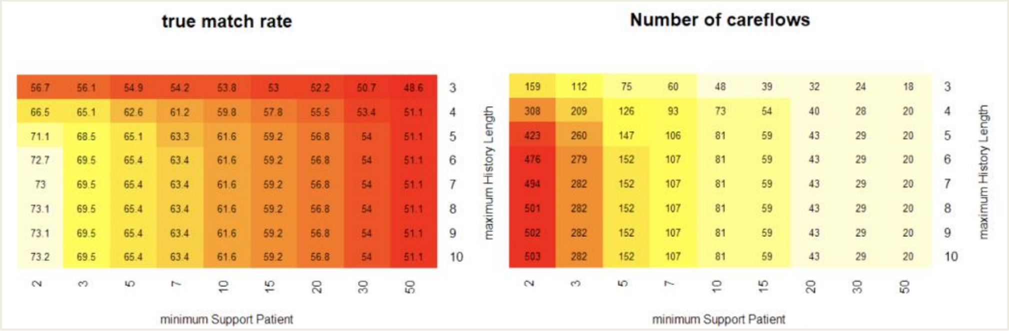


Figure 60 Visualization for the resulting true match rate and number of mined careflows. Red values should be avoided, choosing minimum support and maximum history length accordingly [67].

The result of the CFM application with selected parameters are 81 careflows, containing 160 events, 19 of which are distinct events (Table 7).

The longest careflows comprise of 5 events, and an average history length of 3.33 (SD = 0.9) and a median equal to 3.

Table 7 List of the 19 distinct events resulting after the application of the CFM step. In column 1 is shown the topic label name while in column 2 the most informative word (retrieved by the TM model) of the document containing hospitalization procedures. This describes the final composition of the events name of the event log.

Events Label	
Chemotherapy	biopsy
Chemotherapy	tumor
Lumpectomy	removal
Lumpectomy	prosthesis
Lumpectomy	quadrantectomy
Mastectomy	removal
Mastectomy	unilateral
Mastectomy	prosthesis
Day Hospital	SPU visit
Skin graft / Lymph nodes operations	removal
Skin graft / Lymph nodes operations	skin
Skin graft / Lymph nodes operations	prosthesis
Plastic	prosthesis
Plastic	reconstruction
Other exams and therapies	ultrasound

Other exams and therapies	exams
Other exams and therapies	exercises
Other exams and therapies	tumor
OUT	

In the resulting 81 careflows it is possible to define 9 sub-groups of event patterns with the same meaning [67].

- Cluster 1: Reconstruction/plastic surgery: histories related to cases of one or more occurrences of plastic surgery for breast reconstruction.
- Cluster 2: Surgery + Therapy: histories related to cases one of the breast cancer guidelines of surgery and therapy.
- Cluster 3: Surgery + Therapy + Plastic Surgery: histories of the cluster 2 with the addition of plastic surgery.
- Cluster 4: double surgery: histories of double surgery cases divided in two sub-clusters. Cluster 4a: if the second surgery occurred before two months after the first one. Cluster 4b reports cases related to a second surgery that was performed later in time. The majority of the first cases are second intervention decided after the histopathological exams of the breast samples collected by a first intervention. Instead, in the second case are typically included second cancer episodes.
- Cluster 5: Surgery + Plastic Surgery: history of patients managed by ICSM only for their surgical intervention. Their therapy is decided in another hospital.
- Cluster 6: Neoadjuvant therapy: histories of patients for which a SPU visit is performed before the first surgery. Neoadjuvant therapy consists in the administration of therapeutic agents, such as chemotherapy or hormonal therapy, before the main treatment.
- Cluster 7: surgery + rehabilitation: histories of patients subjected to surgery followed by rehabilitation at the hospital.
- Cluster 8: surgery: histories of patients subjected to surgery at ICSM.
- Cluster 9: surgery + exams: histories of patients underwent surgery and further exams to investigate the clinical outcome of surgery.

Figure 61 shows the original CFM results, with an example of the careflows regrouping operation of Clusters 2 (Surgery and therapy), 4 (double surgery) and 8 (surgery).

The subdivision of the 9 clusters was obtained using only administrative data, but it was also performed the so-called “temporal phenotypes” from the clinical point of view. In fact, the relation between the 9 clusters and the outcome of the patients, through Kaplan-Mayer analysis, had an impact in the description of the evolution of the disease.

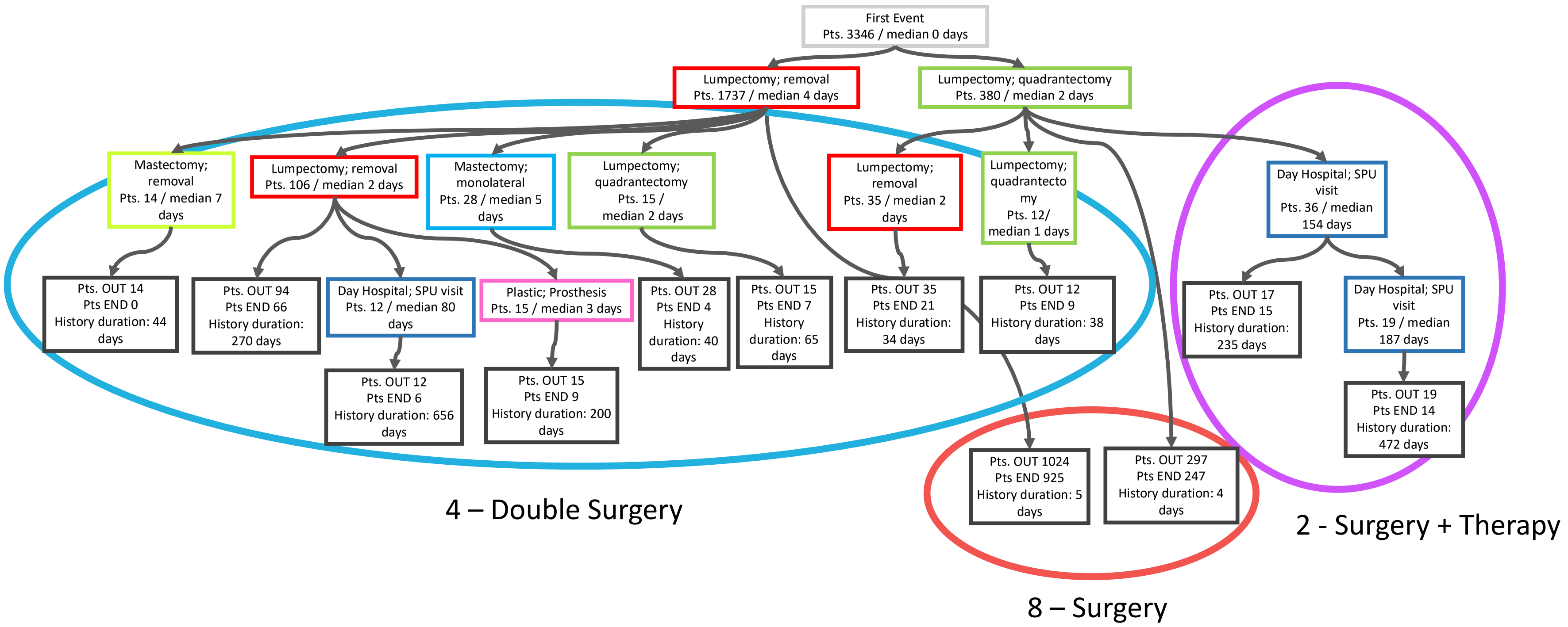


Figure 61 Visualization of the CFM results in the particular case of Clusters 2, 4 and 8 [67].

Chapter 5

Conclusions

In this thesis, many tools have been used to accomplish the creation of an IT infrastructure and to support the clinical research in the oncology field in Papa Giovanni XXIII Hospital.

The organization of the hospital in terms of data managing was already well developed: many types of collected data and applications were available to help daily administration and clinical practice. That was a very good starting point, and dialogue with the Hospital Information System (HIS) technicians permitted to build the different views from many hospital data sources.

Thanks to this data availability, we fully took advantage of the i2b2 platform structure and features. It demonstrated to be capable of integrate information from different electronic databases, already stored in the HIS, and to automatically and retrospectively retrieve them via query tool interface.

During the HER2-BC-FROM project and related studies, i2b2 demonstrated the ability to use its data archive for the reconstruction of relevant cases in breast cancer. The i2b2 platform supported data extraction and phenotyping approach and can be considered as a relevant tool for multiple outcomes research studies.

In this work, also the extraction of information in clinical texts has been addressed. Using an ontology driven Natural Language Process (NLP) approach, many pathology reports regarding breast cancer has been analyzed.

The proposed NLP pipeline was originally developed for molecular cardiology reports but was successfully adapted to extract relevant entities and attributes of oncologic domain from pathology reports. The extracted information has been added to the i2b2 data warehouse, expanding the pool of data available for clinical research on i2b2 query tool.

One of the main application of this work is to support and simplify the activities needed to design a clinical study: very often clinicians need to individually consult different data sources only to verify if they have enough target patients for the study itself. The i2b2 data warehouse unifies all the different data sources in a unique queryable structure.

The manual search can still be used for validation purposes, but the initial response of the query tool is able to avoid time consuming tasks in case of clinical studies that require more patients than those currently hospitalized.

Moreover, this work has provided a novel NLP pipeline to support the extraction of clinical information available in non-structured clinical reports, expanding the pool of possible searches and enlarging the number of concepts represented in the data warehouse.

The attention given to care event patterns in oncology field and the study of electronic phenotyping of patients brought the idea of an i2b2 plugin that could expand the possibility of mining useful patient careflows.

The CareFlow Mining (CFM) algorithm was considered for this aim. The application of CFM, even though applied in another scenario respect HPG23, resulted in careflows that are well matched with existing clinical guidelines and significant from a clinical point of view.

In fact, the algorithm produces real-world evidences, drawing temporal phenotypes across the initial population. Also, the careflows visualization is performed by a graphical output that improve results explanation and patient cohorts' identification.

5.1. Limitations

The work presented in the thesis has several limitations.

First of all, the only textual reports imported in the i2b2 data warehouse are the pathology reports.

This has important implications:

First, it means that the data coming from the Outpatients Services only contains ICD9 procedures, and no additional information related to the outcomes of the procedures and to all other notes is available.

Second, the unique source of information about diagnosis in i2b2 is the ICD9 diagnosis from Hospitalized patients, without all the diagnostic elements reported in imaging procedures reports and in the following visits. The NLP analysis of the pathology reports is a first step to cover this limitation.

Regarding the ETL work used to populate the i2b2 data warehouse, some phases are particularly difficult, therefore they require more attention and manual work by the developers. For example, the translation of the internal laboratory coding system to the LOINC system and of the Outpatients Services SISS procedures into ICD9 procedures were particularly hard to deal with.

Finally, the methods used in the PhD project had some limitations.

In i2b2, the database structure and the query tool have a “patient centric” organization. Sometimes, this is in conflict with the very nature of the data collected in clinical practice. For example, concept and attributes extracted via NLP from pathology reports reflect the “sample centric” focus of the Pathologists, leading to more complex ETL tasks to accomplish data integration.

The NLP method itself, due to its non supervised characteristic, must undergo to an initial step of ontology creation. This is a potential limit, since, even this ontology allowed reaching a tailored extraction of breast cancer data, it does not rely on standard concepts and it is hardly reusable in other contexts and more difficult to maintain.

5.2. Future Works and Perspectives

More efforts are required to complete the implementation and validation of described methods in the HPG23.

The original NLP pipeline has been incorporated in a general NLP procedure (3.2.3) that relies only on the OWL ontology to transfer data from clinical reports directly in i2b2 as observation.

After this general NLP procedure completion, further validations will be necessary to test its reliability, not only in breast cancer, but also for expanding its usage to other cancer types analyzed in pathology reports, as well as other kinds of clinical texts.

The interface part of the i2b2 plugin empowering CFM algorithm to be used directly from i2b2 query tool must be completed.

The i2b2 plugin solution will allow to tackle the crucial CFM preprocessing step, dedicated to the creation of events log. Also, it relies on the data contained and structured in the i2b2 data warehouse, allowing the clinical researchers to construct meaningful evens from i2b2 concepts and taxonomies in the query tool.

Also, an expansion of the CFM itself can be carried out.

A constraint on the temporal durations of events and transitions could be included in the CFM search strategy. For example, time spans differences from two consecutive events could lead to a possible different interpretation of the careflow (Figure 62).

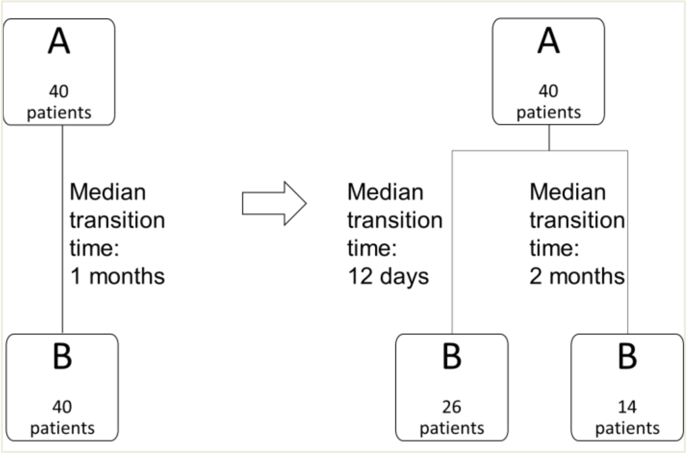

A
40 patients
Median transition time: 1 months
B
40 patients
A
40 patients
Median transition time: 12 days
Median transition time: 2 months
B
26 patients
B
14 patients

Figure 62 Example of careflow split based on transition duration between two events. This would be another user choice offered by the i2b2 plugin.

As medium-term future developments, it would possible to modify data import in i2b2 from the various sources. The LOINC mapping of internal laboratory analysis codes can already be expanded. In a fast and automatic way, the addition of just one row in the mapping table can transform all the observations relative to a laboratory result of interest in LOINC observations.

Also, the detailed information of therapy schedules application available in the drug administration view could be included in the i2b2 data warehouse. Once available, detailed observations of schedule administrations can be used to derive more abstracted observations, describing macro trend of therapy of a patient.

References

[1] ASST-PG23, (n.d.). http://www.asst-pg23.it/ (accessed October 3, 2019).

[2] Biomeris Official Site – Let your data play, (n.d.). https://www.biomeris.it/ (accessed October 3, 2019).

[3] About Breast Cancer | Breast Cancer Overview and Basics, (n.d.). https://www.cancer.org/cancer/breast-cancer/about.html (accessed October 3, 2019).

[4] AJCC - American Joint Committee on Cancer, (n.d.). https://cancerstaging.org/Pages/default.aspx (accessed October 6, 2019).

[5] R. Bellazzi, Big Data and Biomedical Informatics: A Challenging Opportunity, *Yearb. Med. Inform.* **23** (2014) 08–13. doi:10.15265/IY-2014-0024.

[6] N. Peek, J.H. Holmes, and J. Sun, Technical Challenges for Big Data in Biomedicine and Health: Data Sources, Infrastructure, and Analytics, *Yearb. Med. Inform.* **23** (2014) 42–47. doi:10.15265/IY-2014-0018.

[7] W. Raghupathi, and V. Raghupathi, Big data analytics in healthcare: promise and potential., *Heal. Inf. Sci. Syst.* **2** (2014) 3. doi:10.1186/2047-2501-2-3.

[8] i2b2: Informatics for Integrating Biology & the Bedside, (n.d.). https://www.i2b2.org/resrcs/hive.html# (accessed September 23, 2019).

[9] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, and I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), *J. Am. Med. Informatics Assoc.* **17** (2010) 124–130. doi:10.1136/jamia.2009.000893.

[10] A.R. Post, T. Kurc, S. Cholleti, J. Gao, X. Lin, W. Bornstein, D. Cantrell, D. Levine, S. Hohmann, and J.H. Saltz, The Analytic Information Warehouse (AIW): A platform for analytics using electronic health record data, *J. Biomed. Inform.* **46** (2013) 410–424. doi:10.1016/j.jbi.2013.01.005.

[11] D. Segagni, V. Tibollo, A. Dagliati, A. Zambelli, S.G. Priori, and R. Bellazzi, An ICT infrastructure to integrate clinical and molecular data in oncology research, *BMC Bioinformatics*. (2012). doi:10.1186/1471-2105-13-S4-S5.

[12] K.B. Wagholikar, J.C. Mandel, J.G. Klann, N. Wattanasin, M. Mendis, C.G. Chute, K.D. Mandl, and S.N. Murphy, SMART-on-FHIR implemented over i2b2., *J. Am. Med. Inform. Assoc.* **24** (2017) 398–402. doi:10.1093/jamia/ocw079.

[13] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, and J.F. Hurdle,

Extracting information from textual documents in the electronic health record: a review of recent research., *Yearb. Med. Inform.* (2008) 128–44. http://www.ncbi.nlm.nih.gov/pubmed/18660887 (accessed October 6, 2019).

[14] S. Velupillai, D. Mowery, B.R. South, M. Kvist, and H. Dalianis, Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis, *Yearb. Med. Inform.* **24** (2015) 183–193. doi:10.15265/IY-2015-009.

[15] C. Friedman, A broad-coverage natural language processing system., *Proceedings. AMIA Symp.* (2000) 270–4. http://www.ncbi.nlm.nih.gov/pubmed/11079887 (accessed October 6, 2019).

[16] Unified Medical Language System (UMLS), (n.d.). https://www.nlm.nih.gov/research/umls/index.html.

[17] G.K. Savova, J.J. Masanz, P. V Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, and C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications., *J. Am. Med. Inform. Assoc.* **17** (2010) 507–13. doi:10.1136/jamia.2009.001560.

[18] E. Chiaramello, F. Pinciroli, A. Bonalumi, A. Caroli, and G. Tognola, Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes, *J. Biomed. Inform.* **63** (2016) 22–32. doi:10.1016/J.JBI.2016.07.017.

[19] A. Alicante, A. Corazza, F. Isgrò, and S. Silvestri, Unsupervised entity and relation extraction from clinical records in Italian, *Comput. Biol. Med.* **72** (2016) 263–275. doi:10.1016/j.compbiomed.2016.01.014.

[20] A. Mykowiecka, M. Marciniak, and A. Kupść, Rule-based information extraction from patients' clinical data, *J. Biomed. Inform.* **42** (2009) 923–936. doi:10.1016/j.jbi.2009.07.007.

[21] M. Toepfer, H. Corovic, G. Fette, P. Klügl, S. Störk, and F. Puppe, Fine-grained information extraction from German transthoracic echocardiography reports, *BMC Med. Inform. Decis. Mak.* **15** (2015) 91. doi:10.1186/s12911-015-0215-x.

[22] N. Viani, C. Larizza, V. Tibollo, C. Napolitano, S.G. Priori, R. Bellazzi, and L. Sacchi, Information extraction from Italian medical reports: An ontology-driven approach, *Int. J. Med. Inform.* **111** (2018) 140–148. doi:10.1016/j.ijmedinf.2017.12.013.

[23] protégé, (n.d.). https://protege.stanford.edu/ (accessed September 19, 2019).

[24] Á. Rebuge, and D.R. Ferreira, Business process analysis in healthcare environments: A methodology based on process mining, *Inf. Syst.* **37** (2012) 99–116. doi:10.1016/J.IS.2011.01.003.

[25] L. Bouarfa, and J. Dankelman, Workflow mining and outlier detection from clinical activity logs, *J. Biomed. Inform.* **45** (2012) 1185–1190. doi:10.1016/J.JBI.2012.08.003.

[26] S. Panzarasa, S. Maddè, S. Quaglini, C. Pistarini, and M. Stefanelli, Evidence-based careflow management systems: the case of post-stroke rehabilitation, *J. Biomed. Inform.* **35** (2002) 123–139.

doi:10.1016/S1532-0464(02)00505-1.

[27] S. Concaro, L. Sacchi, C. Cerra, P. Fratino, and R. Bellazzi, Mining Health Care Administrative Data with Temporal Association Rules on Hybrid Events, *Methods Inf. Med.* **50** (2011) 166–179. doi:10.3414/ME10-01-0036.

[28] Z. Huang, X. Lu, H. Duan, and W. Fan, Summarizing clinical pathways from event logs, *J. Biomed. Inform.* **46** (2013) 111–127. doi:10.1016/J.JBI.2012.10.001.

[29] Z. Huang, W. Dong, L. Ji, C. Gan, X. Lu, and H. Duan, Discovery of clinical pathway patterns from event logs using probabilistic topic models, *J. Biomed. Inform.* **47** (2014) 39–57. doi:10.1016/J.JBI.2013.09.003.

[30] A. Dagliati, L. Sacchi, A. Zambelli, V. Tibollo, L. Pavesi, J.H. Holmes, and R. Bellazzi, Temporal electronic phenotyping by mining careflows of breast cancer patients, *J. Biomed. Inform.* **66** (2017) 136–147. doi:10.1016/j.jbi.2016.12.012.

[31] A. Dagliati, V. Tibollo, G. Cogni, L. Chiovato, R. Bellazzi, and L. Sacchi, Careflow Mining Techniques to Explore Type 2 Diabetes Evolution, *J. Diabetes Sci. Technol.* **12** (2018) 251–259. doi:10.1177/1932296818761751.

[32] D.J. Albers, N. Elhadad, E. Tabak, A. Perotte, and G. Hripcsak, Dynamical Phenotyping: Using Temporal Analysis of Clinically Collected Physiologic Data to Stratify Populations, *PLoS One*. **9** (2014) e96443. doi:10.1371/journal.pone.0096443.

[33] S.N. Murphy, P. Avillach, R. Bellazzi, L. Phillips, M. Gabetta, A. Eran, M.T. McDuffie, and I.S. Kohane, Combining clinical and genomics queries using i2b2 - Three methods, *PLoS One*. (2017). doi:10.1371/journal.pone.0172187.

[34] C. Friedman, G. Hripcsak, S.B. Johnson, J.J. Cimino, and P.D. Clayton, A Generalized Relational Schema for an Integrated Clinical Patient Database, *Proc. Annu. Symp. Comput. Appl. Med. Care*. (1990) 335. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245527/ (accessed September 23, 2019).

[35] T. Adamusiak, T. Burdett, N. Kurbatova, K. Joeri van der Velde, N. Abeygunawardena, D. Antonakaki, M. Kapushesky, H. Parkinson, and M.A. Swertz, OntoCAT -- simple ontology search and integration in Java, R and REST/JavaScript, *BMC Bioinformatics*. **12** (2011) 218. doi:10.1186/1471-2105-12-218.

[36] Enabling Open Innovation &amp; Collaboration | The Eclipse Foundation, (n.d.). https://www.eclipse.org/ (accessed September 23, 2019).

[37] Mirth Connect Introduction and Tutorial [Patesco], (n.d.). http://wiki.patesco.ca/doku.php?id=hl7:mirth:tutorial (accessed September 15, 2019).

[38] G. Bortis, Experiences with mirth: An open source health care integration engine, in: Proc. - Int. Conf. Softw. Eng., 2008. doi:10.1145/1368088.1368179.

[39] W. Haque, A. Reed, and A. McCann, A framework for secure integration of distributed point-of-care testing results into Electronic

Medical Records, in: Int. Conf. Inf. Soc. i-Society 2013, 2013.

[40] NextGen® Connect Integration Engine, (n.d.). https://www.nextgen.com/products-and-services/integration-engine (accessed September 17, 2019).

[41] Clinical Data Exchange using HL7 and Mirth Connect Lecture 12 - Using JavaScript with Mirth Connect – III - Advanced Message Routing - XSLT transforms. - ppt download, (n.d.). https://slideplayer.com/slide/11646148/ (accessed September 14, 2019).

[42] M. Horridge, H. Knublauch, A. Rector, R. Stevens, C. Wroe, S. Jupp, G. Moulton, N. Drummond, and S. Brandt, Protege 4 Tutorial Version 1.3, *Matrix*. (2011).

[43] D. Ferrucci, and A. Lally, UIMA: An architectural approach to unstructured information processing in the corporate research environment, *Nat. Lang. Eng.* (2004). doi:10.1017/S1351324904003523.

[44] E. Pianta, C. Girardi, and R. Zanoli, The TextPro tool suite, in: Proc. 6th Int. Conf. Lang. Resour. Eval. Lr. 2008, 2008.

[45] Federfarma official site, (n.d.). https://www.federfarma.it/ (accessed September 24, 2019).

[46] G.K. Savova, J.J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, and C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications, *J. Am. Med. Informatics Assoc.* (2010). doi:10.1136/jamia.2009.001560.

[47] H. Harkema, J.N. Dowling, T. Thornblade, and W.W. Chapman, ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports, *J. Biomed. Inform.* (2009). doi:10.1016/j.jbi.2009.05.002.

[48] W. van der Aalst, Process mining: discovering and improving Spaghetti and Lasagna processes, in: 2012. doi:10.1109/cidm.2011.6129461.

[49] M. Leemans, and W.M.P. van der Aalst, Discovery of frequent episodes in event logs, in: Lect. Notes Bus. Inf. Process., 2015. doi:10.1007/978-3-319-27243-6_1.

[50] R. Agrawal, and R. Srikant, Mining sequential patterns, in: Proc. - Int. Conf. Data Eng., 1995.

[51] Transcodifica Codici prestazioni | Open Data Regione Lombardia, (n.d.). https://www.dati.lombardia.it/Sanit-/Transcodifica-Codici-prestazioni/7ugz-vcug (accessed September 27, 2019).

[52] Home – LOINC, (n.d.). https://loinc.org/ (accessed September 27, 2019).

[53] N. Viani, L. Chiudinelli, C. Tasca, A. Zambelli, M. Bucalo, A. Ghirardi, N. Barbarini, E. Sfreddo, L. Sacchi, C. Tondini, and R. Bellazzi, Automatic Processing of Anatomic Pathology Reports in the Italian Language to Enhance the Reuse of Clinical Data., *Stud. Health Technol. Inform.* **247** (2018) 715–719. http://www.ncbi.nlm.nih.gov/pubmed/29678054 (accessed November 23, 2018).

[54] Anatomic Pathology Lexicon - Summary | NCBO BioPortal, (n.d.). https://bioportal.bioontology.org/ontologies/PATHLEX (accessed

September 27, 2019).

[55] C. Daniel, D. Booker, B. Beckwith, V. Della Mea, M. García-Rojo, L. Havener, M. Kennedy, J. Klossa, A. Laurinavicius, F. Macary, V. Punys, W. Scharber, and T. Schrader, Standards and specifications in pathology: image management, report management and terminology., *Stud. Health Technol. Inform.* **179** (2012) 105–22. http://www.ncbi.nlm.nih.gov/pubmed/22925792 (accessed October 18, 2018).

[56] L. Chiudinelli, M. Gabetta, G. Centorrino, N. Viani, C. Tasca, A. Zambelli, M. Bucalo, A. Ghirardi, N. Barbarini, E. Sfreddo, C. Tondini, R. Bellazzi, and L. Sacchi, Ontology-Driven Real World Evidence Extraction from Clinical Narratives., *Stud. Health Technol. Inform.* **264** (2019) 1441–1442. doi:10.3233/SHTI190474.

[57] LINEE GUIDA Neoplasie della mammella 2016 | AIOM, (n.d.). https://www.aiom.it/neoplasie-della-mammella-8/ (accessed September 26, 2019).

[58] Fondazione FROM, (n.d.). http://www.fondazionefrom.it/ (accessed September 30, 2019).

[59] XX Congresso Nazionale AIOM | AIOM, (n.d.). https://www.aiom.it/eventi-aiom/xx-congresso-nazionale-aiom/ (accessed September 30, 2019).

[60] XXI Congresso Nazionale AIOM | AIOM, (n.d.). https://www.aiom.it/eventi-aiom/xxi-congresso-nazionale-aiom/ (accessed September 30, 2019).

[61] A. Zambelli, A. Ghirardi, A. Masciulli, E. Sfreddo, R. Porcino, M. Bucalo, N. Barbarini, L. Chiudinelli, A. Chirco, A. Labianca, T. Barbui, and C. Tondini, Ten-Years Electronic Phenotyping Archive And Automated Reconstruction Of Her2+ Breast Cancer Patients Careflow, Through The Exportable, Open-Source I2b2 Data Ware-Housing Platform, *Tumori J.* **AIOM 2018** (2018). https://www.aiom.it/wp-content/uploads/2018/05/2018_TMJA_AbstractAIOMXX.pdf (accessed September 30, 2019).

[62] I. Jatoi, S.G. Hilsenbeck, G.M. Clark, and C.K. Osborne, Significance of Axillary Lymph Node Metastasis in Primary Breast Cancer, *J. Clin. Oncol.* **17** (1999) 2334–2334. doi:10.1200/JCO.1999.17.8.2334.

[63] J. Richard, C. Sainsbury, G. Needham, J. Farndon, A. Malcolm, and A. Harris, Epidermal-Growth-Factor Receptor Status As Predictor Of Early Recurrence Of And Death From Breast Cancer, *Lancet.* **329** (1987) 1398–1402. doi:10.1016/S0140-6736(87)90593-9.

[64] D.M. Blei, A.Y. Ng, and M.I. Jordan, Latent Dirichlet allocation, *J. Mach. Learn. Res.* (2003).

[65] B. Grün, and K. Hornik, R: Package 'topicmodels,' *CRAN.* (2017).

[66] ICS Maugeri, (n.d.). https://www.icsmaugeri.it/ (accessed October 1, 2019).

[67] L. Chiudinelli, A. Dagliati, V. Tibollo, S. Albasini, N. Geifman, N. Peek, J. Holmes, F. Corsi, R. Bellazzi, and L. Sacchi, Mining post-surgical care processes in breast cancer patients, *Artif. Intell. Med.* **SUBMITTED** (2019).