ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Biomedical Engineering and Medical-Surgical Sciences (30th Cycle)

I2ECR: Integrated and Intelligent Environment for Clinical Research

By

Gian Maria Zaccaria

******

Supervisor(s):
Prof.ssa G. Balestra, POLITO
Prof. M. Boccadoro, UNITO

Doctoral Examination Committee:
Prof. Agostino Accardo, Università di Trieste
Dott.ssa Federica Cavallo, PhD, Università di Torino
Dott.ssa Ilaria Del Giudice, PhD, Università La Sapienza
Dott.ssa Francesca Maria Gay, PhD, Città della Salute e della Scienza di Torino
Prof. Stefano Luminari, Università di Modena e Reggio Emilia

Politecnico di Torino
2018

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Gian Maria Zaccaria.

2018

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

“***So che la sanità***
Può curare i suoi
Grandi numeri ma
Non me…”

Questi ringraziamenti, rigorosamente in italiano, descrivono il grosso sacrificio fatto in questi anni…

A Giovanni, “who reached for the secret too soon, crying for the moon”, un amico che merita di essere ricordato per la sua storia, e per la nostra infinita amicizia…

A Rossella, mia moglie, che non ha mai smesso di capire la mia dedizione a dare il meglio e a non tirarmi indietro.

Ai miei famigliari, che sanno quanto sia importante per me sacrificarsi per un buon risultato.

e agli amici tutti, che hanno saputo cogliere la mia passione nelle cose, senza giudicare le mie tantissime assenze…

Acknowledgment

I would like to acknowledge Prof. Boccadoro for sponsoring this PhD experience.

Thanks to Prof.ssa Balestra, for giving me the right tools for seeking solutions for solving problems.

Thanks to Dr. Simone Ferrero, the perfect physician to work with.

Thanks to Dr. Marco Ladetto and FIL – Fondazione Italiana Linfomi for authorizing me to work on FIL-MCL0208 clinical trial.

Thanks to Dr. Marco Ghislieri, Dr. Cristina Castagneri, Dr. Rebecca Sandrone and Dr. Samanta Rosati, my POLITO colleagues.

A great acknowledgment to Molecular Biology Laboratory of Hematology Unit of Turin, in particular, to Dr. Elisa Genuardi, Dr. Daniela Barbero and both Dr. Mariella Lo Schirico and Dr. Riccardo Moia.

Thanks to Hematology Units of University of Torino, Città della Salute e della Scienza and "Ospedale Civile Santi Antonio e Biagio e Cesare Arrigo" of Alessandria, to the Fo.Ne.Sa (Fondazione Neoplasie del Sangue) organization, to all the clinicians, data-mangers, nurses, clerks for helping me in any situation.

Thanks to Dr. Andrea Evangelista of CPO of Città della Salute e della Scienza di Torino, for sharing his precious expertis.

Abstract

Clinical trials are designed to produce new knowledge about a certain disease, drug or treatment. During these studies, a huge amount of data is collected about participants, therapies, clinical procedures, outcomes, adverse events and so on.

A multicenter, randomized, phase III clinical trial in Hematology enrolls up to hundreds of subjects and evaluates post-treatment outcomes on stratified sub-groups of subjects for a period of many years. Therefore, data collection in clinical trials is becoming complex, with huge amount of clinical and biological variables. Outside the medical field, data warehouses (DWs) are widely employed. A Data Ware-house is a "collection of integrated, subject-oriented databases designed to support the decision-making process". To verify whether DWs might be useful for data quality and association analysis, a team of biomedical engineers, clinicians, biologists and statisticians developed the "I2ECR" project.

I2ECR is an Integrated and Intelligent Environment for Clinical Research where clinical and omics data stand together for clinical use (reporting) and for generation of new clinical knowledge. I2ECR has been built from the "MCL0208" phase III, prospective, clinical trial, sponsored by the Fondazione Italiana Linfomi (FIL); this is actually a translational study, accounting for many clinical data, along with several clinical prognostic indexes (e.g. MIPI - Mantle International Prognostic Index), pathological information, treatment and outcome data, biological assessments of disease (MRD - Minimal Residue Disease), as well as many biological, ancillary studies, such as Mutational Analysis, Gene Expression Profiling (GEP) and Pharmacogenomics. In this trial forty-eight Italian medical centers were actively involved, for a total of 300 enrolled subjects. Therefore, I2ECR main objectives are:

- to propose an integration project on clinical and molecular data quality concepts. The application of a clear row-data analysis as well as clinical trial monitoring strategies to implement a digital platform where clinical, biological and

“omics” data are imported from different sources and well-integrated in a data-ware-house

- to be a dynamic repository of data congruency quality rules. I2ECR allows to monitor, in a semi-automatic manner, the quality of data, in relation to the clinical data imported from eCRFs (electronic Case Report Forms) and from biologic and mutational datasets internally edited by local laboratories. Therefore, I2ECR will be able to detect missing data and mistakes derived from non-conventional data-entry activities by centers.

- to provide to clinical stake-holders a platform from where they can easily design statistical and data mining analysis. The term Data Mining (DM) identifies a set of tools to searching for hidden patterns of interest in large and multivariate datasets. The applications of DM techniques in the medical field range from outcome prediction and patient classification to genomic medicine and molecular biology. I2ECR allows to clinical stake-holders to propose innovative methods of supervised and unsupervised feature extraction, data classification and statistical analysis on heterogeneous datasets associated to the MCL0208 clinical trial.

Although MCL0208 study is the first example of data-population of I2ECR, the environment will be able to import data from clinical studies designed for other onco-hematologic diseases, too.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Personalized, Precise and Translational Medicine

Personalized and precise medicine aim at "just treatment for the right patient at the right time" and “with the optimally dose anticancer agent” (Mendelsohn, 2015). The concept of personalized medicine was first proposed in the 1990s (Shi-kai, 2015) when scientists recognized the close association between the individual genetic features and a phenotype of clinical disease. Personalized medicine provides individual decision of diagnosis and evaluation of the treatment.

In last decade, president Obama announced a research initiative[1] that aimed to accelerate progress toward a new era of precision medicine (Adams and Petersen 2016). Precision medicine initiative requires a more thorough review of disease classification process not focused on a single state of illness, but it includes all relevant molecular information for diseases when confirmed and validated. “This will transform a more holistic (figure 1) view of the disease and may also help to identify latent biological commonality between different disease processes”(Mirnezami, 2012). “Precision medicine requires a strong interdisciplinary collaboration between several stakeholders covering a large continuum of expertise ranging from medical, clinical, biological, technical, and biotechnological know-hows” (Servant, 2014). A first differentiation between

[1] https://obamawhitehouse.archives.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative.

personalized and precision medicine definition is related to level of engineering associated to: personalization of medicine leverages on a social side; physicians shall know very well their patients and may personalize therapy basing on their habits and on exposure agents within environment where they live. However, a precision medicine approach moves through a high level of technological innovation. Again, “the precision medicine considers a model of healthcare that is predictive, preventive, personalized and participatory” (Bellazzi, 2011). Precision medicine therefore extends the concept of personalized medicine. However, personalized and precision medicine implicate translational medicine definition. Translational medicine is the field that integrates genomics and clinical medicine to bridge the gap between basic medical research and clinical care (Bellazzi, 2011).

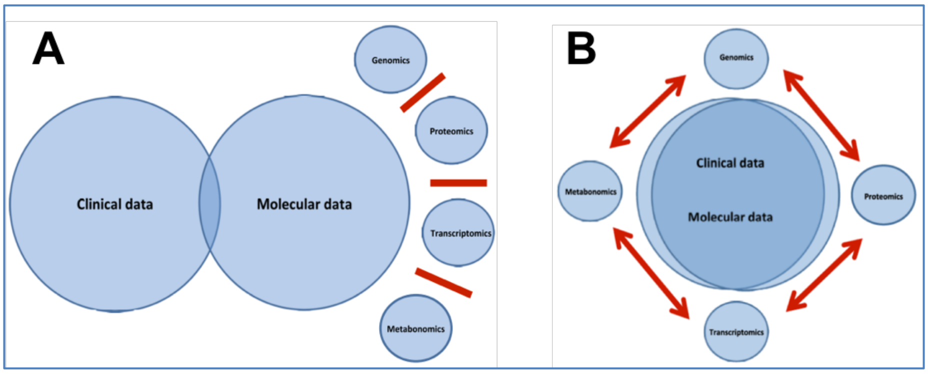


Figure 1: Figure A illustrates current disease classification, which provides insufficient integration of clinical data with disease-relevant biomolecular data (Genomics, Proteomics, Transcriptomics and Metabolomics). Figure B depicts that Precision and Personalized Medicine are bridging to a common domain between molecular and clinical data (Mirnezami, 2012).

1.1.1 “Clinics” vs. “Omics” data: a technological challenge

The next-generation sequencing (NGS) technology firstly came out around ten years ago. The first human genome was sequenced in more than 10 years (2003) at the cost of around 3 billion dollars. Nowadays, it is possible to sequence a genome for a few thousand dollars in a short time (Servant et al. 2014). Since NGS is becoming more and more relevant in the study of cancer diseases, innovative treatment strategies and risk stratification are rising thanks to its affordable costs. Because the whole genome sequencing by NGS is important to the study of complex diseases such as cancer, a decreased cost gives rise to new opportunities for personalized treatment and risk stratification (Andreu-Perez 2015). Target discovery plays a critical role in new drug development. “Genomic (omics) studies indicate that humans have 30.000−40.000 genes and many more proteins and at least 90 percent of the target proteins have not yet been discovered. To discover and validate new drug targets is of great significance for the elucidation of the mechanisms of disease pathology and the effects of drugs" (Shi-kai 2015). Generally, translational medicine projects follow this strategy:

1. A disease model is constructed on biological and clinical samples.
2. Omics analyses are performed, including ancillary studies of genomics.
3. Bioinformatics algorithms are used to process the acquired omics data, and innovative disease-related bio-markers are proposed.
4. Bioengineering methods, however, may be used to retrieve information from related databases, and potential target candidates or disease-related biomolecules are discovered using tools of data mining and network biology. Functional analysis is then performed on these disease-related substances and the functional disease-related biomolecules are proposed as potential targets.
5. Targets are verified by pharmacological studies at the molecular and cellular levels and subsequently in animal models.

"Incorporating omics data into conventional clinical data-sets means that new training paradigms for tomorrow's doctors must be developed" (Mirnezami, 2012). To combine different datasets, three important steps has to be considered: (i) the technical level to develop a powerful computational architecture (software/hardware), (ii) the organizational and management levels to define the procedures to collect data with highest confidence, quality and traceability, and (iii) the scientific level to create sophisticated models to predict the evolution of the disease and risks to the patient. To do this, the development of a seamless information system allowing data integration, data traceability, and knowledge sharing across the different stake-holders is mandatory with the support of a robust architecture, which must warrant the reproducibility of the results.

1.2 Strategies for clinical research

1.2.1 From clinical epidemiology..

"Clinical epidemiology is the science of making predictions about individual patients by counting clinical events in groups of similar patients and using strong scientific methods to ensure that the predictions are accurate. The purpose of clinical epidemiology is to develop and apply methods of clinical observation that will lead to valid conclusions by avoiding being misled by systematic error and the play of chance. Epidemiology is the “study of disease occurrence in human populations” (Fletcher, 2003). Epidemiological evidence pyramids in figure 2 depicts scientific evidence of methodologies in use in clinical epidemiology. Randomized clinical trial (RCT) and Cohort Studies are in the middle of scale.

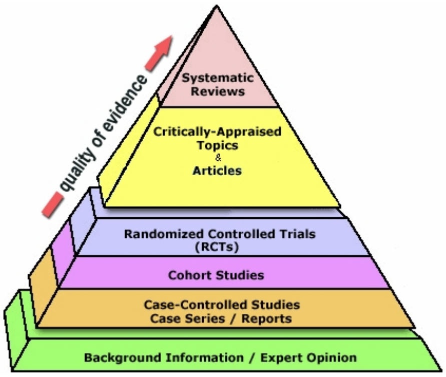

quality of evidence
Systematic Reviews
Critically-Appraised Topics & Articles
Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies Case Series / Reports
Background Information / Expert Opinion

Figure 2: Epidemiological evidence pyramid[2].

“Randomized trials are studies in which a direct comparison is made between two or more treatment groups, one of which serves as a control for the other. Study subjects are randomly allocated into the differing treatment groups, and all groups are followed over time to observe the effect of the different treatments” (Alexander, n.d.). A cohort study is an epidemiological study in which a group of people with a communal characteristic is followed over time to find the percentage of patients that reach a certain health outcome. Health outcomes are the most important events in clinical medicine. They are, discomfort, disability, disease, dissatisfaction and death (Fletcher, 2003). A cohort is defined as a group of persons who share a characteristic, (e.g. smokers), exposed (or less exposed) to determine if the outcome is associated with exposure. The cohort studies are divided in: prospective and retrospective cohorts. Prospective studies follow a cohort for a future health outcome, while retrospective studies trace the cohort back in time for exposure information after the outcome has occurred.

1.2.2 ..to IT infrastructure data-driven projects

In last 20 years, an increasing number of institutions are joining in national and international consortium to integrate clinical and genetic data. Academic, no-profit

[2] https://s3.amazonaws.com/libapps/customers/2440/images/Pyramid_Evidence.JPG

organizations and private companies are boosting on national and international programs of data sharing within involved stakeholders are encouraged to discuss and develop innovative methodologies to extrapolate new health outcomes. Nevertheless, researchers and clinical investigators have understood that translational research shall pass through a data-integration effort. Thus, scientific symposia began to involve actors with no-clinical and biologic skills: among these, legal experts (i) about the anonymized treatment of patients' data, statisticians and mathematicians (ii) to discover and validate models for clinical prediction and engineers (iii) to design IT (Information Technology) infrastructures for collecting data.

Building up the framework for personalized medicine is a challenge for any health care center. Major cancer centers are facing the additional challenge of scale; they need to rapidly incorporate new technologies and offer state-of-the-art cancer care to a massive number of patients in the case of large institutions such as The University of Texas MD Anderson Cancer Center, where 30.000 new patients are seen each year (Meric-Bernstam, 2013). In 2004 Isaac Kohane, Professor of Pediatrics at Harvard Medical School at Children's Hospital Boston founded the Informatics for Integrating Biology and Bedside (i2b2[3]) Center. This Center was funded under a cooperative agreement with the National Institutes of Health (NIH) and initially involved the Harvard-affiliated hospitals, MIT, Harvard School of Public Health, Harvard Medical School and the Harvard/MIT Division of Health Sciences and Technology. The i2b2 Center was thought for developing a scalable computational framework to address the bottleneck limiting the translation of genomic findings and hypotheses in model systems relevant to human health (Murphy, 2006). Focusing on oncology field, in 2008 the Oncotyrol consortium was founded (Siebert, 2015). Oncotyrol is an international and interdisciplinary alliance combining research and commercial competencies to accelerate the development, evaluation and translation of personalized health care strategies in cancer across Tyrol region (among Germany and Austria). "The mission was to establish a consortium to close the gap between basic research, clinical research, and population research, on one hand, and the commercial development of translational approaches and healthcare solutions on the other hand" (Siebert et al. 2015). Finding pan-European initiatives, EU-IMI[4] currently covers a big role in proposing new IT infrastructure programs. The Innovative Medicines Initiative

[3] www.i2b2.org

[4] www.imi.europa.eu

(IMI) is Europe's largest public-private initiative aiming to speed up the development of better and safer medicines for patients. IMI supports collaborative research programs and suites networks of industrial and academic experts in order to enforce pharmaceutical innovation in Europe. Among these projects, EHR4CR and HARMONY initiatives deserve to be cited. The EHR4CR project (i) has shown that it "is now possible to profoundly innovate biomedical research relying on newly. designed IT systems" (Zapletal, 2010). In 2016, IMI and the European Hematology Association (EHA[5]) launched the 5-years HARMONY project. HARMONY captures, integrates, analyzes and harmonizes anonymous patient data from high-quality multidisciplinary sources to unlock valuable knowledges on Hematology Malignancies: multiple myeloma, acute myeloid leukemia, acute lymphoblastic leukemia, chronic lymphocytic leukemia, non-Hodgkin's lymphoma, myelodysplastic syndromes and pediatric HMs. The expected outcome was a better prognosis and quicker life-saving decisions, important for patients suffering from these diseases. The project brings together key participants in the clinical, academic, patient, health technology assessment, regulatory, economical, ethical and pharmaceutical fields to:

1. Projecting a shared platform that empowers clinicians and policy stakeholders to improve decision-making.
2. Defining clinical endpoints and standard outcomes in HM in order to standardize them among the key stakeholders.
3. Providing means for analyzing complex data sets comprising different layers of information.
4. Identifying specific markers for early registration of innovative and effective therapies for HMs.

1.2.3 Technical and scientific approaches for medical data collection

IT infrastructure driven projects are very often the result of integration of single systems which generally fulfill tasks with high-level specialization (vertical software). Traditional health data centers store in Electronic Health Records (EHRs) huge amount of clinical data including diagnostics, laboratory tests, medications, biologic and mutational ones. Importance of EHRs for care delivery has been assumed by scientific boards (Downing, 2009; Servant, 2014). However,

[5] www.eha.org

the “perfect” system able to horizontally manage heterogeneous data in the same environment does not exist. In addition, integration issues dramatically increase if a center is “active” in clinical and molecular research. Currently, electronic Case Report Form (eCRF) platforms for clinical trials data collection are designed by local software-houses and are not integrated to EHRs systems used in single centers. Focusing on this heterogeneousness, integration issues may be divided in two great families, which group different data collection projects (figure 3):

1. Integration issues within hospital – ERP (Enterprise Resource Planning), CPOE (Computerized Physician Order Entry), Laboratory, bio-banking and molecular data systems.
2. Integration issues between hospitals – eCRFs vs. EHRs systems.

A detailed description of above-mentioned systems for health data collection is proposed from paragraph 1.2.1.1 to 1.2.2.3.

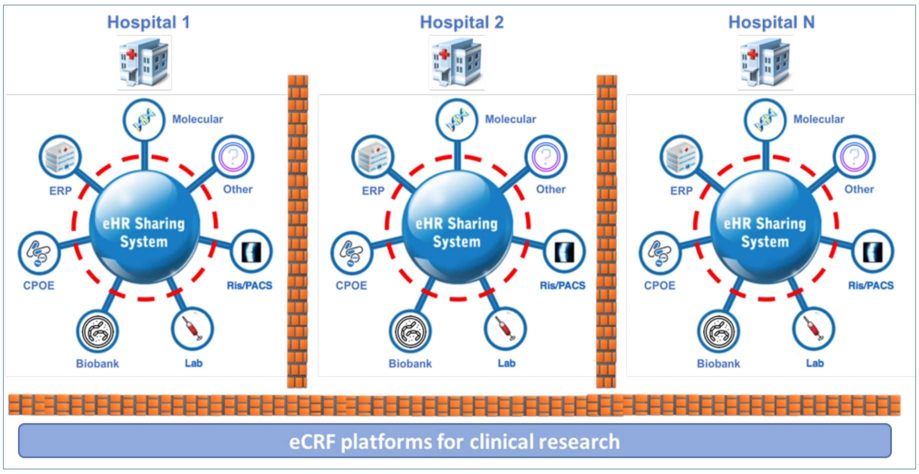


Figure 3: There are two types of SW Integration issues about medical hospitals: one local and one general. Within a hospital, VPN (Virtual Private Network), computerized systems generally suffer of low integration (local). Red dashed circle line indicates this criticism. Horizontal (EHRs) patients' management systems usually group ERP systems, Laboratory Analysis software and RIS/PACS infrastructures. CPOE, biobank and molecular systems are commonly stand-alone software. Clinical Research activities pass through a manual clinical data record on eCRFs platform. Internal privacy policies and low investments on technical solutions obstacle a communication between single centers and central sponsors (global).

1.2.3.1 CPOE Systems

Oncology departments must respect high-standard of security for patients as well as workers managing cytotoxic drugs. Thus, they often acquire CPOE (Computerized Patients Order Entry) systems, specialized in pharmacology electronic prescription (ePrescription), preparation and administration. Therefore, CPOE systems allow high-vertical activities to all actors who manage cytotoxic drugs: physicians, hospital pharmacists, pharmacy technicians, nurses, clinical trial research assistants and clerks. Although in optimistic vision, central EHRs and CPOE system must share information about patient demography, diagnosis, ambulatory reports, procedures and medication, uneasy integration of coexisting electronic and paper-based systems in the correlated phases are usual in most hospitals (Niazkhani, 2011). On the other hand, the transition to EHRs has expanded the reach of medical record–based information but has not markedly improved the quality of the data entered. Although examples of improved clinical practice driven by EHRs can be found, the quality and granularity of the data they record limit their use in research. The inherent variability of clinical data cross institutions is magnified by institution- to-institution differences in EHR systems (Joyner and Paneth 2015).

1.2.3.2 Bio-banking software

Kauffmann et al. defined a biobank as "an organized collection of human biological material and associated information stored for one or more research purposes" (F. Kauffmann 2008). Cancer biobanks represent a key resource for diagnosis and for further use in fundamental and translational cancer research. Generally, are divided in disease-oriented biobanks and population-based biobanks (Luo, 2014). Disease-oriented biobanks are more often based on the hospital and include cancer banks as well as blood collections and other samples from a variety of diseases along with normal controls. However, population-based biobanks are generally located outside the hospitals and sample donors are normal volunteers. By focusing on bio-repository of data through biobanks, associated information may include health data such as clinical information from EHRs, quality of life information acquired through surveys and medical history. Still, bio-samples are used by cancer researchers to study molecular changes between primary tumor and metastatic disease and drug resistance development. Unfortunately, biobanks application systems are often stand-alone. This is generally due to a low interest by hospital management committees to invest public funds.

1.2.3.3 eCRFs, computerized tools for supporting epidemiologic studies

Since last 20 years, sponsors that invest on epidemiologic research are dramatically stimulating centers to use eCRFs platforms for data entry. CRFs are designed to capture the required data at all multicenter trial sites. Public hospitals totally avoid technical integration with eCRF platforms: internal privacy policies and low investments on technical solutions obstacle communication between single centers and sponsors. Hence, physicians, Clinical trial assistants and Data-managers are forced to manually record medical data from hospital EHRs sharing systems to eCRFs. This is a big deal for hospitals that have scientific ambitions.

1.3 Innovative solutions for clinical research

1.3.1 A new era for eCRFs design

Currently, interesting eCRF "ad-hoc" design projects are developing. The Openclinica project is the first example of centralized platform for high-level personalization of eCRFs. Openclinica[6] is a powerful web-based tool where clinical trialists can set-up complex studies to analyze response of treatments on sub-groups of randomized subjects.

In 2004 Vanderbilt University launched the REDCap project. "The REDCap[7] project was developed to provide scientific research teams intuitive and reusable tools for collecting, storing and disseminating project-specific clinical and translational research data. REDcap main strength is: (i) a software generation cycle sufficiently fast to accommodate multiple concurrent projects without the need for custom project-specific programming; and a (ii) model capable of meeting disparate data collection needs of projects across a wide array of scientific disciplines. The concept of metadata-driven application development is well established" (Harris, 2009). The project uses PHP[8] (Hypertext Preprocessor) + JavaScript[9] programming languages and a MySQL[10], a Data-Base Management System (DBMS) for data storage and manipulation. A DBMS is a software system able to manage, with efficacy and efficiency, collections of data that are huge, shared and persistent, ensuring their privacy and reliability. A data base is a data collection managed throw a DBMS (Atzeni, n.d.).

1.3.2 Bioinformatics application on clinical trials

"Omics" together with bioinformatics support must provide real-time data for the therapy. Risk stratification technologies need to be validated and harmonized in order to compare data of different sites. To determine the effectiveness of new treatments the role of observational studies and adaptive tests is continuously growing. Documentation of antitumor treatment efficacy will need relatively small patient groups, while assessment of side-effects will need larger and heterogeneous populations and long-term follow-up (Mendelsohn, 2015). Herein are listed some

[6] https://www.openclinica.com/

[7] https://projectredcap.org/about/

[8] http://php.net/

[9] https://it.wikipedia.org/wiki/JavaScript

[10] https://www.mysql.com/it/

innovative solutions for clinical research: some of these are related to the outcome prediction topic (Kim et al. 2016) (SHIVA clinical trial is considerable first II phase clinical study designed for precision medicine (Servant, 2014)), but also for improving digital security of subjects enrolled in clinical studies (Dernoncourt, 2017; Eubank, 2016). Focusing on bioinformatics tools applied in molecular oncology, Yang and colleagues propose an interesting review (Dancey 2012): Hewett et al. (Hewett and Kijsanayothin 2008) investigated the possibility of making tumor classification more informative by using a method for classification ranking. They applied Multi-Dimensional Ranker on Microarray data of 11 different types and subtypes of cancer. They found that using the classification rankings from Multi-Dimensional Ranker could achieve effective tumor classifications from cancer gene expression data. Wang et al (Wang, 2008) developed a model-based computational approach to detect transcription factors and microRNAs involving the progression of androgen-dependent prostate cancer to androgen-independent prostate cancer. Mehdi et al (Pirooznia, 2008) developed a Java[11] application named Batch Blast Extractor to retrieve data from BLAST[12] output. The tool generates a text file that can be imported into any statistical package such as Excel or SPSS[13] . SMART (Substitutable Medical Applications and Reusable Technology) project needs to be emphasized as a successful open-source application based on the Health IT platform[14]. This is an open-access Application Programming Interface (API) that enables apps to run broadly across the health care ecosystem. "The purpose of the resulting SMART precision cancer medicine app is to present population-level genomic health information to oncologists and their patients in real time as a component of clinical practice" (Warner, 2016). The main objective is showing at the same time demographics data, primary cancer diagnosis, and molecular profile results for clinical consumption. Data were stored in related data ware- houses using an internally developed local code set. A Data Ware-house (DW) is a "collection of integrated, subject- oriented databases designated to support the decision-making process"(Inmon 2005).

[11] www.java.com/
[12] www.blast.ncbi.nlm.nih.gov/Blast.cgi
[13] www.spss.it
[14] www.smarthealthit.org

1.3.3 Data-driven approaches and datamining: the role of Biomedical Engineer

Javier Andreu-Perez and colleagues claim that mining local information included in EHRs data has already been proven to be effective for a wide range of healthcare challenges, such as disease management support, building models for predicting health risk assessment, enhancing knowledge about survival rates, therapeutic recommendation, discovering comorbidities, and building support systems for the recruitment of patients for new trials (Andreu-Perez, 2015). "The term Data Mining identifies a set of tools to search hidden patterns of interest in large and multivariate data sets" (Fayyad, 1996). Medical datamining applications vary from patient outcomes and classification (Fiscon 2014) to image and signal analysis (Rosati, 2014). In 2011, Bellazzi and colleagues proposed an interesting review on the role of biomedical informatics in developing datamining methods. They affirm that "data mining and statistical approaches are no longer seen as alternative ways of dealing with data analysis problems. On the contrary, they are beginning to be fully complementary" (Bellazzi, 2011). Scientific community felt the necessity to determine some rules to attribute reliability to machine learning and datamining methods, to avoid the great risk that a research can fall in "fish expedition" methodologies on data. The traditional idea of knowledge-driven biomedical science should be compared with the evolving data-intensive science where automatic hypotheses are generated among the enormous amount of data available by using computational science with inductive reasoning.

Biomedical Engineer (BE), who follows a pragmatic strategy, may solve the basic dilemma between empiricism and rationalism considering technological constraints and limitations. Moreover, the availability of knowledge repositories in electronic format so strongly empowers bio-medical research that data analysis and knowledge generation steps are now part of a unique, continuous cycle. Bellazzi et al. manly focus on two issues: i) the potential role that BE may have to provide open-access to clinical data; ii) the need for keeping the BE field open to diverse methodological contributions. Thus, BE could play a key role in developing new methods in the field of data mining and machine learning.

1.4 Precision medicine and machine learning vs. clinical research: drawbacks.

1.4.1. Precision medicine, randomized and observational trials

Development of biology is going to improve patients' healthcare and offering targeted treatments based upon "stratification" of patients in small groups. Hence, this will make it possible to precisely tailor healthcare in a personalized manner (Beck, 2012). Precision oncology may be the "Trojan horse" to counter currently obstacles of modern internist medicine: refractory medicine (i), tumors heterogeneity (ii), comorbidities analysis (iii) and cytotoxic drugs interactions. Indeed, study of new cancer genomes is essential to discover innovative tools for molecular diagnosis, to achieve a better understating of cancers. Thus, clinical trials on adult and pediatric people as well as pre-clinical research are a promising strategy for adoption of new therapies. Future researchers are inspired to develop methods to detect, measure and analyze biomedical variables, including molecular, genomic, cellular, clinical, behavioral, physiological and environmental parameters. Moreover, "scientific communities are aware that public and private institutions shall have access to the cohort's data, so that the world's brightest scientific minds can contribute insights and analysis. These data will also enable observational studies of drugs and devices and potentially prompt more rigorous interventional studies that address specific questions"(Francis 2010). However, it seems to be contradictory to affirm the need to strictly stratify patients to personalize healthcare and, on the other side, to encourage international clinical trials and large over-national consortium (as i2B2, Harmony) to evaluate the most promising approaches in a large population over longer periods. Furthermore, how can we face to the paradox based on the issue that "prediction" derives from data of patients enrolled in past studies designed for context where confounding influences may bias results? In addition, some scientists admit that that new markers should be tested on smaller and more homogeneous subgroups of patients to treat with targeted therapy. If the objective is more individualized diagnosis, prognosis, or treatment, one strategy is to "split" the starting population only when there is evidential basis for doing so (Joyner and Paneth 2015). But this is not easy to determine for scientists. The science of discovery presupposes that the individual is isolated from the social context and away from every possible "exposome" and that the cellular data are sufficient to predict the disease. This is impossible: many projects aim at the production of new "biomarkers," but these are not used outside the laboratory environment. These concerns are enlarged by ongoing argument in the oncological community about the proper outcomes for cancer processes and the predictive utility of surrogate endpoints. Moreover, omics data analysis may produce false positive or false negative results in view of such complicated-massive data; due to the limited sensitivity or accuracy of analytical methods, unknown behaviors of molecules which may alter a treatment effectiveness (Shi-kai, 2015). Genomically targeted therapies need to be evaluated by rigorous clinical trials.

1.4.2. Precision medicine, data-mining and machine learning

"Machine-Learning (M-L) predictive algorithms, which can already automatically drive cars, recognize spoken language, and detect credit card fraud, are the keys to unlocking the data that can precisely inform real-time decisions" (Chen and Asch 2017). In healthcare, M-L can improve the ability to establish a prognosis. EHRs and large-scale data ware-houses will provide clinical (as prognostic factors) and biological (as human genomic sequence) data, allowing models to use thousands of rich predictor variables (Obermeyer and Emanuel 2016). Again, computational tools for analyzing large data-sets are enhancing data-cleaning and interpretation of results (Francis, 2010). In addition, M-L will improve diagnostic accuracy. A recent Institute of Medicine report highlighted the alarming frequency of diagnostic errors and the lack of interventions to reduce them (Obermeyer and Emanuel 2016). Clinical medicine has always required doctors to handle enormous amounts of data, from physiology and macro behavior to laboratory and imaging labs. It may be affirmed that M-L approaches problems as a doctor progressing through residency might: by learning rules from data. Starting with patient-level observations, algorithms shift through vast numbers of variables, looking for combinations that reliably predict outcomes. "In recent years, terms such as unsupervised, discovery, and data mining have been used to describe an approach to translational research that proceeds without explicit hypotheses, with conclusions derived from the P values of discovered associations"(Joyner and Paneth 2015). "M-L does not solve any of the fundamental problems of causal inference in observational data sets. Algorithms may be good at predicting outcomes, but predictors are not causes"(Obermeyer and Emanuel 2016). This may be a real risk for clinical research. Hence, machine learning now rides on the "peak of inflated expectations"(Chen and Asch 2017). There are other issues that may influence: (i) EHRs suffer of a lack of a quality in data entry, so that EHRs may retrieve sets with high granularity that obviously effect on data analysis. In addition, the intrinsic variability of clinical data across institutions is magnified by differences in EHRs systems (Figure 3). Consequently, it is critical to figure out how much data to retrieve to not face with fake results. (ii) Validation method must take in account data granularity. Validation on independent data series may be a solution but this does not often limit statistical issues as multiple comparisons theory. Hence, Bonferroni correction is not enough when we compare thousands of features at the same time.

1.5 Straight to the PhD project

This PhD thesis floats in a complex border between clinical onco-hematology and biomedical informatics domains. Heterogeneity of exposed arguments, followed by a strong criticism by part of scientific community, need "simple" messages. The main idea of this thesis is to move backward to a more "clear" management of the onco-hematological data, both clinical and molecular. From data source to data analysis. To do that, the project has been structured on a phase III multicenter clinical trial, with a "translational" vocation. The following chapter is a better contextualization of data quality management methods in clinical trials with the goal to reproduce their main principles in a semi-automated as well as adaptive way.

Chapter 2

I2ECR project

2.1 Background

This chapter proposes a contextualization of data quality management methods in clinical trials with the goal to provide clinical researchers with datasets with a maximized quality. A multicenter phase III hematologic clinical trial enrolls up to hundreds of subjects. In general, a randomized clinical trial evaluates post-treatment outcomes on stratified sub-groups of subjects and it may last several years. Complexity of the study may change and depends by its principal and secondary objectives. However, molecular studies are characterized by gathering of heterogeneous data: clinical, laboratory data, biologic and response to treatment data are only part of the wide types of features. Study sponsors provides centers with either paper-based or paper-less tools to gather data. Data entry is time-consuming, because for a subject hundreds of data are entered. Thus, centers may employees dedicated personnel in data management to internally monitor data-entry. Since last 20 years, sponsors that propose complex clinical trials are dramatically investing to electronic Case Report Forms (eCRFs) platforms to capture the required data at all multicenter trial sites. (ICH 1996). Herein is presented a detailed background about data management, also thought for translational clinical trials. The chapter is developing from the description of MCL0208 clinical trial, portraying, at the end, objectives established during this three years PhD experience.

2.2 Clinical trials monitoring, data Integrity and data type

The International Conference on Harmonization (ICH) of technical requirements for registration of pharmaceuticals for human use defines *Monitoring* within the guideline for Good Clinical Practice (GCP) as well as Good Laboratory Practice (GLP) E6 as in following (ICH Harmonised Tripartite Guideline 1996):

"The act of overseeing the progress of a clinical trial, and of ensuring that it is conducted, recorded, and reported in accordance with the protocol, Standard Operating Procedures (SOPs), Good Clinical Practice (GCP), and the applicable regulatory requirement(s)."

Monitoring starts even before that scientific board opens the enrollment of the trial. Sponsors can prepare centers to optimize data entry so that the actual act of monitoring can begin and happen at any point of the trial value chain (De 2011). Four different monitoring approaches are applicable:

I. Trial oversight committees Monitoring.
II. Central Monitoring.
III. On-sites monitoring.
IV. Adaptive Monitoring.

Committees monitoring are different and operate from the previous phase of the activation of trial to the closing of the enrollment. There are three committees: a committee specialized in preparation of the trial (Trial Management Committee), a committee with strategic skills (Trial Steering Committee) and an independent Data Monitoring Committee for managing central and on-site monitoring. Central monitoring is essential for checking the compliance of centers activity to the central SOPs defined by sponsors. For instance, policies of data harmonization between local laboratories and the central vendor (also called *reviewer laboratory*) must be clear and well-defined in the trial protocol. On-site monitoring is based on a data source verification. Adaptive monitoring is the hybrid monitoring strategy. Committees define the more adaptive monitoring (central or on-sites) basing on data type and centers involved in data entry as well. Quality of data entry is assessed via performance indicators designed by sponsors. The error rate is calculated in (i) base of overall error rate, (ii) relevance of categories in in terms of efficacy, safety, and subject identification and (iii) degree of monitoring observing data source verification protocols (J. R. Andersen et al. 2015).

A verification of compliance of the data presented in the case reports with the source data is carried out to ensure that the collected data are reliable and enable reconstruction and evaluation of the test in reference of accuracy, completeness and verification principles defined by ICH E6. The FDA (Food and Drug Administration) released the "Data Integrity and Compliance with CGMP (Clinical Good Manufacturing Practice) Guidance for industry" documentation. Data **integrity** refers to the completeness, consistency, and accuracy of data" [15]. Officially, the FDA defined clinical data integrity with the A.L.C.O.A. acronym (Woollen 2010). According to ALCOA, data must be:

- Attributable: data should be attributable to user who recorded it after observation.
- Legible: data must be readable.
- Contemporaneous: this element of data quality refers to the timing of data collection with respect of to the time of observation is made.
- Original: data must be compliant to either clinical or laboratory row data.
- Accurate: data must be correct, exact and free from error.

Moreover, eCRFs must be designed in perspective of data analysis. Choice of data type is fundamental for quality control. At a glance, data types may be: numerical, Boolean or logic, of string of characters or comment. In reference of protocol objectives, in perspective of data-entry, data fields designed in eCRFs may be set as mandatory or not. Several eCRFs pages allowed entry in non-mandatory text fields. These text fields are generally excluded from data analysis to avoid bias.

[15] https://www.fda.gov/downloads/drugs/guidances/ucm495891.pdf

2.3 The “FIL-MCL0208” experience: a translational clinical trial for younger patients with Mantle Cell Lymphoma.

2.3.1 Mantle Cell Lymphoma

Mantle cell lymphoma (MCL) is an aggressive neoplasia accounting for 6-8% of all non-Hodgkin's lymphomas and is characterized by the translocation t(11;14)(q13;q32) and the overexpression of cyclin D1. Despite considerable therapeutic progress in the last years, MCL remains a disease difficult to manage, characterized by a poor prognosis in the medium-long term. High-dose chemotherapy, supported by autologous stem cell transplantation (ASCT) is the current standard of care for younger patients, generally providing high response rates and long progression-free survival (PFS), but relapse eventually occurs and patients usually die because of disease progression (Dreyling et al. 2014). However, nowadays MCL has revealed as a highly heterogeneous disease, with some cases extremely aggressive and refractory and others characterized by a better outcome, with stable post-therapy remissions. Therefore, there is urgent need to adapt therapy to the pleomorphic presentation of the disease. The chance to personalize the treatment on the specific characteristics of each patient is made possible by the availability of some validated prognostic tools, such as the MIPI (MCL international prognostic index) and the Ki-67 proliferative index, as well as of early predictors of treatment response, like minimal residual disease (MRD) analysis (Dreyling et al., 2014).

2.3.2 Minimal Residual Disease in MCL.

MRD analysis by allele-specific oligonucleotide (ASO) PCR is able to detect very low levels of residual tumor cells (up to 1 tumor cell out of 100000 healthy cells) in patients achieving complete clinical response (CR) after treatment. This tool, currently applicable to about 90% of MCL patients is an effective early predictor of outcome, showing independent prognostic value in large patients' series and demonstrating superior than the CR achievement in multivariate analysis (Pott et al. 2010). Moreover, MRD prospective assessment is able to early identify patients with increasing risk of upcoming relapse, monitoring those patients experiencing “MRD reappearance” (as to say, again positive MRD results) and thus prone to relapse in the course of the next years (Pott et al. 2014): these “high-risk” patients could be ideal subjects to receive pre-emptive treatments, aimed at avoiding a more challenging full-blown relapse (Andersen et al., 2009). Thus, in the near future MRD analysis might be used to stratify MCL patients into different risk classes, to whom offer a personalized treatment, as already happening in other hematological tumors, such as acute lymphoblastic leukemia (Gökbuget et al. 2014).

2.3.3 The MCL0208 clinical trial.

Multicenter, randomized, Phase III clinical trial MCL0208 sponsored by the Fondazione Italiana Linfomi, FIL (EudraCT code: 2009-012807-25) has been designed to assess MRD by ASO quantitative PCR of bone marrow (BM) and peripheral blood (PB) samples prospectively collected in the molecular biology facilities in Torino from the MCL patients. This trial has recently completed the accrual of 300 young (< 65 years), advanced stage, MCL patients. The therapeutic schedule (figure 4) included a chemo-immunotherapy induction phase (**Restaging 1**), followed by high dose cytarabine treatment, peripheral blood stem cells (PBSC) collection (**Restaging 2**) and ASCT (**Restaging 3**). Finally, responding patients are randomized between maintenance with lenalidomide for 24 months or observation. For this study MRD will be assessed by both qualitative nested PCR and quantitative real time PCR(RQ-PCR) on BM, PB and leukapheresis samples at planned time points: 1) baseline; 2) after R-CHOP induction; 3) leukapheresis; 4) before ASCT; 5) after ASCT; 6) during maintenance/observation at months 6-12-18-24-30-36 (Cortelazzo et al. 2015).

In addition, the BM and PB samples collection by the centralized lab in Torino has already been organized to build a complete biobank of MCL prospective cases and some biological ancillary studies have been already planned and detailed in the MCL0208 protocol (each of these have been performed by experienced hematology vendors in different centers and coordinated by the FIL). In particular the following ancillary studies have been planned:

I. Deep mutational sequencing analysis. Deep sequencing analysis project consisted on performing of this technique on a MCL gene panel (ATM, TP53, CCND1, KMT2D/MLL2, WHSC1, TRAF2, NOTCH1, BIRC3) in the perspective series of patient enrolled in MCL0208 (Francis S. Collins and Harold Varmus, 2010).
II. Gene Expression Profiling (GEP). The aim of this project is to use a GEP approach to identify MCL subsets with peculiar clinical/biological features

in the context of MCL patients treated homogeneously with an autologous transplantation-based program(Fletcher, Robert H.; Fletcher, Suzanne W.; Fletcher 2003).

III. Pharmacogenomics. Pharmacogenomics project main goal is to investigate relationships between antineoplastic drug pharmacodynamics and pharmacogenetics factors (e.g. gene polymorphisms) (Joyner and Paneth 2015).

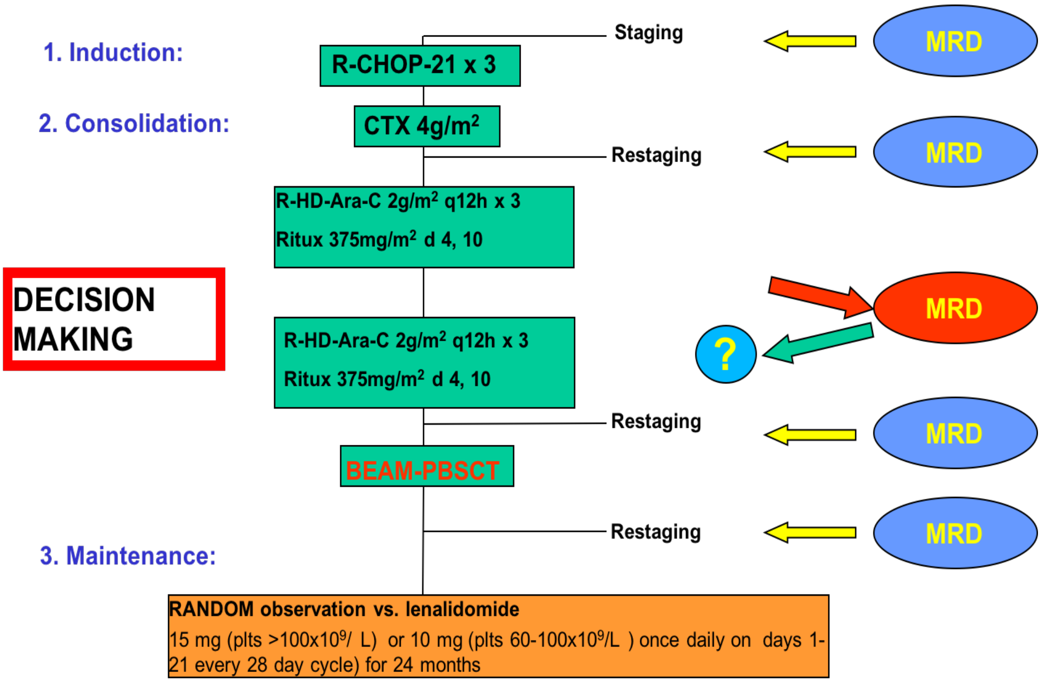

1. Induction:
R-CHOP-21 x 3
Staging
MRD
2. Consolidation:
CTX 4g/m^2
Restaging
MRD
R-HD-Ara-C 2g/m^2 q12h x 3
Ritux 375mg/m^2 d 4, 10
DECISION MAKING
MRD
?
R-HD-Ara-C 2g/m^2 q12h x 3
Ritux 375mg/m^2 d 4, 10
Restaging
MRD
BEAM-PBSCT
Restaging
MRD
3. Maintenance:
RANDOM observation vs. lenalidomide
15 mg (plts >100x10^9/ L) or 10 mg (plts 60-100x10^9/L) once daily on days 1-21 every 28 day cycle) for 24 months

Figure 4: MCL0208 clinical trial general work-flow. Courtesy of FIL (Fondazione Italiana Linfomi).

MCL0208 clinical trial may be considered as a study with a great translational vocation because it collects several ancillary studies over primary objectives declared into the protocol.

2.3.4 MCL0208 data management.

FIL strongly leverages on networking between centers on national proposing both phase II and III clinical trials. The FIL clinical studies may be clinical and/or molecular, covering different level of complexity in the trial management. A central data management office coordinates all management activities of either biological samples or the datasets as defined in the protocol of the clinical trial. However, clinical data monitoring on hundreds of observed subjects which are characterized by clinical, biologic and molecular data is complex and time-consuming. Clinical trials sponsored by Pharmaceutical Companies usually outsource data management to third party organizations called CRO (Contract Research Organizations). However, this solution request high monetary investments that are not easily affordable by no-profit organizations.

MCL0208 clinical trial suffered of a lack of central control on data. Basically, this may also depend by the data-heterogeneity (more than 350 variables for time-point) that, if multiplied for 300 enrolled subjects, reach up to 10^5 orders of magnitude. In absence of a dedicated CRO monitoring on this clinical trial, to automatize remote monitoring has been a good strategy to centrally monitor data-quality and, furtherly, to handle cleaned up dataset ready for statistical analysis. From this need, I2ECR project has been proposed. The aim is to overcome the lacks of data management, using innovative technical tools currently applied in both public or private fields (e.g. marketing to profile customers). Furthermore, this project represents an interesting opportunity to test innovative tools in order to provide clinical researchers a precision medicine methodology integrating clinical and omics data retrieved from a clinical trial.

2.4 I2ECR objectives

I2ECR is an **Integrated and Intelligent Environment for Clinical Research** where clinical and omics data stand together for clinical use (reporting) and for generation of new clinical knowledge. I2ECR is adapted to MCL0208 phase III trial, which is a translational trial with several clinical prognostic factors (e.g. MIPI - Mantle International Prognostic Index) associated to treatment data, biological assessment of disease (MRD - Minimal Residue Disease) and genetic ancillary studies as Pharmacogenomics or Mutational Analysis.

I2ECR primary objective is to propose an integration project on clinical and molecular data.

The application of a clear row-data analysis as well as clinical trial monitoring strategies may guarantee to **implement a digital platform** where clinical, biologic and “omics” data are correctly **imported** from different sources and well-**integrated in a data-ware-house**.

Hence, clinicians, biomedical engineers, biostatisticians, biologist and data-managers will be able to **control**, in a semi-automatic manner, **quality of data**, in relation to the clinical data imported from eCRFs (i), from biologic datasets internally edited by local vendors (ii) and from mutational datasets externally edited by working groups on ancillary studies (iii). Therefore, I2ECR will be able to detect missing data and mistakes derived from some non-conventional data-entry activities by centers.

I2ECR secondary objective is to be a dynamic repository of data congruency rules

These rules will be established by both physicians and biomedical engineers, who can **easily encode** them in the platform. For instance, I2ECR must be able to detect a mistake in the determination of pathologic status (or risk assessment at diagnosis) of a patient, simply comparing all features (clinical, MRD and mutational) that are recognized as negatively prognostic for that malignancy.

I2ECR third objective is to be a platform where researchers can easily design statistical and data mining analysis.

Data Mining (DM) identifies “a set of tools for searching for hidden patterns of interest in large and multivariate datasets”(Fayyad, Piatetsky-Shapiro, and Smyth 1996). “Applications of DM techniques in the medical field range from outcome prediction and patient classification to genomic medicine and molecular biology” (Zaccaria, Rosati, et al. 2017). I2ECR allows clinical stake-holders to propose innovative methods of supervised and unsupervised feature extraction, data classification and statistical analysis on heterogeneous datasets associated to MCL0208 clinical trial. Although MCL0208 study is the first example of data-population of I2ECR, the environment will be able to import clinical studies for all onco-hematologic diseases.

Chapter 3

Methodology

“Clinical trials are designed to produce new knowledge about a disease, drug or treatment” (Gholap et al. 2015). A multicenter phase III hematologic clinical trial enrolls up to thousands of subjects. A randomized clinical trial evaluates post-treatment outcomes on stratified sub-groups of subjects. During these studies, a huge amount of data is collected about participants, therapies, clinical procedures, outcomes, adverse events and so on. Therefore, data collection in clinical trials is becoming complex, with huge amount of clinical and biological variables. Low-quality of the collected clinical data, in terms of incomplete or incorrect values, effects on incorrect calculation of outcome prediction. This means that “quality decision must be based on quality data”(Halkidi, Vazirgiannis, and Batistakis 2000). Han et al. proposed that data preprocessing techniques can be grouped in four main classes (Han, Kamber, and Pei 2012): data integration (i), data cleaning (ii), data transformation (iii) and data reduction (iv). Data integration allows for merging data from multiple sources into a homogeneous dataset. Data cleaning is usually applied to remove noise. Data transformation techniques transform data into forms that are appropriate for the Data-Mining processing. Finally, data reduction eliminates redundant and irrelevant variables. Outside the medical field, data warehouses (DWs) are widely employed to achieve these objectives. A Data Warehouse (DW) is a “collection of integrated, subject-oriented databases designated to support the decision-making process”(Inmon 2005). To verify whether DWs might be useful for data quality and association analysis, a team of biomedical engineers, clinicians, biologists and statisticians developed I2ECR project. I2ECR is an Integrated and Intelligent Environment for Clinical Research. I2ECR is adapted to MCL0208 phase III trial, which is a translational trial with several clinical prognostic factors (e.g. MIPI - Mantle International Prognostic Index (Hoster et al. 2008)) associated to treatment data, biological assessment of disease (MRD - Minimal Residue Disease) and ancillary studies as Pharmacogenomics, Pathology, Mutational Analysis and GEP (Gene Expression Profile). For MCL0208, 48 Italian medical centers were actively involved in the trial, for a total of 300 enrolled subjects (age: 55±8 years). I2ECR main objectives are:

- to propose an integration project on clinical and molecular data quality concepts.
- To be a dynamic repository of data congruency quality rules.
- To provide to clinical stake-holders a platform from where they can easily design statistical and data mining analysis.

Chapter 3 is structured in two main parts:

- section 3.1 – adopted methodology in suiting a pipe-line on data management for clinical trial, from data source to data analysis phases.
- Section 3.2 – adopted methods in designing (i) and implementing (ii) the I2ECR software, which is considered the final object to provide to clinical researchers for applying pipe-line proposed in section 3.1.

3.1 I2ECR project

The I2ECR project is designed on a Data-Ware house. DW management is allowed by Extraction, Transformation and Loading (ETL) processes. “An ETL process is the cornerstone component that supplies the DWs with all the necessary data” (Akkaoui et al. 2011). In I2ECR, the DW is suited on MCL0208 clinical trial. DW has been named **FIL_MCL0208**. FIL_MCL0208 input data has been retrieved from different datasets collected in electronic data sheets (.csv or .xls). A team of biomedical engineers, clinicians, biologists, statisticians re-shaped logical organization of data in sub-groups. The DW has been implemented in MySQL®[16]. ETL processes have been executed via Matlab®[17] (MATrix LABoratory) database toolbox. Loading, transformation and extraction of data stored in DW has been possible through JDBC[18] (Java Data-Base Connectivity) and ODBC[19] (Oracle

[16] https://www.mysql.com/it/

[17] https://it.mathworks.com/products/matlab.html

[18] http://www.oracle.com/technetwork/java/overview-141217.html

[19] http://www.oracle.com/technetwork/database/windows/index-098976.html

Data-Base Connectivity) bridges, which connect Matlab database toolbox to DBMS. The I2ECR ETL environment is depicted in figure 5.

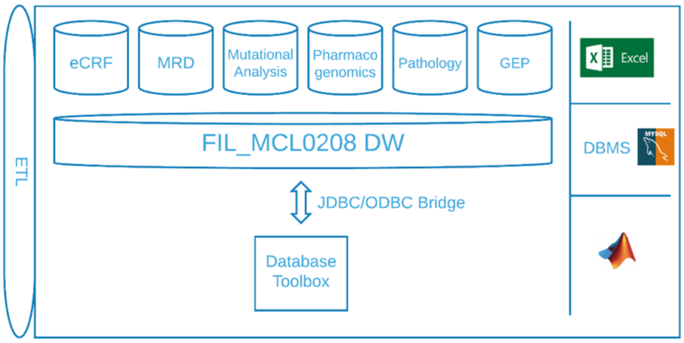


Figure 5: I2ECR environment

FIL_MCL0208 DW has been suited to a large clinical trial from the Fondazione Italiana Linfomi (FIL), including all clinical, biological and mutational features collected both in the trial and in the ancillary studies. During the clinical trial, several variables or features were acquired one CRFs, describing the patient status at the diagnosis and at different time points. Figure 6 shows all steps tailoring I2ECR project. This pipe-line starts from patients' personal health record (MCL0208 clinical study) and it ends with patients' stratification (1). "Small data sources" (2) were Integrated and stored into FIL_MCl0208 DW (3). Once that clinical, molecular and mutational data have been deployed in the DW, a complex cleaning phase was computed (4). At this point, "cleaned" datasets (.csv or .xls format) were exported in IECR output files in order to report to each single center data entry mistakes to be fixed (5). Furtherly, Data transformation (6) has been mandatory to harmonize thousands of data to set-up datasets for statistical and data-mining analysis (7) and eventually extracting new knowledge.

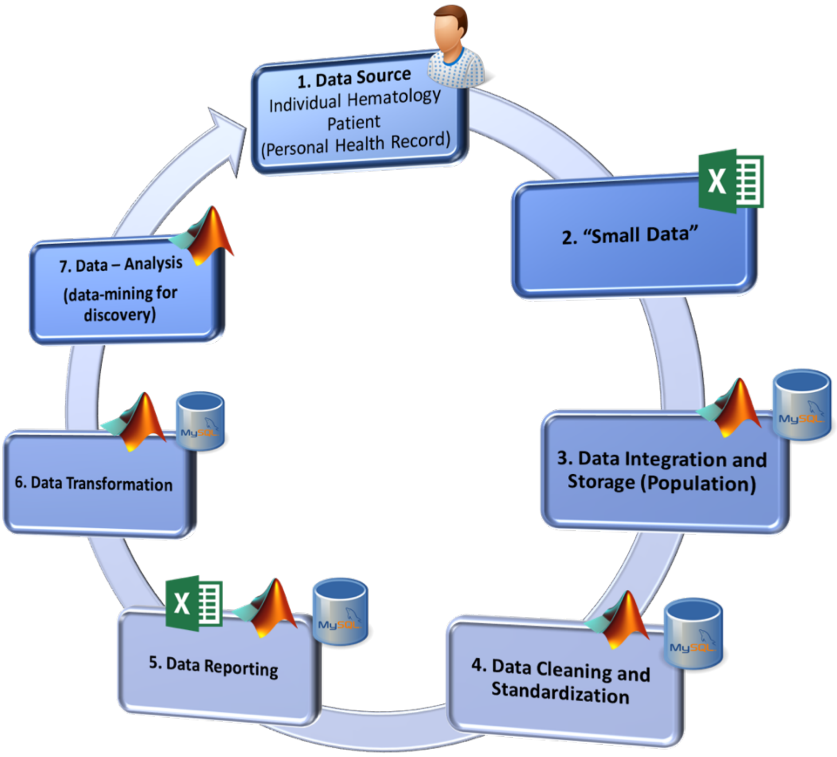

1. Data Source
Individual Hematology Patient
(Personal Health Record)
2. "Small Data"
3. Data Integration and Storage (Population)
4. Data Cleaning and Standardization
5. Data Reporting
6. Data Transformation
7. Data – Analysis
(data-mining for discovery)

Figure 6: The I2ECR pipeline.

3.1.1 Data source

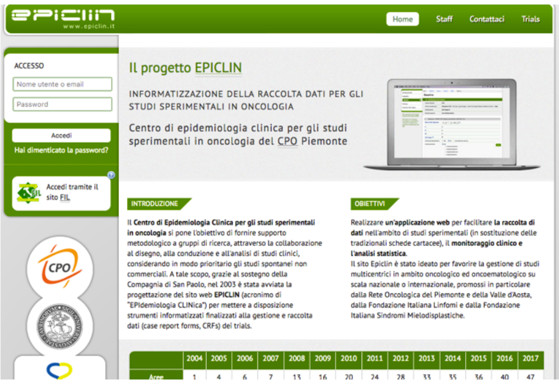
Patients' health records of a translational clinical trial derive from different sources as shown in figure 5. Clinical Health Records for patients enrolled in MCL0208 clinical trial have been retrieved from eCRF web-site (i) and from ancillary studies (ii).

eCRF

The MCL0208 eCRF is a web-based system (figure 7) from whom users can manage:

- Protocol inclusion and exclusion criteria and eventual study interruption GUIs (Graphic User Interfaces).
- Clinical data at diagnosis (Baseline time-point).
- Clinical data at following restaging (up to 3) and follow-ups (up to 5).
- Treatment plan, actual treatment administered and possible Adverse Events (AE).
- MRD recorded at different restaging and follow-ups as well.

eCRF web-site is managed by an administrator, who, in case of MCL0208 is a statistician. This latter also plays the role of data-custodian. Users interacting with eCRF are basically, data-managers, physicians, biostatisticians, research assistants and biologists. There are several authorization levels to protect patients' privacy.



Il progetto EPICLIN

INFORMATIZZAZIONE DELLA RACCOLTA DATI PER GLI STUDI SPERIMENTALI IN ONCOLOGIA

Centro di epidemiologia clinica per gli studi sperimentali in oncologia del CPO Piemonte

INTRODUZIONE

Il **Centro di Epidemiologia Clinica per gli studi sperimentali in oncologia** si pone l'obiettivo di fornire supporto metodologico a gruppi di ricerca, attraverso la collaborazione al disegno, alla conduzione e all'analisi di studi clinici, considerando in modo prioritario gli studi spontanei non commerciali. A tale scopo, grazie al sostegno della Compagnia di San Paolo, nel 2003 è stata avviata la progettazione del sito web **EPICLIN** (acronimo di "EPIdemiologia CLINica") per mettere a disposizione strumenti informatizzati finalizzati alla gestione e raccolta dati (case report forms, CRFs) dei trials.

OBIETTIVI

Realizzare **un'applicazione web** per facilitare **la raccolta di dati** nell'ambito di studi sperimentali (in sostituzione delle tradizionali schede cartacee), il **monitoraggio clinico e l'analisi statistica**.
Il sito Epiclin è stato ideato per favorire la gestione di studi multicentrici in ambito oncologico ed oncoematologico su scala nazionale o internazionale, promossi in particolare dalla Rete Oncologica del Piemonte e della Valle d'Aosta, dalla Fondazione Italiana Linfomi e dalla Fondazione Italiana Sindromi Mielodisplastiche.

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Aree	1	4	6	7	13	16	20	24	28	33	35	36	40	47

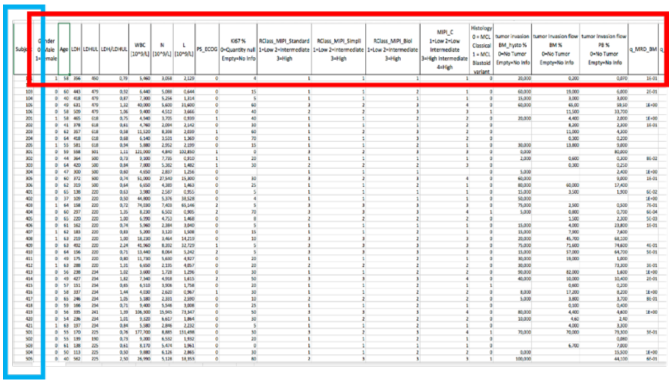
Figure 7: the eCRF platform designed for MCL0208 clinical trial

Ancillary studies

MCL0208 is a translational clinical trial because it involves several ancillary studies: mutational analysis (i), Pharmacogenomics (ii), a centralized Pathology study (iii) and GEP (iv). In order to conduct ancillary studies (laboratories are placed in different cities of the country), patients' samples management has been hardly coordinated by central secretariat of FIL organization. Once results from ancillary studies were assessed from investigators, new data sheets needed to be integrated to clinical health records.

3.1.2 "Small data"

Clinical trial data-custodian is generally the one authorized to export data for both data monitoring and statistical analysis. System exports data in non-proprietary electronic sheets (.csv). A sheet from eCRF is depicted in Figure 8. The first column is the sorted list of Subjects; First row collects all features labeled by administrator during export. Ancillary study sheets shall have the same structure of eCRF sheets to simplify integration.



Subject	Gender 0=Male 1=Female	Age	LDH	LDHUL	LDH/LDHUL	WBC [10^9/L]	N [10^9/L]	L [10^9/L]	PS_ECOG	Ki67 % 0=Quantity null Empty=No Info	RClass_MIPI_Standard 1=Low 2=Intermediate 3=High	RClass_MIPI_Simpli 1=Low 2=Intermediate 3=High	RClass_MIPI_Biol 1=Low 2=Intermediate 3=High	MIPI_C 1=Low 2=Low Intermediate 3=High Intermediate 4=High	Histology 0 = MCL Classical 1 = MCL Blastoid variant	tumor invasion BM_hysto % 0=No Tumor Empty=No Info	tumor invasion flow BM % 0=No Tumor Empty=No Info	tumor invasion flow PB % 0=No Tumor Empty=No Info	q_MRD_BM	q
[illegible]	1	58	356	450	0,79	5,060	3,058	2,129	0	4	1	1	1	1	0	20,000	0,200	0,070	1E-01	
103	0	60	443	479	0,92	6,440	5,088	0,644	0	15	1	1	2	1	0	60,000	19,000	6,000	2E-01	
104	0	40	418	479	0,87	7,300	5,256	1,314	0	5	1	1	1	1	0	15,000	3,000	3,000		
105	0	49	631	479	1,32	40,000	5,600	31,600	0	60	3	2	3	4	0	60,000	65,00	59,50	1E+00	
106	0	58	509	479	1,06	9,400	4,512	3,666	0	40	2	2	3	3	1		11,500	33,700		
201	1	58	465	618	0,75	4,940	3,705	0,939	1	40	1	1	2	2	0	20,000	4,400	2,000	1E+00	
202	0	41	378	618	0,61	4,760	2,094	2,142	0	30	1	1	1	2	0		8,200	2,300	1E-01	
203	0	62	357	618	0,58	11,520	8,308	2,030	1	60	1	2	3	2	0		11,000	4,300		
204	0	64	418	618	0,68	6,140	3,531	1,360	0	70	1	1	3	2	0		0,300	0,200		
205	1	55	581	618	0,94	5,880	2,952	2,199	0	15	1	1	2	1	0	30,000	13,800	3,000		
301	0	59	558	501	1,11	121,000	4,840	102,850	1	0	3	2	3	3	0	0,000		80,000		
302	0	44	364	500	0,73	9,100	7,735	0,910	1	20	1	1	1	1	0	2,000	0,600	0,300	8E-02	
303	0	64	420	500	0,84	7,800	5,382	1,482	1	30	2	2	2	3	0		0,300	0,250		
304	0	47	300	500	0,60	4,650	2,837	1,256	0		1	1	1		0	5,000		2,400	1E+00	
305	0	60	372	500	0,74	51,000	27,540	15,300	0	30	3	2	3	4	0	60,000		9,000	1E-01	
306	0	62	319	500	0,64	6,650	4,389	1,463	0	25	1	1	2	1	0	80,000	60,000	17,400		
401	0	65	138	220	0,63	3,980	2,587	0,955	0	5	1	1	1	1	0	15,000	3,500	1,900	6E-02	
402	0	37	109	220	0,50	44,800	5,376	38,528	0	4	1	1	1	1	0	50,000			1E+00	
403	1	64	158	220	0,72	74,030	7,403	65,146	2	5	3	3	3	3	0	75,000	2,500	0,500	7E-01	
404	0	60	297	220	1,35	8,230	6,502	0,905	2	70	3	3	3	4	1	5,000	0,800	0,700	6E-04	
405	0	65	220	220	1,00	6,990	4,753	1,468	0	8	2	2	2	2	0		1,500	2,300	5E-03	
406	0	61	162	220	0,74	5,960	2,384	3,040	0	5	1	1	1	1	0	15,000	4,000	23,800	1E-01	
407	1	62	183	220	0,83	5,200	3,120	1,508	0	15	1	1	2	1	0	15,000	7,900	7,600		
408	1	63	219	220	1,00	18,230	3,464	14,219	0	10	3	3	2	3	0	20,000	45,700	68,100		
409	0	63	492	220	2,24	41,960	8,392	32,729	1	5	3	3	3	3	0	75,000	71,600	74,600	4E-01	
410	0	64	156	220	0,71	13,440	8,064	5,242	2	5	3	3	3	3	0	15,000	57,000	64,700	5E-01	
411	0	49	175	220	0,80	11,730	5,630	4,927	0	20	1	1	2	1	0	30,000	19,000	1,000		
412	1	63	288	220	1,31	6,650	2,135	4,057	0	20	2	2	2	2	0	30,000		73,300	3E-01	
413	0	56	238	234	1,02	3,600	1,728	1,296	0	30	1	1	2	2	0	90,000	82,000	1,600	1E+00	
414	0	49	427	234	1,82	7,340	4,918	1,615	2	50	3	3	3	4	0	40,000	10,000	10,400	2E-01	
415	0	57	151	234	0,65	6,510	3,906	1,758	0	20	1	1	2	1	1		0,600	0,200		
416	0	58	337	234	1,44	4,030	2,620	0,967	1	30	1	1	2	2	0	8,000	17,200	8,200	1E+00	
417	0	65	246	234	1,05	5,180	2,331	2,590	0	10	2	2	2	2	0	5,000	3,800	3,700	8E-01	
418	0	59	166	234	0,71	9,400	5,546	3,008	0	25	1	1	2	1	0		0,100	0,400		
419	0	56	335	241	1,39	106,900	15,945	73,947	0	50	3	3	3	4	0	80,000	4,400	4,600	1E+00	
420	0	54	236	234	1,01	9,320	6,617	1,864	0	30	1	2	2	2	0	10,000	4,60	2,40		
421	1	63	197	234	0,84	5,580	2,846	2,232	0	5	1	1	2	1	0		4,000	3,300		
501	0	55	170	225	0,76	177,700	8,885	133,498	0	50	3	2	3	4	1	70,000	70,000	79,300	5E-01	
502	0	55	139	190	0,73	9,200	6,532	1,932	0	20	1	1	2	1	0			0,080		
503	0	61	138	225	0,61	8,170	5,474	1,961	0	0	1	1	1	1	0		6,700	7,000		
504	0	50	113	225	0,50	9,880	6,126	2,865	0	30	1	1	2	2	0	0,000		15,500	1E+00	
505	0	40	562	225	2,50	26,990	5,128	18,953	0	80	2	3	3	3	1	100,000		44,100	6E-01	

Figure 8: a data sheet exported from eCRF: rows (blue box) are subjects, columns (red box) are the features included in dataset.

3.1.3 Data integration and storage (population)

Data integration is usually necessary in case of data extracted from different sources, each collecting parameters with different orders of magnitude, units of measurement, or ranges of validity. A team of biomedical engineers, clinicians, biologists, statisticians studied each dataset from corresponding electronic sheet. Sharing their different skills, they though to a new organization of data. The target was to shape an innovative structure of tables where attributes' congregation was driven by a logic proximity.

3.1.3.1 FIL_MCL0208 data ware-house

FIL_MCL0208 DW has a "snowflake" architecture (Han, Kamber, and Pei 2012). Figure 9 shows the logic Entity Relational (ER) schema, designed via MySQL Workbench®. Center of the model (red) is composed by tables containing information about enrolled Subjects (i), active Centers (ii) and Protocol information (iii). Identifying Relationship: an identifying relationship is one "where the child table cannot be uniquely identified without its parent". For this DW, a child table must be filled up by records whom secondary keys are identified by mother's attribute codeSubject. Practically, codeSubject attribute of child tables must be the same ID recorded in "codeSubject" attribute recorded in the "Subject" table. So that, in case of recording a new subject which is not included in codeSubject records (within Subject table), DBMS rejects the SQL command with an error message. Children tables of Subject are:

- Protocol_info (red).
- Clinical_Data_Baseline and Clinical_Data_Restaging (yellow).
- Laboratory_Data (light green).
- Pathologic_Data (green).
- Diag_Procedures (purple).
- MRD_Baseline, MRD_LK and MRD_Restaging (magenta).
- Gene, GEP and Pharmacogenomics (strawberry).
- Treatment and Transplant (light blue).
- Toxicity (blue).

Relationships between Toxicity table and toxicity children tables (more details in following) follow the same approach. Cardinality: cardinality between tables depends from time-points associated to input datasets that populated the DW. MCL0208 multi-centric clinical trial was designed on 3 restaging and 5 follow-ups. Focusing on the ER diagram, this temporal requirement has been implemented adding the attribute "N_timepoint". For tables that include N_timepoint attribute, cardinality (from Subject to children tables) is set **0:many** because data for the same subjects may be repeated up to 8 times. The list of tables including "N_timepoint" attribute is:

- Laboratory_Data (light green).
- Pathologic_Data (green)
- Diag_Procedures (purple).
- Clinical_Data_Restaging (yellow).
- MRD_Restaging (magenta).
- Treatment and Transplant (light blue).

However, relationships between Subject and under-cited tables are set with **0:1** cardinality. This is because children tables involve data collected only at one timepoint:

- Clinical_Data_Baseline (yellow).
- MRD_Baseline, MRD_LK (magenta).
- Gene, GEP and Pharmacogenomics (strawberry).

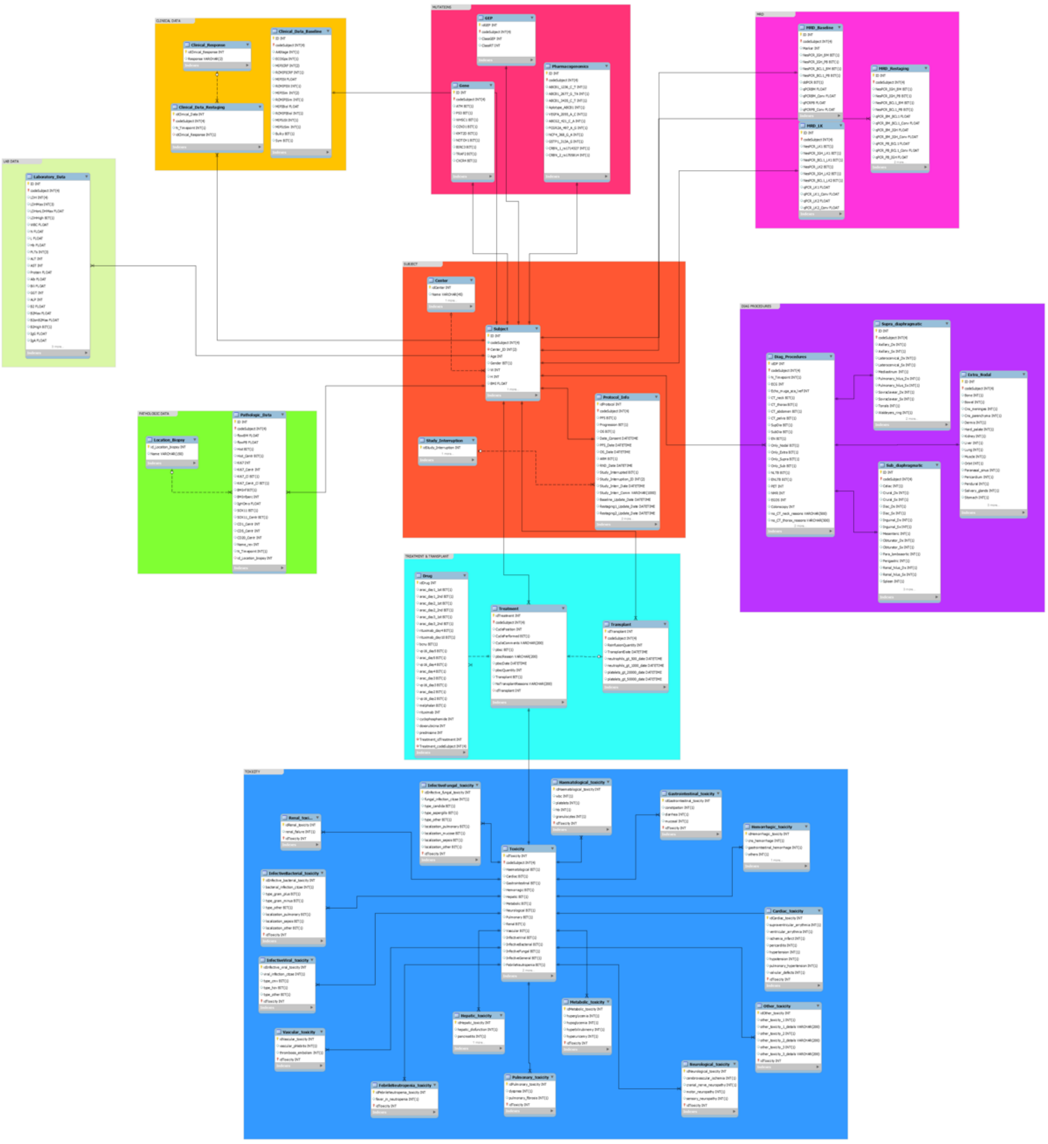


Figure 9: FIL_MCL0208 data warehouse.

SUBJECT sub-group.

SUBJECT group of tables (red in figure 9) is the backbone of the DW. This group includes tables Subject (i), Protocol_info (ii), Center (iii) and Study_interruption (iv). A subject is associated to a center through a many:1 cardinality. In fact, one center may enroll one or more subjects. Protocol_info is an extension of Subject table: this table is externally identified from codeSubject. A patient may interrupt clinical trial for several reasons. This information must be recorded in Protocol_info table. In case of interruption, interruption cause is expressed through Study_Interruption table (many:1 cardinality). Therefore, an interruption cause may be associated to several subjects.

- Subject table is characterized by:
  - Age, Gender: demographics attributes.
  - W, H, BSA, BMI: morphometric attributes (W – weight, H – height, BSA – Body Surface Area (Mosteller 1987) and BMI – Body Mass Index (Diehr Diane E.Harris, 1998)).
- Center: this table includes all centers active in enrollment. Each center is expressed through both CenterID and Location attributes. Centers is a mandatory table. Table of centers is shown in ANNEX 1.
- Protocol_info includes:
  - Progression, PFS, OS: outcome variables (PFS - Progression Free Survival, OS – Overall Survival). Every subject assumes 1 if outcome is defined or 0 if not.
  - Date_Consent, PFS_Date, OS_Date, RND_Date: timestamp attributes related to consent signature by patient (Date_Consent), outcomes (PFS_Date, OS_Date), enrollment randomization (RND_Date). Date_Consent is assumed as the official patient's enrollment starting point.
  - ARM: the treatment ARM (A or B) randomly assigned to patient after 3rd Restaging.
  - Baseline_Update_Date, Restaging1_Update_Date, Restaging2_Update_Date and Restaging3_Update_Date: operational timestamps of dataset loading (Baseline_Update_Date, Restaging1_Update_Date, Restaging2_Update_Date and Restaging3_Update_Date). Those timestamps are helpful for monitoring the DW update status from eCRF sheets.
  - Study_Interr_Date and Study_Interruption_ID: information about study interruption status.

- Study_Interruption table is a list of interruption status as ruled in the study protocol. Reasons of interruption study for a patient are listed in table 1.

Table 1: Study Interruption classes

Study_Interruption	Name
1	Adverse Event
2	Withdrawal of consent
3	Poor Compliance
4	Serious breach of protocol
5	Progression
6	Decision of responsible of the study
7	Dispersed during the study
8	Other
9	Death

LAB_DATA sub-group.

Laboratory_Data table (light green in figure 9) collects data from several blood-draws gathered at different timepoints (attribute N_timepoint). This table includes attributes derived from:

- WBC, N, L, Hb, PLTs: Complete Blood Count (CBC): WBC (White Blood Cells) and its components (Neutrophils – N and Lymphocytes – L), Hb (Hemoglobin), PLTs (Platelets).
- LDH: Lactate Dehydrogenase.
- Alb, Bili, GGT, ALP, AST, ALT: albumin, bilirubin, gammaGT (GGT), Alkaline Phosphatase (ALP), Aspartate Transaminase (AST) and alanine transaminase (ALT) are indexes of liver status.

- Protein, ß2, IgG, IgA, IgM: total proteins (Protein), ß2 Macroglobulin's and Immunoglobulins (IgG, IgA, IgM) levels.
- Uricemy: toxicity index.
- ß2Max, ß2onß2Max, LDHMax, LDHonLDHMax: "input variables ß2 and LDH are laboratory measurement whose values are obtained according to the local vendor. This means that the threshold for discriminating between normal and altered values (ß2onß2Max and LDHonLDHMax) can be different.

PATHOLOGIC_DATA sub-group.

PATHOLOGIC_DATA sub-group (green in figure 9) includes variables of pathologist's assessment. Pathologic_Data (i) and Location_Biopsy (ii) tables are included:

- Pathologic_Data table is defined by following attributes:
  - N_timepoint: timepoint associated to pathology analysis. As a matter of fact, a patient may repeat pathology investigations several times during protocol development.
  - BMInf, BMInfperc: bone marrow tumor invasion by immunochemistry. Bone marrow samples are drawn from patients via biopsy. Biopsy location is described from id_Location_biopsy attribute, externally identified in Location_Biopsy table.
  - Hist: histology evaluation has high prognostic impact on MCL affected patients (Tiemann et al. 2005). It assumes normal or blastoid classification. Within MCL0208 clinical study, histology has been assessed by both local (Hist) and centralized (Hist_Centr) pathologists. In I2ECR Hist assumes 0 if normal or 1 if blastoid.
  - SOX11: SOX11 (Vegliante et al. 2013) is a protein responsible of neural transcription, found to be over-expressed in leukemic MCL cells. In MC0208 clinical trial, SOX11 is assessed by both local (SOX11) and centralized (SOX11_Centr) pathologists.
  - Ki_67: Ki_67 is a proliferation marker expressed in high level among MCL effected patients (Jares, Colomer, and Campo 2012). Ki67 is measured in % and categorized in two classes (Ki67_Cl: 0 for values <30% vs. 1 for values >=30%). During MCL0208 study, this value has been assessed by both local (Ki67 and Ki67_CI) and centralized (Ki67_Centr and Ki67_Centr_CI) pathologists.
  - CD1, CD5 and CD20: CD1, CD5 and CD20 biomarkers are considered prognostic in MCL (Swerdlow and Williams 2002). Herein, these

    biomarkers are assessed by centralized pathologists (CD1_Centr, CD5_Centr, CD20_Centr).
  - flowBM, flowPB: disease assessment via flow-cytofluorimetry on either bone marrow (flowBM) or peripheral blood samples (flowPB).
  - IgHOmo: omology to IgH "germline" configuration was assessed by both local pathologist and biologists (Alamyar et al. 2012).
  - Name_rev: it records the name of the reviewer pathologist.
- Location_Biopsy: this table includes the anatomic location (attribute Name) by whom biopsy has been executed (ANNEX 2). Pathologic_Data and Location_Biopsy have been associated with a cardinality of **many:1**. In fact, biopsy procedure is mandatory in base of what established by trial protocol.

CLINICAL_DATA sub-group.

CLINICAL_DATA sub-group (yellow in figure 9) is composed by Clinical_Data_Baseline, Clinical_Data_Restaging and Clinical_Response tables.

- Clinical_Data_Baseline table includes attributes derived from non-Hodgkin lymphoma clinical prognostic factors assessed at baseline:
- ECOG: Eastern Cooperative Oncology Group performance status (Ghielmini et al. 2013). In I2ECR, ECOG may assume a discrete value from 0 to 4.
- Bulky: often used to describe large tumors in the chest. In MCL0208, Bulky is 1 when the detected mass is >5cm, otherwise the attribute assumes 0.
- Sym: it indicates the class of symptoms recorded by clinical staff. For mantle cell lymphoma (MCL), symptoms may be grouped in class A or class B (Mallick, Lal, and Daugherty 2017). In I2ECR Sym attribute is 0 for class A symptoms or 1 for class B of symptoms.
- AAstage: Ann Arbor stage (AAstage) shows whether the mantle cell lymphoma is in one area of body (localized) or has spread to other areas. AAstage may assume a discrete value from 1 to 4 (Vose 2015).
- MIPI indexes. MIPI (MCL International Prognostic Index (Hoster et al. 2008)) is a prognostic index of overall survival that groups patients into 3 classes (low, intermediate and high risk) based on four independent clinical variables: age, ECOG performance status, LDH and WBC. In addition, MIPI extensions were developed: biologic MIPI (MIPIb) and MIPIc involve the same independent variables that compose MIPI plus Ki67 additive value (Hoster et al. 2016). In FIL_MCL0208, MIPI indexes were encoded as follows:
  - MIPICRF: MIPI standard imported from eCRF.

  - RCMIPICRF: risk classes imported from eCRF.
  - MIPISt: MIPI standard.
  - RCMIPISt: risk classes derived from MIPI standard index.
  - MIPISim: MIPI simplified.
  - RCMIPISim: risk classes derived from MIPI simplified index.
  - MIPIb: MIPI biologic.
  - RCMIPIb: risk classes derived from MIPI biologic index.
  - MIPICSt: MIPIC obtained from MIPI standard index.
  - MIPICSim: MIPIC obtained from MIPI simplified index.
- Clinical_Data_Restaging: in this table, clinical responses to treatment for each restaging timepoint were recorded (Bruce D. Cheson, 1999). Time-points are provided by N_timepoint attribute. idClinical_Response attribute is externally identified by Clinical Response's primary key with a cardinality of **many:0**.
- Clinical_Response: it collects the list of clinical response options. Each idClinical_Response is associated to its Response attribute which contains the clinical response names (table 2).

Table 2: Classes of clinical responses

idClinical_Response	Response	Description
1	CR	Complete Response
2	PR	Partial Response
3	SD	Stable Disease
4	PD	Progression Disease

DIAG_PROCEDURES sub-group.

DIAG_PROCEDURES (purple in figure 9) sub-group involves tables designed to manage information on diagnostic instrumental procedures. Diag_Procedures (i), Supra_diaphragmatic (ii), Sub_diaphragmatic (iii) and Extra-Nodal tables were included. Subject and Diag_Procedures table are associated with a cardinality of **0:many**. Supra_diaphragmatic, Sub_diaphragmatic and Extra-Nodal tables are extensions of Diag_Procedures tables. Those are externally identified by as idDP, which is the Diag_Procedures primary key, as code Subjects, which is Diag_Procedures secondary key.

- Diag_Procedures has been designed with following attributes:
  - N_timepoint: it describes timepoint associated to that diagnostic procedure. As a matter of fact, a patient may repeat a diagnostic imaging analysis several times during protocol development.
  - ECG: it indicates the execution (value 1) or not (value 0) of ECG (Electro Cardio-Graph) diagnostic exam.
  - Echo_muga_scan_lvef: Multiple Gated Acquisition scan to assess left ventricular ejection fraction. It assesses patient's cardiac performance quantifying the % of volume of blood ejected from left ventricle on its total volume.
  - CT_neck, CT_thorax, CT_abdomen, CT_pelvis: these attributes indicate if there is an involvement (value 1) or not (value 0) in a precise anatomical zone.
  - SupDIA, SubDia, EN: attributes that testify the lymph-nodes involvement detected by CT scan. Every attribute assumes 1 if there is involvement, 0 if not.
  - Only_Supra, Only_Sub, Only_Extra: these attributes define if there is only a supra-diaphragmatic, sub_diaphragmatic or extra-nodal involvement (1 vs 0).
  - PET: it indicates if PET (Positron Emission Tomography) has been executed (it assumes 1 value) or not (0). In case of the exam has been executed, the variable may assume 2 if there is an involvement.
  - NMR: it indicates if NMR (Nuclear Magnetic Resonance) has been executed (it assumes 1 value) or not (0). In case of the exam has been executed, the variable may assume 2 if there is at list one involvement.
  - EGDS: it indicates if EGDS (Esophagus-Gastro-Duodenoscopy) has been executed (it assumes 1 value) or not (0). In case of the exam has been executed, the variable may assume 2 if there is at least one involvement.
  - Colonoscopy: it indicates if Colonoscopy has been executed (it assumes 1 value) or not (0). In case of the exam has been executed, the variable may assume 2 if there is at least one involvement.
  - NLTB: Nodal Low Tumor Burden. It is an aggregate variable that assumes 1 if conditions were verified: there is (i) a nodal involvement (SupraDia =

  1 or SubDia = 1 and EN = 0), (ii) infiltrated bone marrow (BMInf = 1) and (iii) Bulky > 5 [cm] (Bulky = 1). Otherwise the attribute assumes value 0.
  - ENLTB: Extra-Nodal Low Tumor Burden. It is an aggregate variable that assumes 1 if conditions were verified: there is (i) an extra-nodal involvement (SupraDia = 0 and SubDia = 0 and EN = 1), (ii) infiltrated bone marrow (BMInf = 1) and (iii) Bulky > 5 [cm] (Bulky = 1). Otherwise the attribute assumes value 0.
- Supra_diaphragmatic: in case of supra-diaphragmatic lymph-node involvement by CT scan, at least one of attributes must be set at 1. Each attribute is a specific anatomic localization (e.g. Axillary DX).
- Sub_diaphragmatic: in case of sub-diaphragmatic lymph-node involvement by CT scan, at least one of attributes must be set at 1. Each attribute is a specific anatomic localization (e.g. Inguinal_DX).
- Extra_nodal: in case of extra nodal involvement by CT scan, at least one of attributes must be set at 1. Each attribute is a specific anatomic localization (e.g. Liver).

MRD sub-group.

MRD sub-group (magenta from figure 9) involves by MRD_Baseline (i), MRD_LK (ii) and MRD_Restaging (iii) tables. Each table is externally identified by codeSubject:

- MRD_Baseline: it gives information about baseline tumor burden assessment with PCR.
  - Marker: it indicates the marker used for tumor burden assessment (0=BCL1, 1=IgH, 2=both).
  - NesPCR_IGH_BM, NesPCR_IGH_PB, NesPCR_BCL1_BM, NesPCR_BCL1_PB: qualitative tumor burden assessment via PCR. These attributes testify if nested PCR on either bone marrow (BM) or peripheral blood (PB) samples through IGH or BCL1 marker were executed (value 1) or not (0).
  - ddPCR: it checks if digital PCR technique has been performed on baseline samples (Drandi et al. 2016).
  - qPCRBM, qPCRPB: quantity of tumor burden assessed via PCR on BM or PB. This quantity is expressed on logarithmic scale.
  - qPCRBM_Conv, qPCRPB_Conv: linear equivalent scale of qPCRBM and qPCRPB.

- MRD_LK: it collects information of tumor burden assessment by PCR at intermediate timepoint previous to pre-chemotherapy leukapheresis processes (Strunk et al. 2005).
  - NesPCR_LK1, NesPCR_LK2: qualitative tumor burden assessment via PCR. PCR has been executed in two-time points called LK1 and LK2. Attributes assume 1 if the test has been done or 0 if not.
  - NesPCR_IGH_LK1, NesPCR_IGH_LK2, NesPCR_BCL1_LK1, NesPCR_BCL1_LK2: qualitative tumor burden assessment via PCR. PCR has been executed in two timepoints called LK1 and LK2 analyzing 2 different biomarkers (IGH and BCL1). Attributes assume 1 if the test has been done or 0 if not.
  - qPCR_LK1, qPCR_LK2: quantity of tumor burden assessed via PCR at LK1 and LK2 time-points. This quantity is expressed on logarithmic scale.
  - qPCR_LK1_Conv, qPCR_LK2_Conv: linear equivalent scale of qPCR_LK1 and qPCR_LK2.
- MRD_Restaging: it collects information of baseline tumor burden assessment with PCR at time-points after baseline.
  - N_timepoint: time-point associated to MRD analysis.
  - NesPCR_IGH_BM, NesPCR_IGH_PB, NesPCR_BCL1_BM, NesPCR_BCL1_PB: qualitative tumor burden assessment via PCR. These attributes testify if nested PCR on either bone marrow (BM) or peripheral blood (PB) samples through IGH or BCL1 marker have been executed (value 1) or not (0).
  - qPCR_BM_BCL1, qPCR_PB_BCL1, qPCR_BM_IGH, qPCR_PB_IGH: quantity of tumor burden assessed both via PCR on BM or PB and through IGH or BCL1 marker analysis. This quantity is expressed on logarithmic scale.
  - qPCR_BM_BCL1_Conv,qPCR_PB_BCL1_Conv,qPCR_BM_IGH_Conv, qPCR_PB_IGH_Conv: linear equivalent scale of qPCR_BM_BCL1, qPCR_PB_BCL1, qPCR_BM_IGH and qPCR_PB_IG.

MUTATIONS sub-group.

This group (strawberry in figure 9) collects Gene (i), GEP (ii) and Pharmacogenomics (iii) tables.

- Gene table includes mutational analysis contribution for each MCL0208 subject at baseline (if available). ATM, P53, WHSC1, KMT2D, NOTCH1, BIRC3, TRAF2, CXCR4: each attribute is a dummy variable that indicates if that gene has mutated (1) or not (0).

- GEP table includes the Gene Expression Profile unsupervised classification for a sample at baseline (if available):
    - ClassGEP attribute is 0 if a subject was classified in class 1 or 1 if he was classified in class 2.
    - ClassRT attribute describes the Real-Time classification methodology used as best practice. It assumes 0 if a subject was classified in class 1 or 1 if was classified in class 2.
- Pharmacogenomics: this table collects polymorphism analyzed on targeted genes (ABCB1, VEGFA, ABCG2, FCGR2A, NCF4). Polymorphisms are classified as not present: 0=WT (Wild Type); 1: heterozygote polymorphism; 2: homozygote polymorphism.

TREATMENT & TRANSPLANT sub-group.

TREATMENT & TRANSPLANT sub-group (light blue in figure 9) is composed by Treatment (i), Transplant (ii) and Drug (iii) tables. A Subject may receive 0 or more pharmacologic treatment (cardinality **0:many**) as well as 0 or 1 bone marrow transplant (**0:1**). Each treatment may be composed by 1 or more drug (cardinality **1:many**). Treatment and Transplant tables are externally identified by codeSubject attribute (from Subject table) and Drug is externally identified by Traetment_idTreatment and Treatment_codeSubject as well. In that way, DBMS is set to both associate a treatment and a transplant to a codeSubject (previously recorded in Subject table) and a drug to idTreatment and to a codeSubject (included in Treatment table). Treatment and Transplant have a relationship of **1:0** of cardinality (a transplant is associated to a treatment at least once because a bone marrow transplantation follows treatment).

- Treatment: it is composed by attributes inherited from treatment sheets and exported from eCRF.
    - CyclePosition: it indicates at what cycle the treatment belong.
    - CyclePerformed: this attribute defines if a cycle has been administered or not.
    - pbsc: peripheral stem cell transplantation. This attribute testifies if patient received a stem cell transplantation of peripheral blood (value 1) or not (0).
    - pbscDate: date of peripheral blood stem cells transplantation.
    - pbscQuantity: total quantity of stem cells transplanted.

  - Transplant: it indicates if autologous bone marrow transplant was executed (value 1) or not (value 0).
  - IdTransplant: externally key inherited from Transplant.
- Transplant: this table is composed by attributes that define the transplant.
  - Reinfusion_Quantity: total quantity of stem cells infusion.
  - TransplantDate: date time of stem cells transplantation.
  - neutrophils_gr_500_date, neutrophils_gt_1000_date, platelets_gt_20000_date, platelets_gr_50000_date: signs of engraftment. Timestamp recorded when a patient, in a post-transplant status, reaches both quantities of neutrophils at 500 or at 1000 [10^9/L] and quantities of platelets at 20000 or at 50000 [10^9/L].
- Drug: this table collects information about drug administration.
  - arac (from day1 to day5): treatment with Ara-C administration in a planned day (value 1) or not (0).
  - rituximab (day4 and day10): treatment with Rituximab administration in a planned day (value 1) or not (0).
  - vp16 (from day 2 to day 6): treatment with etoposide administration in a planned day (value 1) or not (0).
  - bcnu: treatment with carmastine administration (value 1) or not (0).
  - melphalan: treatment with melphalan administration (value 1) or not (0).
  - rituximab, cyclophosphamide, doxorubicine, prednisone: administered dose to patients.

TOXICITY sub-group.

Toxicity sub-group (blue from figure 9) includes Toxicity (i), Hematological_toxicity (ii), InfectiveFungal_toxicit (iii), Renal_toxicity (iv), InfectiveBacterial_toxicity (v), InfectiveViral_toxicity (vi), Vascular_toxicity (vii), FebrileNeutropenia_toxicity (viii), Pulmonary_toxicity (ix), Hepatic_toxicity (x),

Metabolic_toxicity (xi), Neurological_toxicity (xii), Gastrointestinal_toxicity (xiii), Hemorrahagic_toxicity (xiv), Cardiac_toxicity (xv), Other_toxicity (xvi) tables. About a design point of view, this “star” structure defines Toxicity table as mother and other tables as children. Children tables are externally identified by idToxicity with a cardinality of **0:many**. In fact, considering that a toxicity is identified by a treatment (idTreatment), it is possible that for a treatment are associated 0 toxicities or more:

- Toxicity table: it collects all toxicities recordable from Adverse Event (AE) recorded during a treatment. Each toxicity may be activated (value 1) or not (value 0). Toxicities are harmonized on CTCAE (Common Terminology Criteria for Adverse Event) international standard[20].A toxicity grade may assume several levels as shown in table 3.

Table 3: Classification of toxicities in reference of CTCAE international standard.

Grade	Description
1	Mild; asymptomatic or mild symptoms; clinical or diagnostic observations only; intervention not indicated.
2	Moderate; minimal, local or noninvasive intervention indicated; limiting age-appropriate instrumental ADL.
3	Severe or medically significant but not immediately life-threatening; hospitalization or prolongation of hospitalization indicated; disabling; limiting self-care ADL**.
4	Life-threatening consequences; urgent intervention indicated.
5	Death related to AE

The grade associated to each single toxicity is included into children tables, where each toxicity was classified in attributes following a top-down strategy:

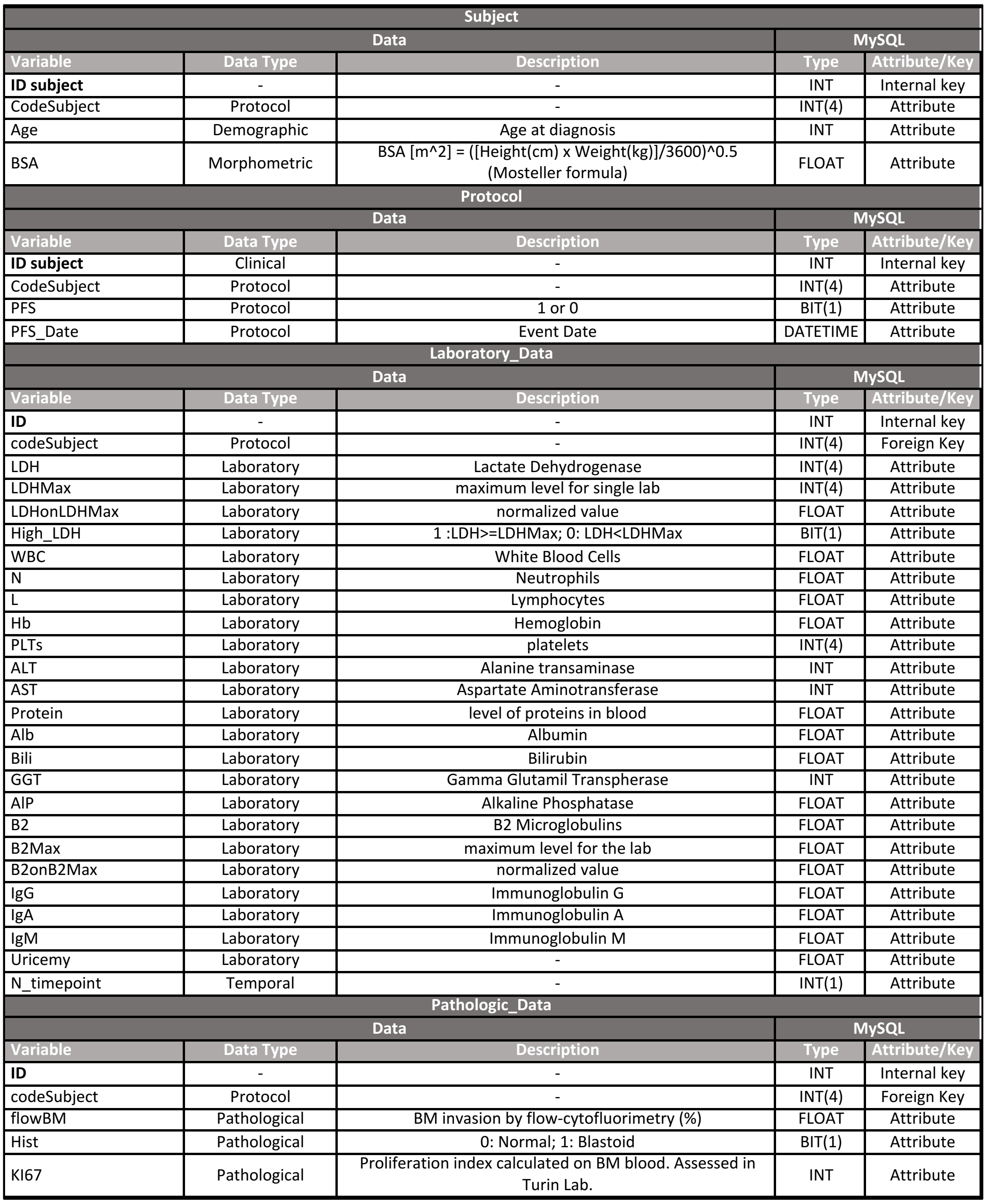
- Hematological_toxicity: WBC, PLTs, Hb, granulocytes.
- Renal_toxicity: renal_failure.
- Vascular_toxicity: vascular_phlebitis and thrombosis_embolism.
- FebrileNeutropenia_toxicity: fever_in_neutropenia.
- Pulmonary_toxicity: dyspnea and pulmonary_fibrosis.
- Hepatic_toxicity: hepatic_disfunction, pancreatitis.
- Metabolic_toxicity: hyperglycemia, hypoglycemia, hyperbilirubinemy, hyperuricemy.

[20] https://evs.nci.nih.gov/ftp1/CTCAE/CTCAE_4.03_2010-06-14_QuickReference_8.5x11.pdf

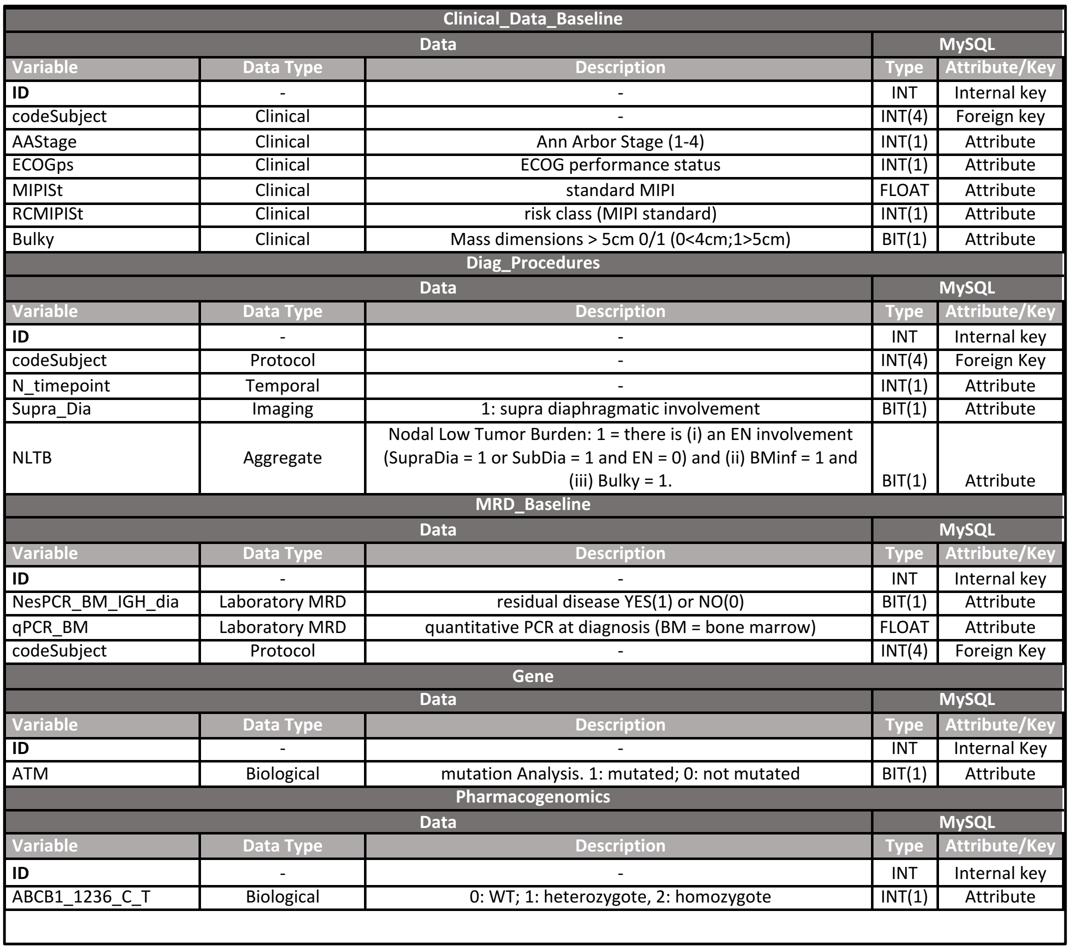
- Neurological_toxicity: cerebrovascular_ischemia, cranial_nerve_neuropathy, motor_neuropathy, sensory_neuropathy.
- Gastrointestinal_toxicity: constipation, diarrhea, mucosal.
- Hemorrahagic_toxicity: cns_hemorrhage, gastrointestinal_hemorrhage.
- Cardiac_toxicity: supraventricular_arrythmia, ventricular_arrythmia, ischemia_infarct, pericarditis, hypertension, hypotension, pulmonary_hypertension, valvular_defects.
- InfectiveFungal_toxicity: it gives information about eventual fungal infection.
  - fungal_infection_ctcae: grade of infection basing on CTCAE scale.
  - type_candida, type_aspergillo, type_other: type of infection.
  - localization_pulmonary, localization_mucose, localization_sepsis: localization of infection.
- InfectiveBacterial_toxicity:
  - bacterial_infection_ctcae: grade of infection basing on CTCAE scale.
  - type_gram_plus, type_gram_minus, type_other: type of infection.
  - localization_pulmonary, localization_sepsis: localization of infection.
- InfectiveViral_toxicity:
  - viral_infection_ctcae: grade of infection basing on CTCAE scale.
  - Type_cmv, type_hzv, type_other: type of infection.

ANNEX 2 collects FIL_MCL0208 DW details about each table and attributes. Table 4 is an extracted and it shows a sub-group of attributes: the most relevant are those associated to a measure.

Table 4: Extract from FIL_MCL0208 pool of variables



Subject				
Data			MySQL	
Variable	Data Type	Description	Type	Attribute/Key
ID subject	-	-	INT	Internal key
CodeSubject	Protocol	-	INT(4)	Attribute
Age	Demographic	Age at diagnosis	INT	Attribute
BSA	Morphometric	BSA [m^2] = ([Height(cm) x Weight(kg)]/3600)^0.5 (Mosteller formula)	FLOAT	Attribute
Protocol				
Data			MySQL	
Variable	Data Type	Description	Type	Attribute/Key
ID subject	Clinical	-	INT	Internal key
CodeSubject	Protocol	-	INT(4)	Attribute
PFS	Protocol	1 or 0	BIT(1)	Attribute
PFS_Date	Protocol	Event Date	DATETIME	Attribute
Laboratory_Data				
Data			MySQL	
Variable	Data Type	Description	Type	Attribute/Key
ID	-	-	INT	Internal key
codeSubject	Protocol	-	INT(4)	Foreign Key
LDH	Laboratory	Lactate Dehydrogenase	INT(4)	Attribute
LDHMax	Laboratory	maximum level for single lab	INT(4)	Attribute
LDHonLDHMax	Laboratory	normalized value	FLOAT	Attribute
High_LDH	Laboratory	1 :LDH>=LDHMax; 0: LDH<LDHMax	BIT(1)	Attribute
WBC	Laboratory	White Blood Cells	FLOAT	Attribute
N	Laboratory	Neutrophils	FLOAT	Attribute
L	Laboratory	Lymphocytes	FLOAT	Attribute
Hb	Laboratory	Hemoglobin	FLOAT	Attribute
PLTs	Laboratory	platelets	INT(4)	Attribute
ALT	Laboratory	Alanine transaminase	INT	Attribute
AST	Laboratory	Aspartate Aminotransferase	INT	Attribute
Protein	Laboratory	level of proteins in blood	FLOAT	Attribute
Alb	Laboratory	Albumin	FLOAT	Attribute
Bili	Laboratory	Bilirubin	FLOAT	Attribute
GGT	Laboratory	Gamma Glutamil Transpherase	INT	Attribute
AlP	Laboratory	Alkaline Phosphatase	FLOAT	Attribute
B2	Laboratory	B2 Microglobulins	FLOAT	Attribute
B2Max	Laboratory	maximum level for the lab	FLOAT	Attribute
B2onB2Max	Laboratory	normalized value	FLOAT	Attribute
IgG	Laboratory	Immunoglobulin G	FLOAT	Attribute
IgA	Laboratory	Immunoglobulin A	FLOAT	Attribute
IgM	Laboratory	Immunoglobulin M	FLOAT	Attribute
Uricemy	Laboratory	-	FLOAT	Attribute
N_timepoint	Temporal	-	INT(1)	Attribute
Pathologic_Data				
Data			MySQL	
Variable	Data Type	Description	Type	Attribute/Key
ID	-	-	INT	Internal key
codeSubject	Protocol	-	INT(4)	Foreign Key
flowBM	Pathological	BM invasion by flow-cytofluorimetry (%)	FLOAT	Attribute
Hist	Pathological	0: Normal; 1: Blastoid	BIT(1)	Attribute
KI67	Pathological	Proliferation index calculated on BM blood. Assessed in Turin Lab.	INT	Attribute



Clinical_Data_Baseline				
Data			MySQL	
Variable	Data Type	Description	Type	Attribute/Key
ID	-	-	INT	Internal key
codeSubject	Clinical	-	INT(4)	Foreign key
AAStage	Clinical	Ann Arbor Stage (1-4)	INT(1)	Attribute
ECOGps	Clinical	ECOG performance status	INT(1)	Attribute
MIPISt	Clinical	standard MIPI	FLOAT	Attribute
RCMIPISt	Clinical	risk class (MIPI standard)	INT(1)	Attribute
Bulky	Clinical	Mass dimensions > 5cm 0/1 (0<4cm;1>5cm)	BIT(1)	Attribute
Diag_Procedures				
Data			MySQL	
Variable	Data Type	Description	Type	Attribute/Key
ID	-	-	INT	Internal key
codeSubject	Protocol	-	INT(4)	Foreign Key
N_timepoint	Temporal	-	INT(1)	Attribute
Supra_Dia	Imaging	1: supra diaphragmatic involvement	BIT(1)	Attribute
NLTB	Aggregate	Nodal Low Tumor Burden: 1 = there is (i) an EN involvement (SupraDia = 1 or SubDia = 1 and EN = 0) and (ii) BMinf = 1 and (iii) Bulky = 1.	BIT(1)	Attribute
MRD_Baseline				
Data			MySQL	
Variable	Data Type	Description	Type	Attribute/Key
ID	-	-	INT	Internal key
NesPCR_BM_IGH_dia	Laboratory MRD	residual disease YES(1) or NO(0)	BIT(1)	Attribute
qPCR_BM	Laboratory MRD	quantitative PCR at diagnosis (BM = bone marrow)	FLOAT	Attribute
codeSubject	Protocol	-	INT(4)	Foreign Key
Gene				
Data			MySQL	
Variable	Data Type	Description	Type	Attribute/Key
ID	-	-	INT	Internal Key
ATM	Biological	mutation Analysis. 1: mutated; 0: not mutated	BIT(1)	Attribute
Pharmacogenomics				
Data			MySQL	
Variable	Data Type	Description	Type	Attribute/Key
ID	-	-	INT	Internal key
ABCB1_1236_C_T	Biological	0: WT; 1: heterozygote, 2: homozygote	INT(1)	Attribute

3.1.3.2 Data storage (population)

Basing on I2ECR architecture, explained in figure 5, population of DW was allowed through a bridge between DBMS and Matlab®. Data storage in the DW consisted in implementation of several functions. Functions were classified in 3 types:

- Functions of Data Loading and Updating.
- Functions of Cleaning and Harmonization.
- Functions for calculation of aggregate variables.

Data Loading and Update: these functions were created to populate the DW. Functions activities consisted in:

- Loading a .csv and .xls file.
- Recognizing subjects and features.
- Importing data from file in pre-allocated tables.
- Connecting with DBMS sending INSERT (or UPDATE) SQL commands:

```
%DB Conncection
conn = database('FIL','root','admin2','Vendor','MySQL','Server','localhost');

%Data import from source file (.xls)
[NUMERIC_Clinics,TXT_Clinics]=xlsread('MCL0208_clinics');

%Subject table population (from exdata matrix)
colnames={'IDSubjectcol','codeSubject','ARM', 'Date_Consent' ,'RND_Date','PFS',
'Progression',...
'Death','PFS_Date','Last_OS_Date','Age_Dia','Gender','Study_Interrupted','Study_Interr_Date'
,...
'Study_Interr_Spec','Study_Interr_Comm','Study_interruption_idStudy_Interruption',...
'Baseline_Update_Date','Restaging1_Update_Date','Restaging2_Update_Date','Restaging3_Update_
Date', ...
'Treatment_Update_Date','Center_idCenter'};
fastinsert(conn,'subject',colnames,exdata);

%”Date_Consent” attribute update from Subject table
colnames={'Date_Consent'};
id=cell2mat(C1(:,1));
pk_value=cell2mat(C1(:,1));
where = arrayfun(@(id) sprintf('WHERE codeSubject = %d', id), pk_value, 'UniformOutput',
false);
update(conn,'Subject',colnames, C1(:,2), where);
```

Data Cleaning and Harmonization: this part will be deeply analyzed in Paragraph 3.1.4.

Calculations of aggregate variables: data management with ware-house approach gives the opportunity to easily create aggregate variables: I2ECR includes variables obtained implementing mathematical models (of different level of complexity. Table 5 shows all aggregate variables and their sources.

Table 5: Aggregate variables from FIL_MCL0208 and related sources

Derived Variables	Source Variables
BMI	W, H
BSA	W, H
MIPISt, MIPISim	Age, LDH, WBC, ECOGps
MIPIb	Age, LDH, WBC, ECOGps, Ki67
RClassMIPISt, RClassMIPISim, RClassMIPIBiol	MIPISt, MIPISim, MIPIb
MIPICSt, MIPICSim	MIPISt, MIPISim, Ki67
Only_Nodal, Only_Extra, Only_Supra, Only_Sub	SupraDia, SubDia, EN
NLTB	BMInf, Bulky, Only_Nodal
ENLTB	BMInf, Bulky, Only_Extra

MIPI index is one example of the automatic calculation of an aggregate variable, this is the part of Matlab code:

```
%% MIPI STANDARD CALCULATION- MIPI_Std

% MIPI score = [0.03535 * age (years)] +
% 0.6978 (if ECOG > 1) +
% [1.367 * log10(LDH/ULN)] +
% [0.9393 * log10(WBC count 106/l)]

MIPI_Std=NaN(length(mipi(:,1)),2);
MIPI_Std(:,1)=mipi(:,1);

for i=1:1:length(mipi(:,1))
    if mipi(i,3)>1
        A=0.03535*(mipi(i,2)); % Age
        B=0.6978; % ECOG
        C1=(mipi(i,4)/mipi(i,5));
        C=1.367*log10(C1); %LDH/ULN
        D=0.9393*log10(mipi(i,6)*1000);%WBC 106/L
        MIPI_Std(i,2)=round(A+B+C+D,2);
    else
        A=0.03535*(mipi(i,2)); %Age
        C1=(mipi(i,4)/mipi(i,5));
        C=1.367*log10(C1); %LDH/ULN
        D=0.9393*log10(mipi(i,6)*1000);%WBC 106/L
        MIPI_Std(i,2)=round(A+C+D,2);
    end
    clear A B C C1 D;
end
```

Example of risk class calculation from MIPI standard in reference of Hoster et al. (Hoster et al. 2008):

```
% RCMIPIStd calculation
RCMIPI_Std=NaN(length(mipi(:,1)),2);
RCMIPI_Std(:,1)=mipi(:,1);

for i=1:1:length(mipi(:,1))
    if MIPI_Std(i,2)<5.7
       RCMIPI_Std(i,2)=1;
    end
    if (MIPI_Std(i,2)>5.69 && MIPI_Std(i,2)<6.2)
       RCMIPI_Std(i,2)=2;
    end
    if MIPI_Std(i,2)>6.19
       RCMIPI_Std(i,2)=3;
    end
end
```

3.1.4 Data cleaning and standardization

In this subsection, explanation of data cleaning strategies for I2ECR is proposed. Row datasets are affected by several data-entry mistakes. Here, cleaning strategies may be classified in two main classes:

- I level controls.
- II level controls.

I Level Controls

I level controls are strictly focused on a single variable. For each variable, I level controls have the objective to find errors on each single data imported from input dataset. In this section, a simple classification of errors is provided:

a) Missing value.
b) Nulls.
c) Ranging errors.

a) Missing Value (MV): a missing value is a data considered mandatory for data analysis, but actually not filled in eCRFs by centers. Each trial protocol establishes what are mandatory data for a clinical study. These data are fundamental for statistical analysis for study endpoints achievement and each center must correctly fill in eCRF platforms.

In case of temporary MV (e.g. due for a temporary unavailability of laboratory data or for a delay of sample shipment between center and vendor), data-managers avoid eCRF system constraints, entering unconventional data as. ‘0’, ‘-’, ‘-1’, ‘-9’ and so on. If the eCRF platform does not include implemented back-office controls, a manual control should be necessary for maximizing data quality.

b) Null: nulls not allowed. Biologic, pathology and physiologic variables cannot assume a 0 value. Table 6 lists I2ECR variables from whom a NULL value is not admissible.
c) Ranging errors: validity ranging errors in data entry. In nature, defining a validity range for a biologic value actually is problematic. Nevertheless, in collaboration with clinicians and taking in consideration Unit of Magnitude of each variable, validity ranges with large intervals were defined (table 7 and ANNEX 2).

WBC variable is a representative example of I level control. Generally, WBC count is expressed with both unit of magnitude equal to [10^9/L] or [10^6/L]. This divergence may reflect on data entry management. Despite sponsors previously stimulate centers to be aware to choose a unique data-entry strategy in eCRF, uniformities are commons. In MCL0208 eCRFs, 41 subjects with a WBC expressed in [10^6/L] (13% of total) were found, reflecting on MIPI automatic calculation provided to users. To avoid that, in I2ECR an automatic conversion has been implemented:

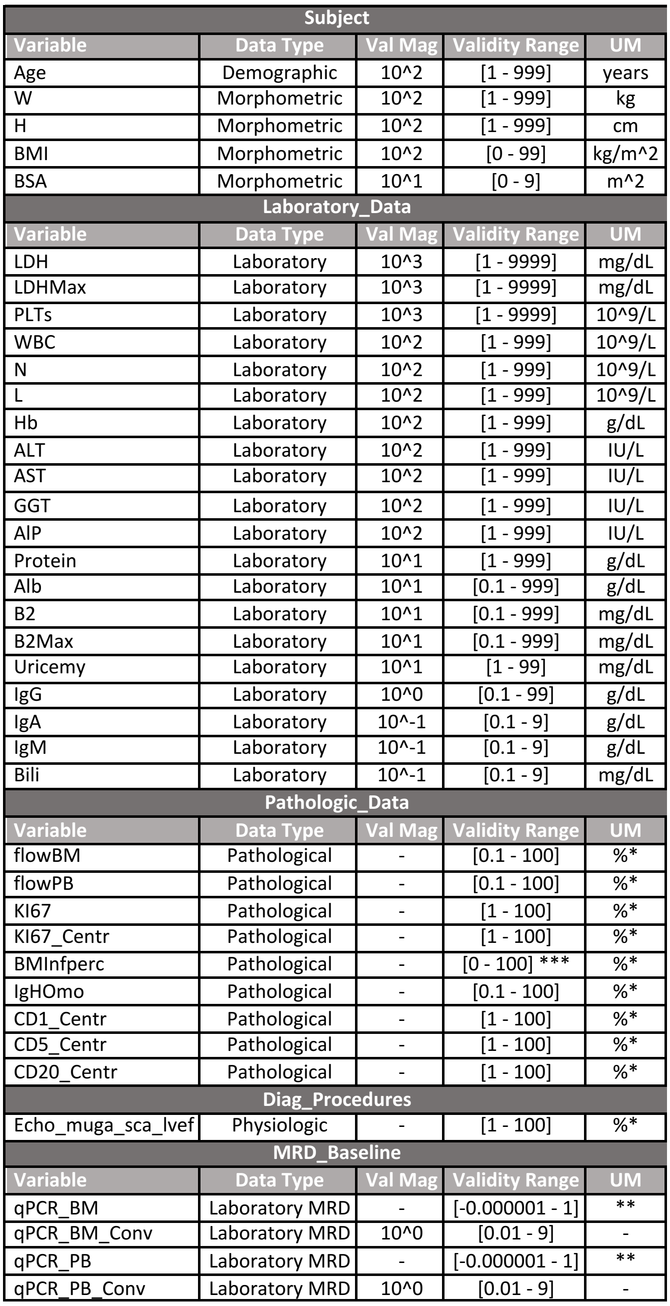
```
for i=1:1:length(WBC)
      if WBC(i,2)/1000 > 1
          WBC(i,2)=round((WBC(i,2)/1000),2);
      end
end
```

In this case, if WBC/1000 ratio is more than 1, WBC count has been recorded in [10^6/L] and automatically converted in [10^9/L] (a patient with a count of WBC equal to 1.000 [10^9/L] is not likelihood).

Table 6: Variables on whom I2ECR I Level Controls have been applied

I2ECR Tables	Variables
Subject	Age, W, H, BMI, BSA
Laboratory_Data	LDH, WBC, N, L, Hb, PLTs, ALT, AST, Protein, Alb, Bili, GGT, ALP, B2, IgG, IgH, IgM, Uricemy
Pathologic_Data	flowBM, flowPB, Ki67, IgHOmo
Clinical_Data_Baseline	AAstage, MIPIb
Diag_Procedures	Echo_muga_sca_lvef
MRD_Baseline	qPCRBM, qPCRPB

Table 7: List features grouped for tables implemented in DW, every feature is described data type, Validity Magnitude, Validity Range and Unit of Magnitude. *: features expressed in percentage. **= features expressed in quantity of millions cells *= quantity of bone marrow infiltration (0: not infiltrated, >0: infiltrated).**



Subject				
Variable	Data Type	Val Mag	Validity Range	UM
Age	Demographic	10^2	[1 - 999]	years
W	Morphometric	10^2	[1 - 999]	kg
H	Morphometric	10^2	[1 - 999]	cm
BMI	Morphometric	10^2	[0 - 99]	kg/m^2
BSA	Morphometric	10^1	[0 - 9]	m^2
Laboratory_Data				
Variable	Data Type	Val Mag	Validity Range	UM
LDH	Laboratory	10^3	[1 - 9999]	mg/dL
LDHMax	Laboratory	10^3	[1 - 9999]	mg/dL
PLTs	Laboratory	10^3	[1 - 9999]	10^9/L
WBC	Laboratory	10^2	[1 - 999]	10^9/L
N	Laboratory	10^2	[1 - 999]	10^9/L
L	Laboratory	10^2	[1 - 999]	10^9/L
Hb	Laboratory	10^2	[1 - 999]	g/dL
ALT	Laboratory	10^2	[1 - 999]	IU/L
AST	Laboratory	10^2	[1 - 999]	IU/L
GGT	Laboratory	10^2	[1 - 999]	IU/L
AlP	Laboratory	10^2	[1 - 999]	IU/L
Protein	Laboratory	10^1	[1 - 999]	g/dL
Alb	Laboratory	10^1	[0.1 - 999]	g/dL
B2	Laboratory	10^1	[0.1 - 999]	mg/dL
B2Max	Laboratory	10^1	[0.1 - 999]	mg/dL
Uricemy	Laboratory	10^1	[1 - 99]	mg/dL
IgG	Laboratory	10^0	[0.1 - 99]	g/dL
IgA	Laboratory	10^-1	[0.1 - 9]	g/dL
IgM	Laboratory	10^-1	[0.1 - 9]	g/dL
Bili	Laboratory	10^-1	[0.1 - 9]	mg/dL
Pathologic_Data				
Variable	Data Type	Val Mag	Validity Range	UM
flowBM	Pathological	-	[0.1 - 100]	%*
flowPB	Pathological	-	[0.1 - 100]	%*
KI67	Pathological	-	[1 - 100]	%*
KI67_Centr	Pathological	-	[1 - 100]	%*
BMInfperc	Pathological	-	[0 - 100] ***	%*
IgHOmo	Pathological	-	[0.1 - 100]	%*
CD1_Centr	Pathological	-	[1 - 100]	%*
CD5_Centr	Pathological	-	[1 - 100]	%*
CD20_Centr	Pathological	-	[1 - 100]	%*
Diag_Procedures				
Echo_muga_sca_lvef	Physiologic	-	[1 - 100]	%*
MRD_Baseline				
Variable	Data Type	Val Mag	Validity Range	UM
qPCR_BM	Laboratory MRD	-	[-0.000001 - 1]	**
qPCR_BM_Conv	Laboratory MRD	10^0	[0.01 - 9]	-
qPCR_PB	Laboratory MRD	-	[-0.000001 - 1]	**
qPCR_PB_Conv	Laboratory MRD	10^0	[0.01 - 9]	-

II Level Controls

Once that I Level data-cleansing was processed on features, II level controls were implemented. II level controls were based on cross controls between clinical and biologic variables which express information assuming common clinical hypothesis. Physicians and biologists' supervision was necessary to apply this monitoring strategy on data. Table 8 includes rules on MCL0208 cross-monitoring control.

Table 8: Cross controls between variables supervised by clinical expertise.

Rules	Variables
1	BMInf > 0 AND AAstage < IV
2	EN>0 AND AAstage < IV
3	AAstage < IV AND BMInf > 0 AND EN > 0
4	AAstage < IV AND BMInf > 0 AND (flowBM > 15% OR flowPB > 15%)
5	AAstage < IV AND BMInf > 0 AND ($qPCR_{BM} > 10^{-5}$ OR $qPCR_{PB} > 10^{-5}$)
6	AAstage < IV AND BMInf > 0 AND (flowBM > 15% OR flowPB > 15%) AND ($qPCR_{BM} > 10^{-5}$ OR $qPCR_{PB} > 10^{-5}$)
7	AAstage < IV AND L > 5 [10^9/L] AND BMInf > 0
8	PR or SD not allowed after CR

An interesting example of II Level control is rule number 6. Rule goal is to catch either incorrect AAstage or BMInf record on eCRF. Bone marrow infiltration of a mantle cell lymphoma can be detected via immunochemistry on bone marrow sample (BMInf) or via molecular biology advanced techniques (cytofluorimetry or quantitative PCR) (sub-section 3.1.3.1.). Rule's aim is to detect observations with AAstage minor than IV, with a bone marrow assessed as infiltrated via both standard and molecular available techniques. In case of cytofluorimetry, clinicians consider a bone marrow infiltrated with values over 15%, whereas in case of quantitative PCR, clinicians define a bone marrow infiltrated if at least 10^{-5} cells are malignant. Both flow cytometry and quantitative PCR techniques provide same analysis on peripheral blood (Ferrero et al. 2011). Moreover, rule n. 8 is a typical data-management issue: considering the i^{th} timepoint after pharmacotherapy administration, it is clinically unlikelihood that a patient has a both PR and SD at ($i^{th} + 1$) time point after that he had a CR at i^{th} time point, but possible for incorrect data-entry.

3.1.5 Data reporting

A strategy to assess the effect of remote monitoring on MCL0208 clinical trial

Once that data set was cleaned after the application of I and II Level controls, I2ECR environment was able to generate high-optimized datasets. In order to assess the effect of remote monitoring through I2ECR on MCL0208 clinical trial, a set of features have been chosen (table 9). The idea was to compare a quality level among the I2ECR output well-optimized dataset and three input datasets retrieved in 3 different time-points: 1st in late 2015, 2nd in middle 2016, 3rd in early 2017 (figure 10A).

Table 9: Features retrieved from I2ECR for quality control assessment.

Table	Variables
Laboratory Features	LDH, LDHMax, WBC, N, L, Hb, PLTs, B2, B2Max
Diagnostic Procedures Features	Sub_diaph Involvements, SupraDia Involvements, Extra-Nodal, PET
Pathology Features	Ki67, BMInf, BMInfperc, Hist
Clinical Features	AAstage, MIPIcrf, Bulky, ECOGps
MRD Features	flowBM, flowPB, $qPCR_{BM}$, $qPCR_{PB}$

4 classes of data-entry errors have been chosen to standardize the errors' count associated to each input dataset with respect to the reference (table 10): I level controls (M, N and R) were distinguished from II level controls (R). To assess the total effect of errors in data-entry 2 indexes have been constructed. Index_1 measures data quality for each feature, whereas Index_2 measures quality data-entry for each active center. To compare centers, mistakes detected from each subject belonging to same center were normalized on total number of subjects enrolled by that center times the total number of analyzed variables. Example in figure 10B.

$$\frac{total\ mistakes\ detected\ from\ i^{th} center}{total\ enrolled\ subjects\ by\ i^{th} center * number\ of\ analyzed\ variables}$$

Table 10: classes of errors established by I2ECR team.

Errors Encoding	Description
M	Missing Value
N	Nulls
R	Range errors
R	Crossing Errors (II Level controls)

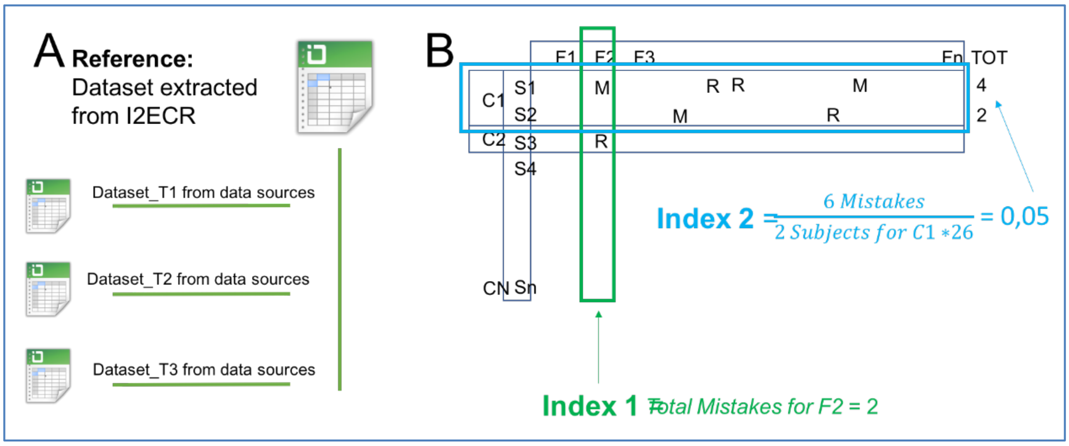

A
Reference:
Dataset extracted from I2ECR
Dataset_T1 from data sources
Dataset_T2 from data sources
Dataset_T3 from data sources
B
F1
F2
F3
Fn
TOT
C1
S1
S2
C2
S3
S4
CN
Sn
M
R
R
M
4
M
R
2
R
Index 2 = $\frac{6\ Mistakes}{2\ Subjects\ for\ C1 * 26}$ = 0,05
Index 1 Total Mistakes for F2 = 2

Figure 10: Methodology assumed for quality control in I2ECR. Figure A describes that 3 row datasets are compared to a reference dataset, which is the output dataset extracted from I2ECR. Figure B shows both indexes suited for datasets quality assessment.

Moreover, data quality effect was applied on aggregate MIPI clinical prognostic factor. Hoster et al. modeled a MIPI as mathematical sum of variables weighted by constants.

$$
\begin{gathered}
MIPI\ score = [0.03535 * age\ (years)] * \ age\ (years)] \\
+0.6978\ (if\ ECOG > 1) \\
+ \left[1.367 * log10\left(\frac{LDH}{LDHMax}\right)\right] \\
+ [0.09393 * log10(WBC\ count)]
\end{gathered}
$$

The effect of quality improvement on subjects' stratification by MIPI score has been related to PFS (Progression Free Survival) post-treatment outcome, updated at September 2017. PFS curves have been calculated from T1 to T3 to visually demonstrate data quality improvement. PFS curves were obtained from application of Kaplan-Meier model with a P value defined by Log-rank test.

3.1.6 Data transformation

Data transformation techniques transform or consolidate data into forms that are appropriate for DM processing. In I2ECR, data transformation methods used have been:

A. Normalization: because variables may present very large ranges, to allow a significant numerical comparison they have been scaled into similar intervals. In this case, min-max method has been applied.

B. Missing values imputation. Imputation methods are dived in statistical (i) and machine learning (ii) (García-Laencina, 2010):
   - Statistical Methods: Conditioned (mean/median methods) and no-conditioned (multiple imputation) methods (Horton and Lipsitz 2001).
   - Machine learning Methods: k-NN (k nearest neighborhood), SOM (Kohonen Self-Organized Maps), GA (Genetic Algorithms).

   Table 11 and ANNEX 2 show missing values (MV) count for a subgroup of features extracted from I2ECR. In I2ECR project, given a patient a MV for a specific input variable, a statistical conditioned imputation method has been chosen basing on the median assessed between subjects belonging to the same MIPI risk classes.

C. Discretization based on MIPI classification, where the raw values of a numeric attribute have been replaced by interval labels. The Chi-Merge algorithm (Kerber 1992) was chosen and implemented for discretization. It is a supervised and bottom-up method that discretizes each variable separately using the χ^2 statistics. It iteratively merges adjacent elements until the χ^2 value exceeds a defined threshold. "In this case, the threshold is determined as the χ^2 value for a significance level of 0.95 and a number of degrees of freedom equal to the number of MIPI classes minus one, that is 2"(Zaccaria, Rosati, et al. 2017).

D. Categorization. For I2ECR, continues values of a numeric attribute have been replaced by categorical values following three approaches:
   - Normality interval ranges defined by clinicians.
   - Normality interval on the normality maximum level defined by laboratories which processed biologic samples. Valid for B2 Macroglobulin's' and LDH values (ß2/ß2Max, LDH/LDHMax).
   - High vs low values defined on median assessed by data retrieved from study.

The list of intervals is described in table 11.

Table 11: List features extracted from DW, every feature is described in data type, Unit of Magnitude (UM), quantity of Missing Values (MVs) updated at Summer of 2017 and both normality range and cut-off thresholds defined by clinicians for categorization. For more details, see also ANNEX 2.

Variable	Data Type	UM	MV	Normality Range	Cut-off
Age	Demographic	years	0	-	Median
W	Morphometric	kg	0	-	Median
H	Morphometric	cm	0	-	Median
BMI	Morphometric	kg/m^2	0	-	25
BSA	Morphometric	m^2	0	-	Median
LDH	Laboratory	mg/dL	0		LDH/LDHMax
PLTs	Laboratory	10^9/L	1	[150 - 450]	-
WBC	Laboratory	10^9/L	0	-	Median
N	Laboratory	10^9/L	1	-	Median
L	Laboratory	10^9/L	2	-	Median
Hb	Laboratory	g/dL	0	[11.7 - 18]	-
ALT	Laboratory	IU/L	5	[7 - 56]	-
AST	Laboratory	IU/L	5	[10 - 40]	-
GGT	Laboratory	IU/L	20	[8 - 65]	-
AlP	Laboratory	IU/L	26	[44 - 147]	-
Protein	Laboratory	g/dL	18	[6 - 8.3]	-
Alb	Laboratory	g/dL	36	[3.4 - 5-4]	-
B2	Laboratory	mg/dL	51	-	B2/B2Max
IgG	Laboratory	g/dL	78	[0.7 -1.6]	-
IgA	Laboratory	g/dL	78	[0.07 - 0.4]	-
IgM	Laboratory	g/dL	77	[0.04 - 0.23]	-
Bili	Laboratory	mg/dL	19	[0.2 - 1.2]	-
flowBM	Pathological	%	48	-	Median
flowPB	Pathological	%	15	-	Median
KI67	Pathological	%	29	-	30
IgHOmo	Pathological	%	85	-	Median
qPCR_BM	Laboratory MRD	-	139	-	Median
qPCR_PB	Laboratory MRD	-	130	-	Median

3.1.7 Data-analysis projects

Feature Selection

Feature selection (or data reduction) aim is to remove noise effect on data (Han, Kamber, and Pei 2012) improving the performance of mining in terms of result comprehensibility (Fahrudin et al. 2017). In this project, the Quick-Reduct Algorithm (QRA) (Shen and Chouchoulas 2000) was used to select the most important features. It is a supervised tool based on the Rough Set Theory that allows for solving FS problems without generating all the possible subsets. QRA uses the *dependency degree* $\gamma_R(D)$ value to measure the importance of a given subset of input features R with respect to the class attribute D (MIPI class risk). The main idea of the algorithm is to iteratively add to the actual features subset those attributes producing the largest increase in the dependency degree.

- **Subset R1**: Age, LDH, WBC, PLTs, Hb, B2, Protein, Albumin, IgG, IgA, IgM, AST, ALT, GGT, ALP, Bilirubin, $qPCR_{BM}$, $qPCR_{PB}$, flowBM, flowPB, BMInfperc, IgHOmo (table 11).
- **Subset R2**: includes R1 features minus clinical variables determining MIPI (Age, LDH, WBC), as they could bias the final results.
- **Class D**: for each patient, the corresponding MIPI has been extracted. According to the MIPI (Hoster et al. 2008):
    - 182 subjects were classified as low risk subjects,
    - 73 intermediate risk subjects,
    - 43 high risk subjects.

A double approach has been followed:

A. QRA applied on subset R1 that includes variables composing MIPI value.
B. QRA applied on subset R2, which is R1 subset without MIPI independent composing features.

Therefore, rough dataset contained 298 subjects characterized by 22 (or 18) input continues variables and one class variable (MIPI value). For this project, defined pipeline of data processing followed those steps:

I. Data cleaning via I2ECR (rough dataset).
II. Data transformation following A, B and C methods (section 3.1.6).
III. Application of QRA algorithm on both subsets R1 and R2.
IV. Methodology validation.

Selected features of R2 subset were related to post-treatment outcomes in order to stratify subjects in novel survival classes. To do this, curves were derived from application of Kaplan-Meier model with a P value defined by Log-rank test.

Moreover, to validate the pipeline capability of the datasets obtained after each step to correctly classify the subjects has been assessed. Therefore, results were compared to classification accuracy reached by the initial raw dataset. Hierarchical classification (K-nearest neighbor) to measure the quality variation among datasets has been applied. K values from 3 to 10 and different distance metrics (Euclidean, the Chebyshev and the City-block distances) have been tested. K=7 and the City-block distance better performed. The leave-one-out validation has been employed to assess the classification performances.

DELPHI

DELPHI is a statistical univariate project for Data Elaboration to Predict Hypothetical AssocIations. The goal of this project is to discover novel putative associations between baseline variables of MCL0208 clinical study. Features included have been clinical, laboratory, pathological, mutational and GEP (Gene Expression Profiling). A team of 3 lymphoma experts was involved for assessing the expected associations between 62 variables extracted from baseline of MCL0208 study via I2ECR. Each clinician underlined each possible couple between features (e.g. association between ALT and ALP as well as between Ki-67 and MIPIc are expected (Hoster et al. 2016)). Contributions from experts were merged to extrapolate common associations. DELPHI automatically found the expected association following a statistical approach**:**

1. Continues features were categorized to allow a statistical comparison between categorical and continues variables (basing on D transformation point from subsection 3.1.6).
2. Patients were divided in a discovery-set (200 subjects from center 10 to 52) and a validation set (100 subjects from center 1 to 9). The split strategy was applied balancing the distribution of missing data between features analyzed.
3. Not significant categorical couples (p>0,05) were discarded using a χ^2 and a Fisher exact test as appropriate. Table 12 lists for each qualitative feature the statistical test chosen to screen significant associations. χ^2 test has been applied to independent categorical couples of variables (more than two

categories) as well as Fisher exact test has been applied to independent nominal couples (Chan 2003).

4. Strength of Associations via Cramer's V coefficient was used to assess the strength of associations of significant couples (Bergsma 2013).

Table 12: Statistical tests applied to DELPHI dataset (p<0.05).

Test	Variables
χ^2	ECOGps, MIPISt, MIPISim, MIPIb, MIPICSt, MIPICSim, SOX11, AAstage, ABCB-1236_C>T, ABCB-2677_C>T, ABCB-3435_C>T, Aplotype_ABCB1; VEGFA-2055_A>C; ABCG2-421_C>A; FCGR2A-497_A>G; NCF4-368_G>A.
Fisher	Gender, Age, W, H, BSA, BMI, LDH, WBC, N, L, Hb, PLTs, ALT, AST, Creatinine, Protein, Alb, Bili, GGT, ALP, B2, IgG, IgA, IgM, Ki67, Histology, BMInf, NesPCR_BM_IGH_dia, NesPCR_BM_IGH_dia, NesPCR_BM_BCL1_dia, NesPCR_PB_BCL1_dia, flowBM, flowPB, qPCRBM, qPCRPB, IgHOmo_High, Sym, Bulky, Supra_Inv, Sub_Inv, EN_Inv, NLTB, ENLTB, PET, ATM, P53, WHSC1, CCND1, KMT2D, NOTCH1, BIRC3, TRAF2, CXCR4, GEP.

Couples from discovery set confirmed by validation-set ha selected. Therefore, selected couples were overlapped to the previously declared "clinical expectations" to extrapolate the unexpected ones, suggested by the analysis (figure 11). Data visualization was performed via Circos (Krzywinski 2009).

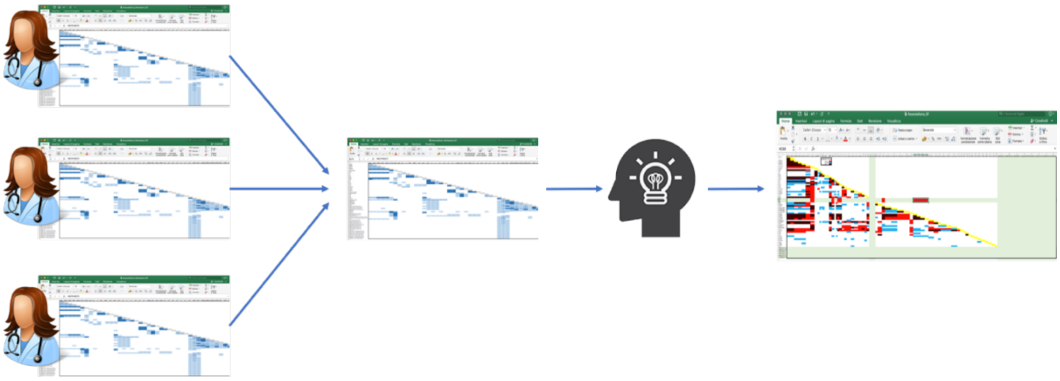


Figure 11: DELPHI methodology. 3 lymphoma experts expcted associations between 62 variables extracted from MCL0208 baseline. Sheets obtained were merged and compared to DELPHI analysis.

3.2 The I2ECR environment design

The 2nd part of the chapter describes all phases composing I2ECR environment design and implementation. The main goal of designing the I2ECR system has been to provide clinical researchers a tool for managing all activities described in previous sections. Clinical research involves several stakeholders, characterized by different skills. Among these:

- **Principal Investigator (PI)** is the scientific and legal responsible of the research.
- **Clinical trialists** are physicians who follows enrolled patients during their entire course.
- **Clinical assistants** are clerks involved in clinical trials management, they are in charge of trials authorization by hospital ethical and steering committee.
- **Data-Managers** are responsible of data integrity and coordinate data flux between centers who enroll subjects and central-hubs.
- **Biostatisticians** are responsible of data analysis.
- **Biologist** is responsible of molecular biology data collection, storage and delivery to central hubs.

Figure 12 depicts a general workflow defining macro-activities of I2ECR environment (Aalst and van Hee 2004). Once that a DW has been designed (and already implemented in this PhD experience basing on MCL0208 clinical trial), **I2ECR system provides to:**

- Setup a DW basing on requirements defined by clinical trial protocol: target number of subjects to be enrolled (i), number of time-points (ii) and data constraints (iii).
- Encode cleaning requirements from whom saved data in a DW can be automatically cleaned up: definition of protocols to set standardized quality controls (I and II Level of controls) on data as described in Section 3.1.4.
- Load a dataset on a DW and update them whenever necessary.
- Semi-automatically generate queries to send to centers active in trial enrollment in order to catch up with a correct data entry.

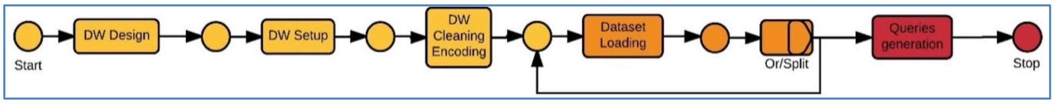


Figure 12: I2ECR general workflow

A preliminary activity analysis has been provided through swim lane business process management diagram (Jun et al. 2009). Figure 13 shows the “DW design” process swim lane. This diagram underlines the business interactions between involved actors: Biomedical Engineer, Principal Investigator (PI) and DBMS (Database Management System). “DW Setup”, “DW Cleaning Encoding”, “Dataset Loading” and “Queries generation” are included in ANNEX 3.

3.2.1 The I2ECR software

I2ECR software was designed via UML® (Unified Modeling Languages[21]). Interactions between system and users have been modelled via Use Case Diagram (figure 14).

[21] http://www.uml-diagrams.org/

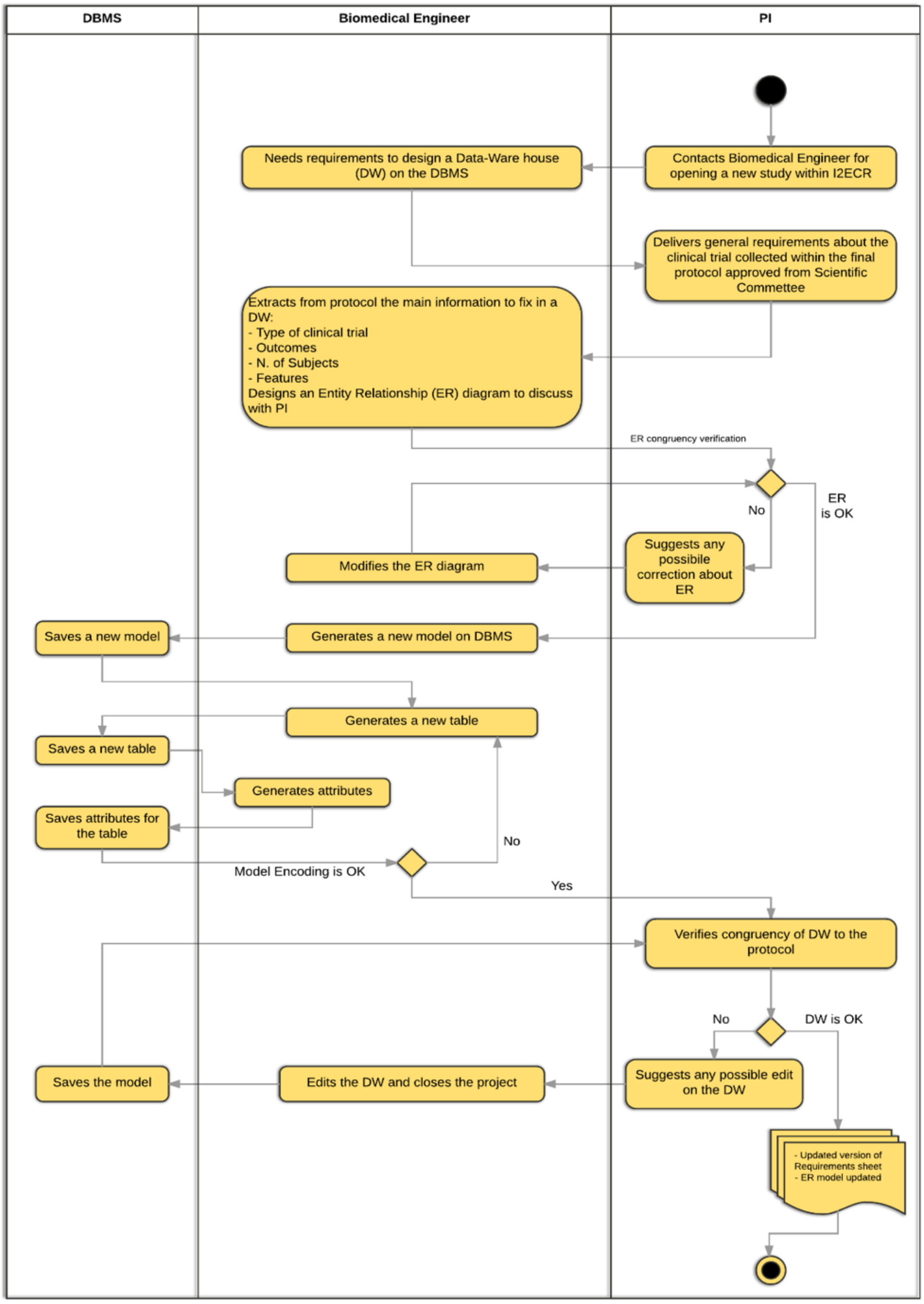

DBMS
Biomedical Engineer
PI
Needs requirements to design a Data-Ware house (DW) on the DBMS
Contacts Biomedical Engineer for opening a new study within I2ECR
Delivers general requirements about the clinical trial collected within the final protocol approved from Scientific Commettee
Extracts from protocol the main information to fix in a DW:
- Type of clinical trial
- Outcomes
- N. of Subjects
- Features
Designs an Entity Relationship (ER) diagram to discuss with PI
ER congruency verification
No
ER is OK
Modifies the ER diagram
Suggests any possibile correction about ER
Saves a new model
Generates a new model on DBMS
Generates a new table
Saves a new table
Generates attributes
Saves attributes for the table
No
Model Encoding is OK
Yes
Verifies congruency of DW to the protocol
No
DW is OK
Saves the model
Edits the DW and closes the project
Suggests any possible edit on the DW
- Updated version of Requirements sheet
- ER model updated

Figure 13:Swim-lane of “DW_design” activity

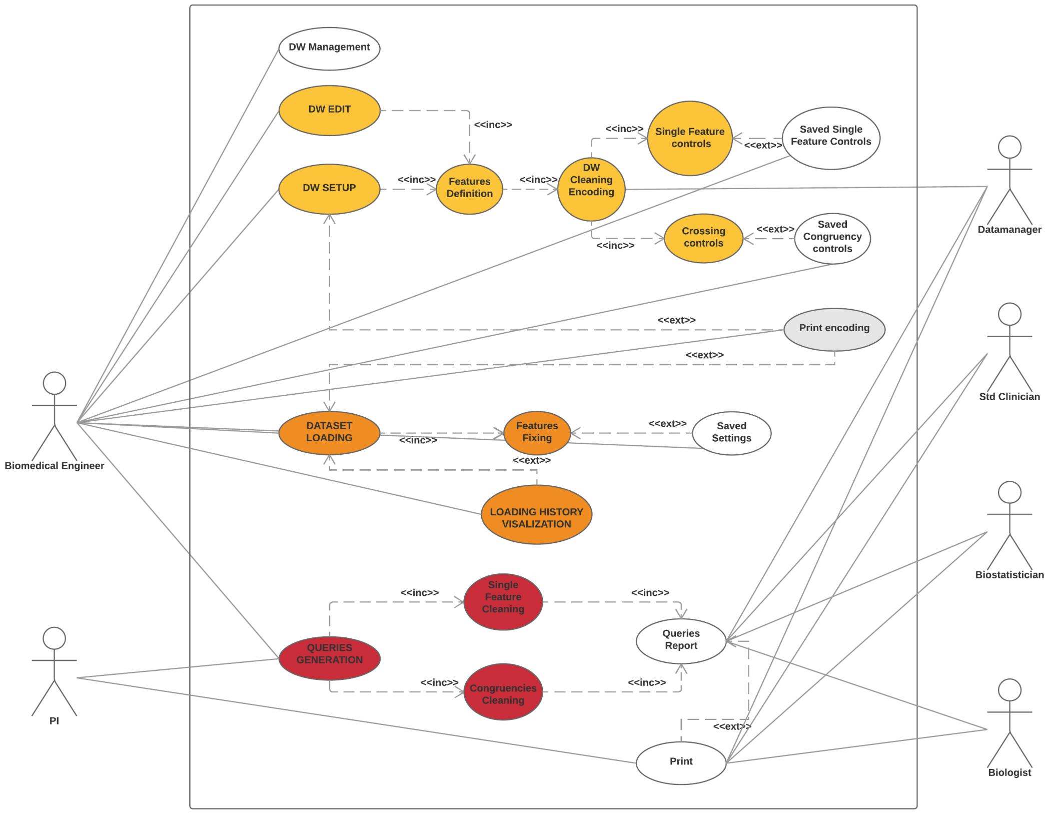


Figure 14: I2ECR Use Case Diagram.

Users

I2ECR users are listed in table 13. A “Key user” is the user that has high interaction with the system. Although a “secondary user” interacts with a system, he accesses to limited use cases.

- Biomedical Engineer: he is the I2ECR key user. His skills are to collect requirements from clinical researchers and translate them to uses cases. He has high technology experience due to his professional background.
- Principal Investigator: he is the scientific and legal responsible of clinical trial. He approves the final protocol to be authorized from hospital’s ethical and scientific committees. He is a key user as he has high knowledge of the clinical trial. Despite of his clinical experience, he is not expert in technology.
- Standard Clinician: He accesses to I2ECR reports in order to evaluate data cleaning status of clinical trials. He generally has a medium technology experience.

- Data-Manger: He has the duty to interface to centers active in clinical trials for data monitoring. Therefore, he accesses to I2ECR reports for data reporting to deliver to satellite centers. He generally has a medium technology experience.
- Biostatistician: He accesses to I2ECR to extract reports in order to evaluate data cleaning status. He generally has a medium technology experience.
- Biologist: He accesses to I2ECR to extract reports in order to evaluate data cleaning status on molecular and biologic data. He generally has a medium technology experience.

Table 13: I2ECR's users

User	Technology Experience	User Priority
Biomedical Engineer	High	Key User
Principal Investigator	Low/Medium	Key User
Standard Clinician	Low/Medium	Secondary User
Data Manager	Medium	Secondary User
Biostatistician	Medium	Secondary User
Biologist	Medium	Secondary User

Use cases

Table 14 lists use cases designed for I2ECR software. However, figure 15 shows a use case detail of use case 10 “Dataset Loading”. For this use case, user loads a rough dataset (as introduced in paragraph 3.1.2). Moreover, he defines a description of data included in dataset, such as the type (baseline, outcome). Once that dataset has been loaded, I2ECR analyzes subjects format in reference of what attended by clinical trial DW requirements. For MCL0208, subjects are encoded in a 4 digits format. Hence, I2ECR detects unconventional formats, for instance if the dataset includes a subject’s ID which includes patient initials (GZ_1234). Finally, I2ECR requires to user to associate features to each column detected from input file. In this way, system is able to fill data into the table from DW (in our case FIL_MCL0208). Most relevant detailed use cases are collected in ANNEX 3.

Table 14: I2ECR’s Use Cases

N.	Use-Case	Actor/s
1	DW Setup	Biomedical Engineer
2	DW Edit	Biomedical Engineer
3	Features Definition	Biomedical Engineer
4	DW Cleaning Encoding	Biomedical Engineer, Data-Manager
5	Single features controls	Biomedical Engineer
6	Crossing controls	Biomedical Engineer
7	Saved single features controls	Biomedical Engineer
8	Saved Crossing controls	Biomedical Engineer
9	Print Encoding	Biomedical Engineer
10	Dataset Loading	Biomedical Engineer
11	Feature Fixing	Biomedical Engineer
12	Saved Settings	Biomedical Engineer
13	Loading history visualization	Biomedical Engineer
14	Queries generation	Biomedical Engineer, PI
15	Single features cleaning	Biomedical Engineer, PI
16	Congruencies (crossing) cleaning	Biomedical Engineer, PI
17	Queries Report	Biomedical Engineer, PI, Data-Manager, Std Clinician, Biostatistician, Biologist
18	Print	Biomedical Engineer, PI, Data-Manager, Std Clinician, Biostatistician, Biologist

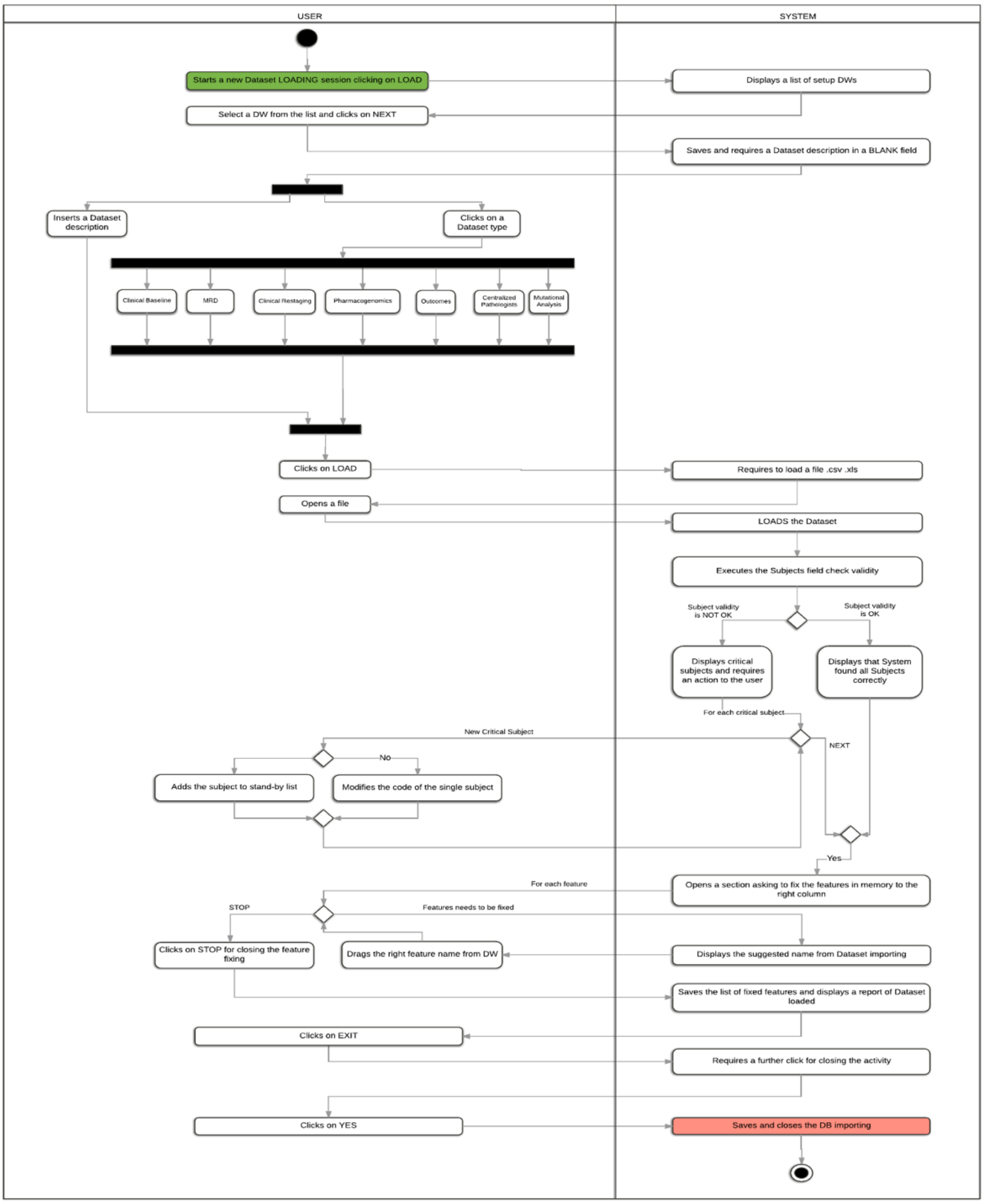

USER
SYSTEM
Starts a new Dataset LOADING session clicking on LOAD
Displays a list of setup DWs
Select a DW from the list and clicks on NEXT
Saves and requires a Dataset description in a BLANK field
Inserts a Dataset description
Clicks on a Dataset type
Clinical Baseline
MRD
Clinical Restaging
Pharmacogenomics
Outcomes
Centralized Pathologists
Mutational Analysis
Clicks on LOAD
Requires to load a file .csv .xls
Opens a file
LOADS the Dataset
Executes the Subjects field check validity
Subject validity is NOT OK
Subject validity is OK
Displays critical subjects and requires an action to the user
Displays that System found all Subjects correctly
For each critical subject
New Critical Subject
NEXT
No
Adds the subject to stand-by list
Modifies the code of the single subject
Yes
For each feature
Opens a section asking to fix the features in memory to the right column
STOP
Features needs to be fixed
Clicks on STOP for closing the feature fixing
Drags the right feature name from DW
Displays the suggested name from Dataset importing
Saves the list of fixed features and displays a report of Dataset loaded
Clicks on EXIT
Requires a further click for closing the activity
Clicks on YES
Saves and closes the DB importing

Figure 15: Use Case detail n.10 – Swim-Lane BPM diagram.

I2ECR architecture

I2ECR software (SW) is a web-based application for clinical research. A clinical trial involves stakeholders from different centers. Satellite centers, molecular laboratories and hub centers are often placed on the national or communitarian territory. Easily reachable platforms via the internet are necessary to rapidly connect scientists. The software has been developed via XAMPP® by Apache. XAMPP includes MySQL® and PHP language. Mock-up Graphic User Interfaces (GUIs) have been designed via Lucidart®, a web-browser extension. Whereas, SW GUIs have been implemented in PHP and JavaScript. Used coding editor has been Eclipse® with PHP extension mounted. PHP and JavaScript programming language has been chosen for following reasons:

- PHP allows to develop server-side applications as well JavaScript is optimized to develop user-side applications.
- PHP is designed to connect with a DBMS to rapidly interrogate databases and retrieve data (ETL).
- PHP allows to implement easy to use and ergonomic GUIs because its high-level integration with markup (HTML[22] files) as well as format (CSS[23] files) levels (Daniele Bochiccio 2015). PHP, JavaScript, HTML and CSS are established as "Public Domain" languages by World Wide Web Consortium (W3C).

Even if MATLAB® environment provides higher data-elaboration power, PHP and JavaScript allow to implement user-oriented web-browser apps. I2ECR software environment is implemented on 4 databases (DBs) (figure 16):

- Clinical Trials DW: in this case FIL_MCL0208 (figure 9). The number of encoded DWs is equal to all clinical trials included in I2ECR project.
- I2ECR_protocol_setup (figure 17): this is the database that collects all protocols encoded. There is a **1:1** cardinality between a protocol and a clinical trial DW. This DB includes information on protocol:
  - Name, description, number and name of time points associated (through <u>protocol_timepoints</u> table), number of enrolled subjects and number of active centers.

[22] https://it.wikipedia.org/wiki/HTML

[23] https://it.wikipedia.org/wiki/CSS

- I2ECR_dataset_setup (figure 18) collects all input dataset loaded into I2ECR. Different datasets must be associated to one protocol (cardinality **0:many**). Once that a dataset is associated to a protocol, included data are imported into the clinical trial DW (figure 9). This DB is designed to manage history of data loading as well to track critical subjects' management.
    - <u>datasets</u>: dataset table collects loading. Every dataset is associated to a name, a description and a timestamp of loading to track the data entry flow.
    - <u>dataset_standby_subjects</u>: it collects all subjects' ID suggested as unconventional by I2ECR. The user moves those subjects manually.
    - <u>dataset_adjusted_subjects</u>: it collects all subjects 'ID that I2ECR suggests to user for fixing ID format. Adjusted subject are externally identified by a correspondent stand-by subject (cardinality **1:1**).
- I2ECR_feature_encoding (figure 19): this DB collects all settings from whom to set data cleaning. PHP queries this DB to launch 1st and 2nd level controls (both introduced in paragraph 3.1.4). In detail, <u>Features</u> table included data information: units of magnitude (<u>Unit</u>, <u>Magnitude</u>), encoding about missing
- values (<u>Missing</u>) and validity and normality ranges (<u>ValidityRange_Min, ValidityRange_Max, NormalityRange_Min, NormalityRange_Max</u>).

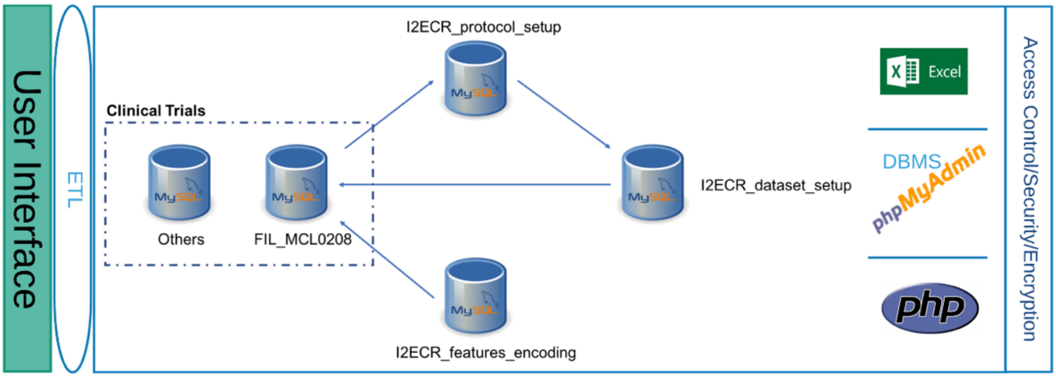


Figure 16: I2ECR architecture environment

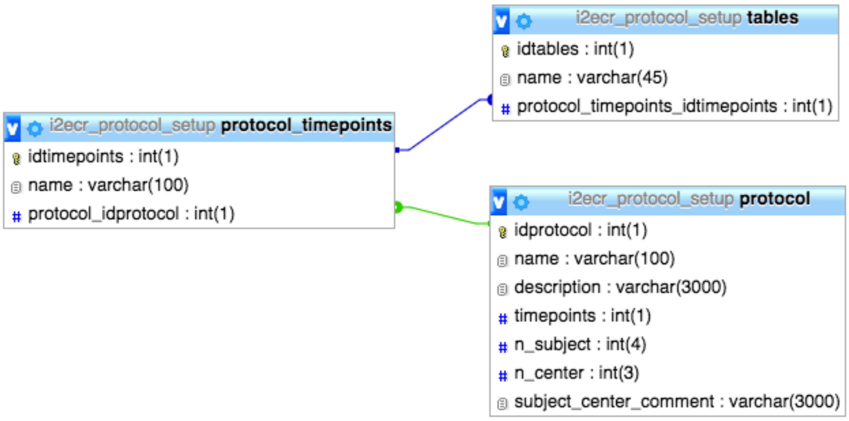

i2ecr_protocol_setup tables
idtables : int(1)
name : varchar(45)
protocol_timepoints_idtimepoints : int(1)
i2ecr_protocol_setup protocol_timepoints
idtimepoints : int(1)
name : varchar(100)
protocol_idprotocol : int(1)
i2ecr_protocol_setup protocol
idprotocol : int(1)
name : varchar(100)
description : varchar(3000)
timepoints : int(1)
n_subject : int(4)
n_center : int(3)
subject_center_comment : varchar(3000)

Figure 17: I2ECR_protocol_setup DB

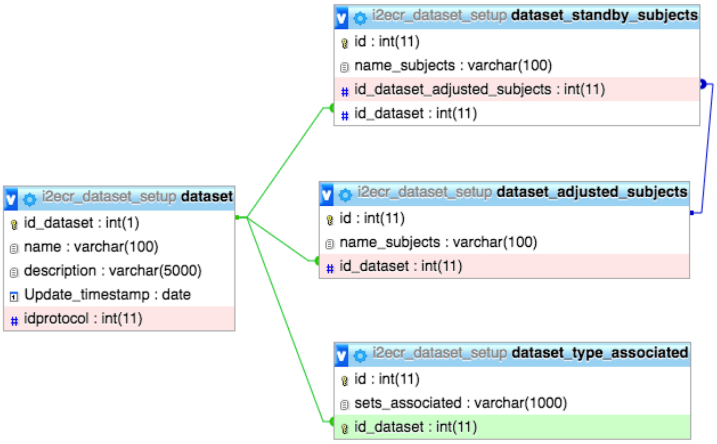
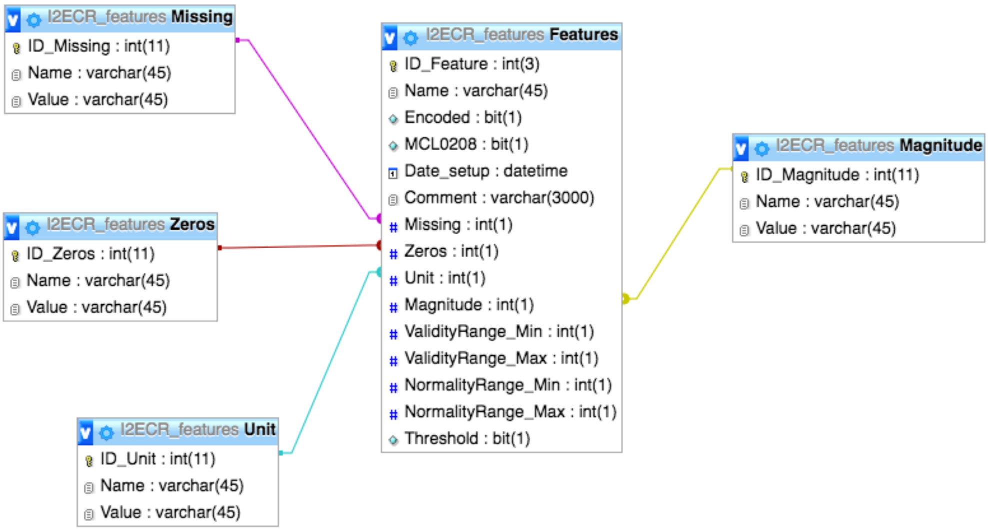

i2ecr_dataset_setup dataset_standby_subjects
id : int(11)
name_subjects : varchar(100)
id_dataset_adjusted_subjects : int(11)
id_dataset : int(11)
i2ecr_dataset_setup dataset
id_dataset : int(1)
name : varchar(100)
description : varchar(5000)
Update_timestamp : date
idprotocol : int(11)
i2ecr_dataset_setup dataset_adjusted_subjects
id : int(11)
name_subjects : varchar(100)
id_dataset : int(11)
i2ecr_dataset_setup dataset_type_associated
id : int(11)
sets_associated : varchar(1000)
id_dataset : int(11)

Figure 18: I2ECR_dataset_setup DB



I2ECR_features **Missing**
ID_Missing : int(11)
Name : varchar(45)
Value : varchar(45)

I2ECR_features **Zeros**
ID_Zeros : int(11)
Name : varchar(45)
Value : varchar(45)

I2ECR_features **Unit**
ID_Unit : int(11)
Name : varchar(45)
Value : varchar(45)

I2ECR_features **Features**
ID_Feature : int(3)
Name : varchar(45)
Encoded : bit(1)
MCL0208 : bit(1)
Date_setup : datetime
Comment : varchar(3000)
Missing : int(1)
Zeros : int(1)
Unit : int(1)
Magnitude : int(1)
ValidityRange_Min : int(1)
ValidityRange_Max : int(1)
NormalityRange_Min : int(1)
NormalityRange_Max : int(1)
Threshold : bit(1)

I2ECR_features **Magnitude**
ID_Magnitude : int(11)
Name : varchar(45)
Value : varchar(45)

Figure 19: I2ECR_feature_encoding DB

Chapter 4

Results & Discussions

4.1 I2ECR project

I2ECR is an **Integrated and Intelligent Environment for Clinical Research** where clinical and omics data stand together for clinical use and for generation of new clinical knowledge.

I2ECR idea is structured on Data Ware-house (DW) implementation. DWs ensure high level integration of translational data with the centrality of subject ID (Han, 2012). However, application of data “warehousing” concepts on medical informatics is still uncommon (Zapletal, 2010). Focusing on clinical research, there is no evidence of data where-housing usage in clinical trials sponsored by no-profit organizations.

I2ECR is adapted to MCL0208 phase III trial, which is a translational trial with several clinical prognostic factors associated to treatment data, biological assessment of disease and ancillary studies as Pathology, Mutational Analysis and GEP (Gene Expression Profile). Chapter 4 can be dived in 3 sections:

- Section 4.2: data warehousing in I2ECR, description of the architecture.
- Section 4.3: results derived from data cleaning and data analysis sub-projects on the wide pool of data of MCL0208 clinical trial retrieved from FIL_MCL0208 data ware-house (DW).
- Section 4.4: results in terms of implementation of the I2ECR software environment.

I2ECR project was developed on several steps (figure 6). Some of those steps produced several results which explain the novelty of using I2ECR approach in managing a clinical trial. First of all, effects of implementing a DW were assessed in terms of data cleaning (figure 20A). Data-management of a multicenter phase III clinical trial may be tricky if not supported by a central monitoring. Moreover, MCL0208 data-management was cumbersome because this study collects several ancillary studies as a centralized pathology review, a mutational analysis, pharmacogenomics and GEP. Every study implicates enormous quantity of rough data sheets to be integrated with clinical and molecular observations recorded in eCRF. Has been estimated that MCL0208 baseline data amount reached the order of 10^5 (350 features times 300 enrolled subjects). This huge pool of data increased up to 8 times if treatment and post-treatment data were involved. If a DW is associated to a DBMS, data extraction is fast. This implicates clinical researchers to easily extract reports as (i) "to query" mistakes on data due to unconventional data-entry by centers as (ii) to allow data-driven discovery strategies (figure 20B). A typical report of queries sent to each center is shown in ANNEX 4.

Furthermore, if data-driven discovery strategies are mixed up to clinical knowledge, I2ECR can be a powerful tool that allows to overcome human limitation in data elaboration, despite if actors have a big clinical expertise. In this PhD experience, two data analysis projects (figure 20C) have been proposed to evidence potential of I2ECR in clinical research: a feature selection project and DELPHI (Data ELaboration to Predict Hypothetical assocIations) project, both adapted on MCL0208 baseline variables. Data analysis project results are described in section 4.3.3.

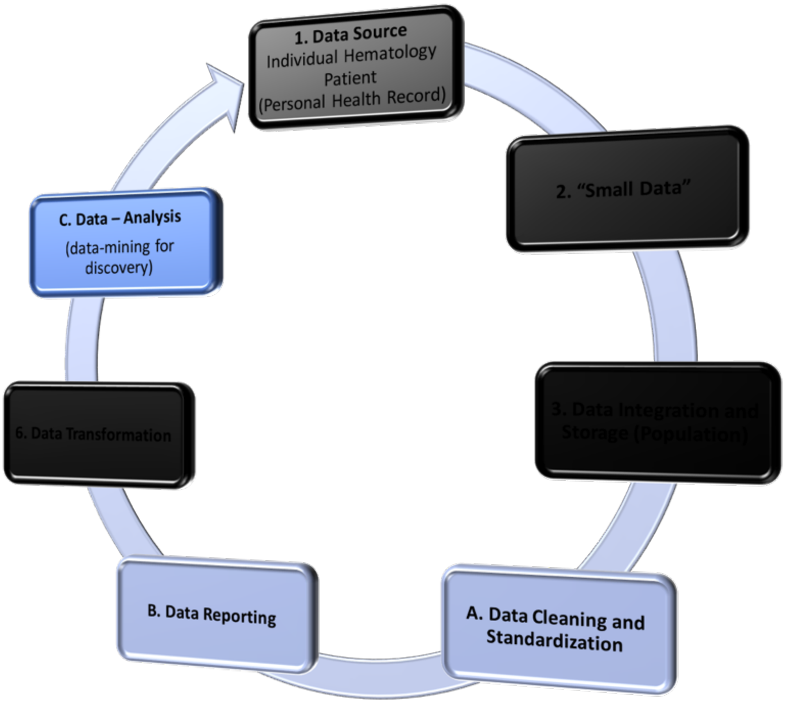

1. Data Source
Individual Hematology Patient
(Personal Health Record)
2. "Small Data"
3. Data Integration and Storage (Population)
A. Data Cleaning and Standardization
B. Data Reporting
6. Data Transformation
C. Data – Analysis
(data-mining for discovery)

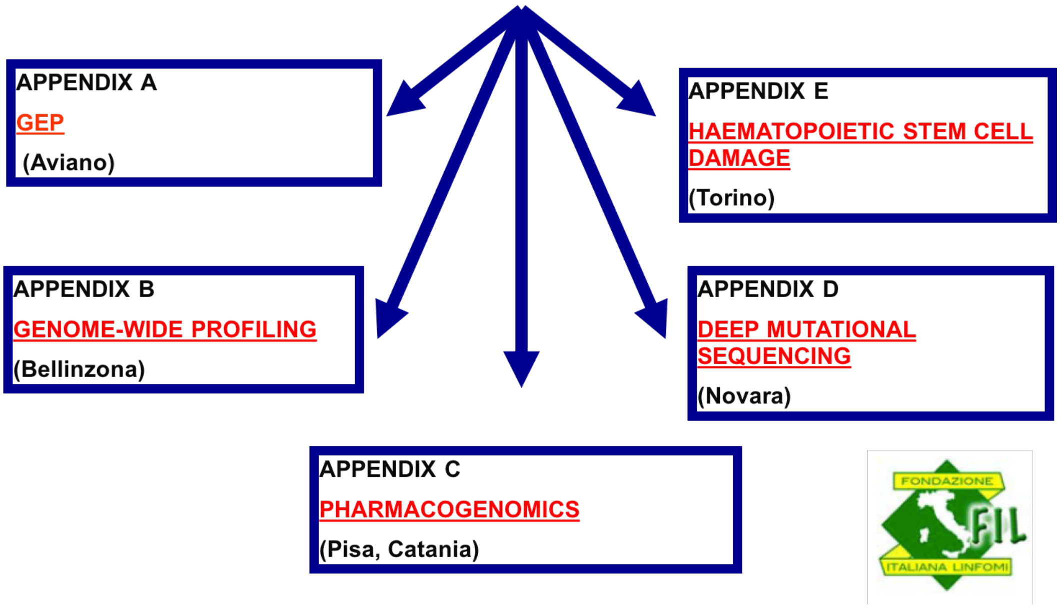
Figure 20: I2ECR pipe-line for data management. Focusing on results, I2ER allows to provide effect of both data quality improvement and data elaboration from steps A – data cleansing, B – data reporting and C – statistical analysis and data mining methods on clinical data.

4.2 FIL_MCL0208: a dynamic architecture.

MCL0208 clinical trial is a multicenter phase III study where clinical practice and molecular biology techniques are combined to evaluate survival from 1st line high-dose therapy up to post stem cells transplantation maintenance therapy (Revlimid®) for mantle cell lymphoma young patients (<60 years aged). The study workflow (figure 4) is composed by 3 post-diagnosis restaging during high dose therapy followed by 5 follow-up post maintenance. Therefore, ancillary laboratory studies were designed to assess secondary clinical trial outcomes. Figure 21 shows all ancillary studies of MCL0208. Among these, appendix A (Gene Expression Profile), appendix C (Pharmacogenomics) and appendix D (Deep Mutational Sequencing) have already been included in FIL_MCL0208. Both genome-wide profiling (appendix B) and hematopoietic stem cell damage (appendix E) databases inclusion is ongoing. This fact indicates that in clinical research, several years are needed to achieve primary and secondary clinical outcomes: a clinical trial life-cycle depends from drug development and, for a phase III study, duration may reach 4 years[24]. Consequently, for a DW applied to a clinical study we need at least the entire duration of study for data completeness, not considering delays due by legal amendments released by ethical steering committee.

However, I2ECR data warehouse modelling uses both top-down (TD) and as a bottom-up (BU) approach. This is the case of FIL_MCL0208 DW. At first, FIL_MCL0208 construction followed a top-down modeling: this approach has consisted in a general overview of the problem despite of integration quality and flexibility. With a practical vision, clinical team adopted this strategy facing with two main databases:

- one extracted from eCRF (Epiclin from figure 7) collecting both clinical, laboratory and treatment data.
- One extracted from molecular laboratory including both internal pathologic and Minimal Residual Disease (MRD) data.



[24] https://www.fda.gov/ForPatients/Approvals/Drugs/ucm405622.htm#Clinical_Research_Phase_Studies

Figure 21: MCL0208 ancillary studies. Currently, Genome-Wide profiling is the last excel database added to FIL_MCL0208. Figure courtesy of FIL (Fondazione Italiana Linfomi).

On the other hand, a bottom-up modeling has been used dealing with databases derived from both ancillary studies and central pathology reviewers. In that way, data warehouse development of independent data-sheets "provides flexibility, low-cost and rapid return of investment" (Han, Kamber, and Pei 2012).

Table 15 lists databases imported to I2ECR for MCL0208 study quantifying both number of import for the same database (equal to the number of updates for data included) and frequency during a 1 year. An interesting data is provided by total amount of imports computed to I2ECR via ETL (Extraction, Transformation, Loading) layer. In fact, 23 datasheets have been imported into I2ECR with high-level of complexity within each data-sheet as well as different versions of the same data-sheet. Hence, last column of table 15 describes this level of complexity, taking in account variability (type of features included) and sample-size (both total number of subjects and features included in data-sheet). Complexity levels have been categorized in 4 general classes:

- L – Low: low sample size and low variability and quantity between characteristics.
- M - Medium: medium/high sample size, medium variability and quantity of characteristics.
- H – High: medium/high sample/size, medium variability and high quantity of characteristics.
- VH – Very High: medium/high sample/size, high variability and medium/high quantity of characteristics.

Table 15: pool of datasheets used to both model and populate FIL_MCL0208 DW for I2ECR. For each input dataset the modeling strategy, total number and frequency of imports are listed. Last column defines level of complexity for each single data-sheet.

DB name	**Model Strategy**	**N. of import**	**Freq. per y**	**Complexity**
eCRFs_dataset	TD	8	4	H
MRD_lab_dataset	TD	4	2	VH
GEP_dataset	BU	3	1.5	L
mutational_dataset	BU	2	1	M
pathology_dataset	BU	1	0.5	VH
Pharmacogenomics_dataset	BU	4	2	M
genome_wide_profiling_dataset	BU	1	0.5	M
TOT	-	**23**	-	

4.3 Data Cleaning, Reporting and Analysis

4.3.1 Data cleaning

I Level controls

I Level controls are defined in section 3.1.4. In table 16 are listed most-relevant baseline features. Features are extracted from FIL_MCL0208 during I2ECR project:

- LDH/LDHMax, WBC, Hb, PLTs, B2/B2Max from Lab_Data table at N_timepoint=0 (baseline).
- ECOGps, AAstage, Bulky, Hist and Sym from Clinical_data_Baseline table.
- Age from Subjects table.
- BMInf, flowBM, flowPB and Ki67 from Pathologic_Data table at N_timepoint=0 (baseline).

Quantities shown are the total of data-recovered since early 2016 conveniently classified in Missing values, Null not allowed and Ranging mistakes.

Table 16: TOT mistakes recovered since Early 2016 during I2ECR project for a subgroup of most-relevant features.

Features	Mistakes		
	Classes	N	TOT
LDH/LDHMax	Missing Values	21	25
	Null	3	
	Ranging	1	
WBC	Missing Values	8	49
	Null	0	
	Ranging	41	
ECOGps	Missing Values	1	2
	Null	1	
	Ranging	0	
Age	Missing Values	0	0
	Null	0	
	Ranging	0	
Hb	Missing Values	8	8
	Null	0	
	Ranging	0	
PLTs	Missing Values	7	16
	Null	0	
	Ranging	9	
B2/B2Max	Missing Values	8	18
	Null	0	
	Ranging	10	
Ki67	Missing Values	16	38
	Null	22	
	Ranging	0	
BMInf	Missing Values	8	9
	Null	1	
	Ranging	0	
flowBM	Missing Values	133	135
	Null	2	
	Ranging	0	
flowPB	Missing Values	154	154
	Null	0	
	Ranging	0	
AAStage	Missing Values	0	0
	Null	0	
	Ranging	0	
Bulky	Missing Values	0	1
	Null	1	
	Ranging	0	
Sym	Missing Values	0	0
	Null	0	
	Ranging	0	
Hist	Missing Values	8	15
	Null	7	
	Ranging	0	
All Features	**Missing Value**	**372**	**470**
	Null	**37**	
	Ranging	**61**	

I2ECR allowed to dramatically decrease mistakes in data-entry from starting of the project (figure 22). Values have been measured in reference of the start of the project (in early 2016). In more detail, flowBM e flowPB (Bone Marrow and Peripheral Blood infiltration via flow cytofluorimetry), both retrieved from local laboratory sheets and later imported in FIL_MCL0208 DW, record highest quantity of correct observations (135, 154). Mistakes from both laboratory baseline features LDH/LDHMax, WBC from eCRFs detected a drop-off (25, 49). Moreover, table 16 shows a classification of mistakes (Missing Values, Null and Range mistakes). Observing the classification, Ki67 values (retrieved from pathology table) show a big classification within Null mistakes (22/38) whereas WBC values were affected by incorrect ranging fill-in (41/49). Table 16 final row indicates that Missing Values (MV) are the big deal in data-entry: in fact, MV represents about 80% of total mistakes recorded (372/470). Finally, Hist (histology), AAstage (Ann Arbor stage), Sym (symptoms), Age and ECOGps (ECOG performance status) features are represented by few mistakes.

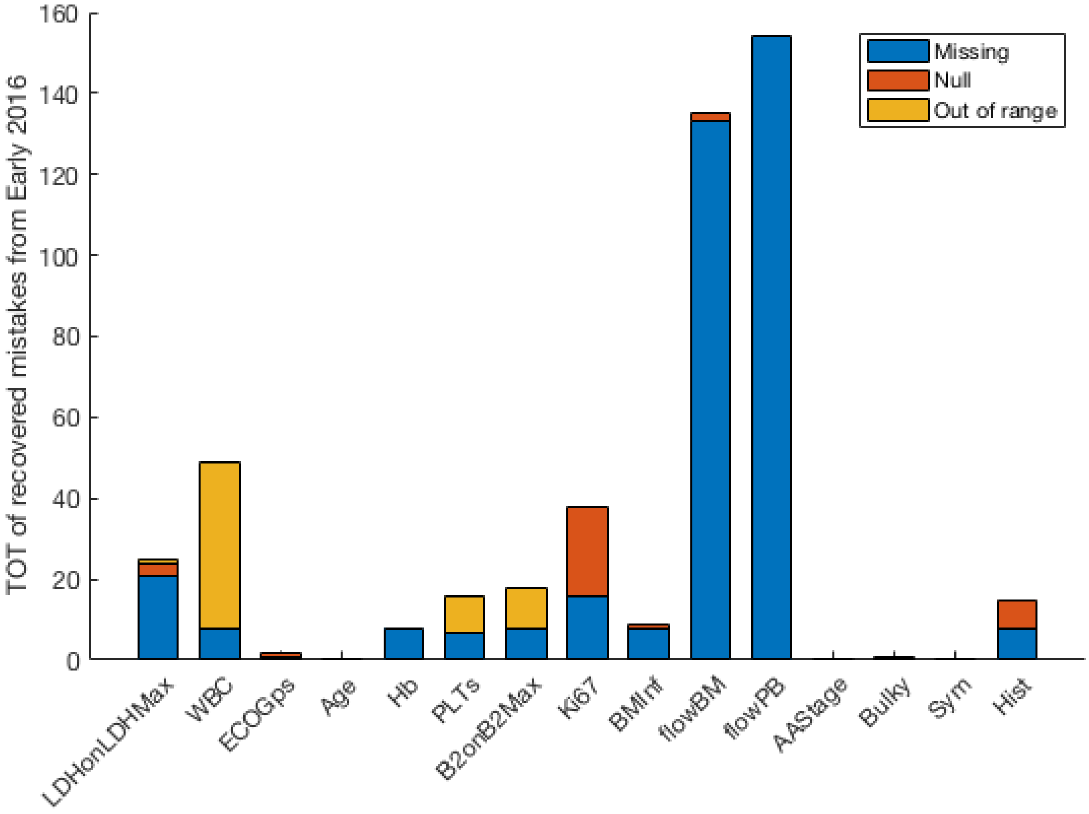


Figure 22: Number of corrections via I2ECR for MCL0208 clinical trial for a subset of selected features.

II Level controls

Table 8 from section 3.1.4 lists cross controls assumed as I2ECR II Level controls. Table 17 shows subjects obtained from the applications of those controls. Observing rule n. 1, 11 subjects with incongruent data entry are listed. Qualitatively, several observations are commonly reported from the application of different rules: subjects 522, 1508 are both reported in columns 1 and 5; subjects 608, 609 and 1607 are listed in columns 1 and 6; 3504, 3702 and 3705 observations are reported in columns 3 and 5.

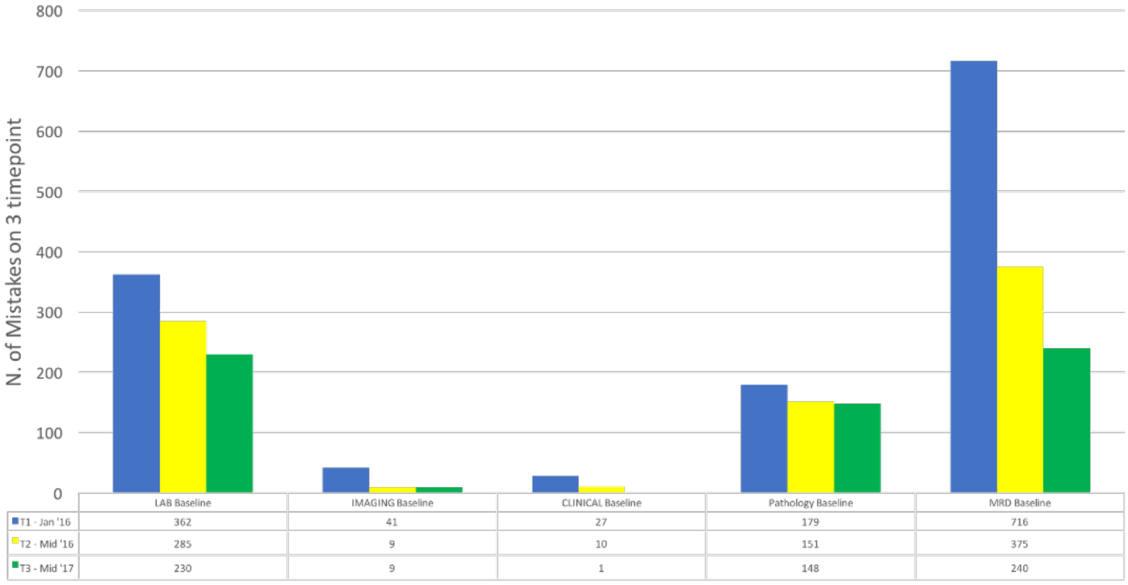
Table 17: qualitative analysis of cross controls by I2ECR

Rules	1	2	3	4	5	6	7
	BMinf AND AAStage	EN AND AAStage	AAStage AND BMInf AND EN	AAStage AND BMInf AND flow	AAStage AND BMInf AND qPCR	AAStage AND BMInf AND flow AND qPCR	AAStage AND L AND BMInf
Observations	522	521	1206	4304	522	608	1104
	608	3502	1311		1508	609	1606
	609		1410		3504	1206	
	1508		1606		3702	1410	
	1607		3204		3705	1606	
	1901		3504			1607	
	2503		3702			1901	
	2604		3705				
	2805						
	4302						
	4304						
TOT	11	2	8	1	5	7	2

4.3.2 Data Reporting

Source data verification (SDV) may reach 25% of entire cost for a clinical trial management. For this reason, clinical trials' sponsors may invest a quote of budget in tools for remote monitoring on data (J. R. Andersen et al. 2015). I2ECR 2nd main purpose is to boost on quality controls' implementation on data. In section 3.1.4 quality controls are dived in controls focused on a single feature (1st level controls) as well as between features (2nd level controls) describing a clinical phenomenon from a different point of view: e.g. bone marrow infiltration of disease may be assessed either via immunochemistry techniques (BMinf) or molecular biology techniques (cytofluorimetry - flowBM, flowPB). Quality controls must be in series: application of 1st level controls on data reflects on 2nd level controls.

Data cleaning effect has been globally assessed on DW FIL_MCL0208 tables listed in table 9. Moreover, in this PhD experience, a temporal analysis of improvement of data quality via I2ECR has been proposed. Figure 23 describes that a general decrease of mistakes detected by I2ECR over three time-points is detectable. Time points were T1: early 2016, T2: middle of 2016 and T3: early 2017. Highest decrease of number of mistakes is for MRD baseline features (from 716 of T1 to 240 of T3). Data-entry mistakes of LAB baseline variables drop-down of 36%. Moreover, some lowest change on mistakes is observable for both IMAGING and CLINICAL baseline groups.



	LAB Baseline	IMAGING Baseline	CLINICAL Baseline	Pathology Baseline	MRD Baseline
■T1 - Jan '16	362	41	27	179	716
■T2 - Mid '16	285	9	10	151	375
■T3 - Mid '17	230	9	1	148	240

Figure 23: Number of mistakes reported in 3 timepoint – T1: early '16, T2: middle '16, T3: early '17. Features graphed are organized in sub-groups: LAB baseline, IMAGING baseline, ClINICAL baseline, PATHOLOGY baseline, MRD baseline.

Figure 24 shows data cleaning effect among three different time point on a sub-group of features. Both flowBM and flowPB features are detected a strong decrease of data-entry errors from T1 to T3. The main reason can be addressed to data-recovery by laboratories. Out of range mistakes on WBC feature passed from more than 40 in T1 to 0 in T3. B2onB2Max feature constantly reduced the rate of about 28% (from 125 to 91) and Ki67 has the lowest reduction of mistakes (7%).

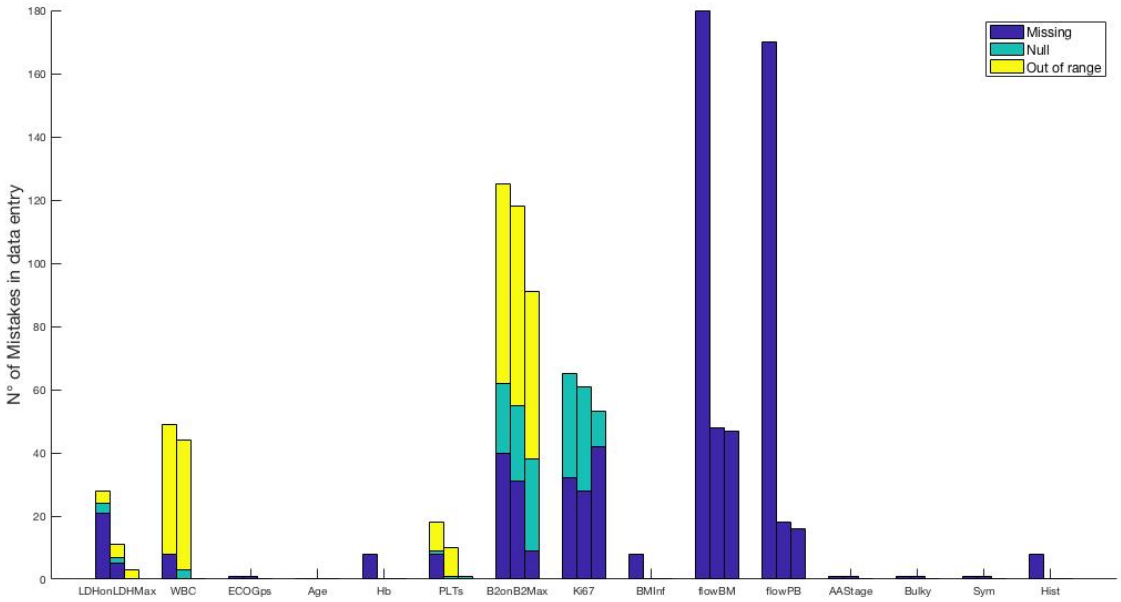


Figure 24: Number of mistakes reported in 3 timepoint – T1: early '16, T2: middle '16, T3: early '17. Data are shown for a subset of features.

Cleaning effect on MIPI value

Data quality improvement via I2ECR allowed to recover with prognostic factors calculated from statistical models which take as input baseline variables (subsection 3.1.3.2). In case of MIPI calculation, I and II level controls on LDHonLDHMax, WBC, ECOGps and Age features allowed to restore **23 MIPI values** (Hoster et al. 2008). Figure 25 shows PFS (Progression Free Survival) curves observed for subjects grouped in Low (green curve), Intermediate (blue) and High (red) risk classes in base of both MIPI Standard and MIPI simplified classification. PFS curves are detected for both T1 and T3 timepoint data, maintaining the same data of clinical outcome. Comparing T3 to T1 curves for both MIPI classifications, both significances among red and blue as well as blue and green lines rise. In fact, probability for a subject to be classified in a different class decreases:

- from 0.0025 to 0.0007 for p values assessed in case of Low vs Intermediate curves from MIPI Simplified.
- Up to the half for p values calculated from MIPI Standard.

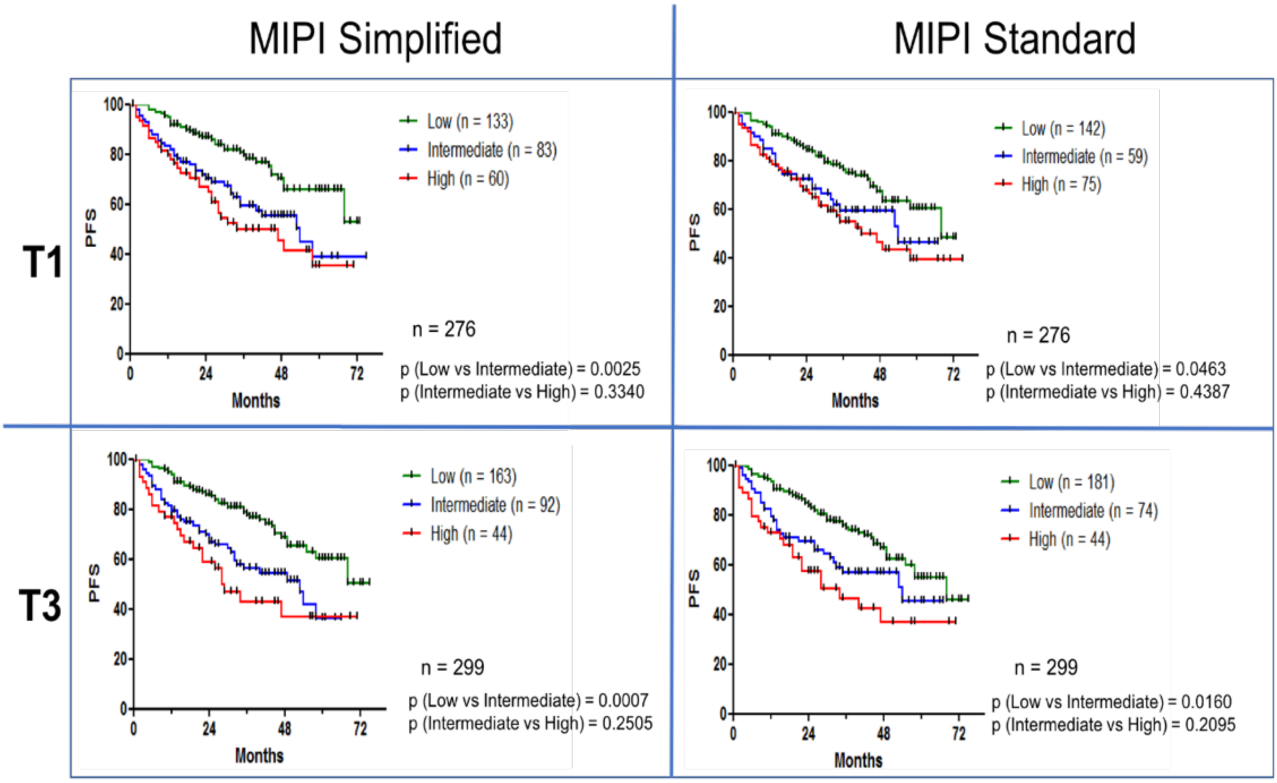


Figure 25: PFS curves observed for subjects grouped in Low (green curve), Intermediate (blue) and High (red) risk classes in base of both MIPI Standard and MIPI simplified classification. PFS curve are detected for both T1 and T3 time-point data, maintaining the same data of clinical outcome (Sept '17).

To perform a well-done data-management strategy from the starting of clinical trial shall allow a lower SDV effort in following (De 2011). For a translational point of view, data quality management via I2ECR allows a more significant stratification of patients in terms of PFS outcome (figure 26). Ki67 feature is an interesting example: in early 2016, some centers associated to 22 subjects a value of Ki67 equal to 0. Patients with a diagnosis of mantle cell lymphoma are characterized by an over-expression of this marker (Jares et al. 2012). A 0 value of Ki-67 is not acceptable by a pathologic point of view (Jares, Colomer, and Campo 2012). However, not expert stakeholders in pathology can easily corrupt veracity of data (Viceconti et al. 2015). Therefore, 22 subjects have been misclassified in "low" Ki-67 class (less than 30%), whereas they actually belong to "no-info" subjects (figure 26). This issue does not pour on stratification between low Ki-67 subjects and high Ki-67 subjects (p values does not change significantly), but a correct classification allows to evidence that subjects with "no-info" of Ki-67 (black line from figure 26) behave as "low" Ki-67 subjects (green line). Hence, clinicians may focalize in seeking of clinical reasons of this behavior.

Moreover, SDV verification via I2ECR reflects on veracity of aggregate features (table 5). MIPI re-calculation is the most important example of remote monitoring via I2ECR (figure 26). In detail, even if eCRF system automatically calculate MIPI from Hoster et al., a lack of quality control on independent variables that define MIPI as WBC and LDH consequently translate in loss of veracity on MIPI index (23 incorrect values– figure 26).

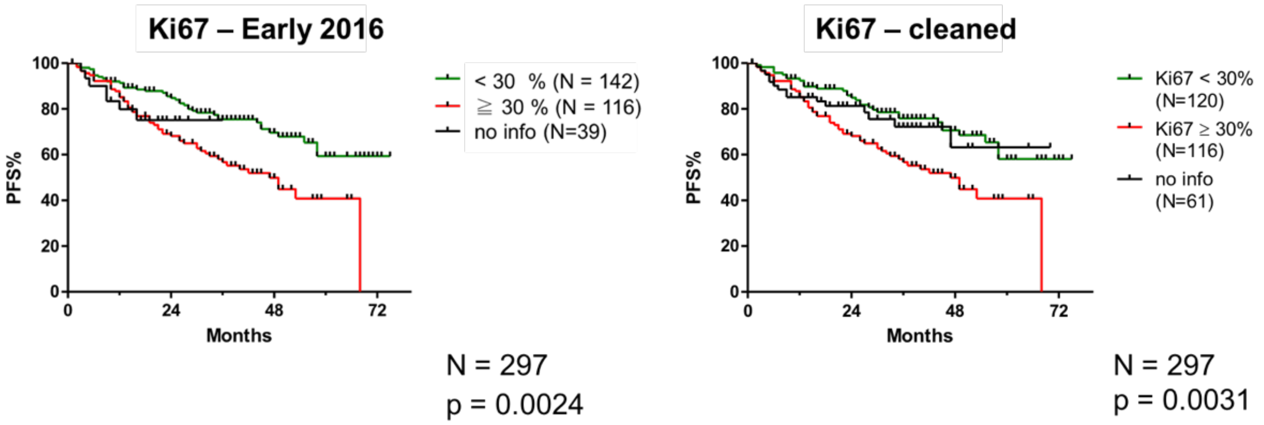
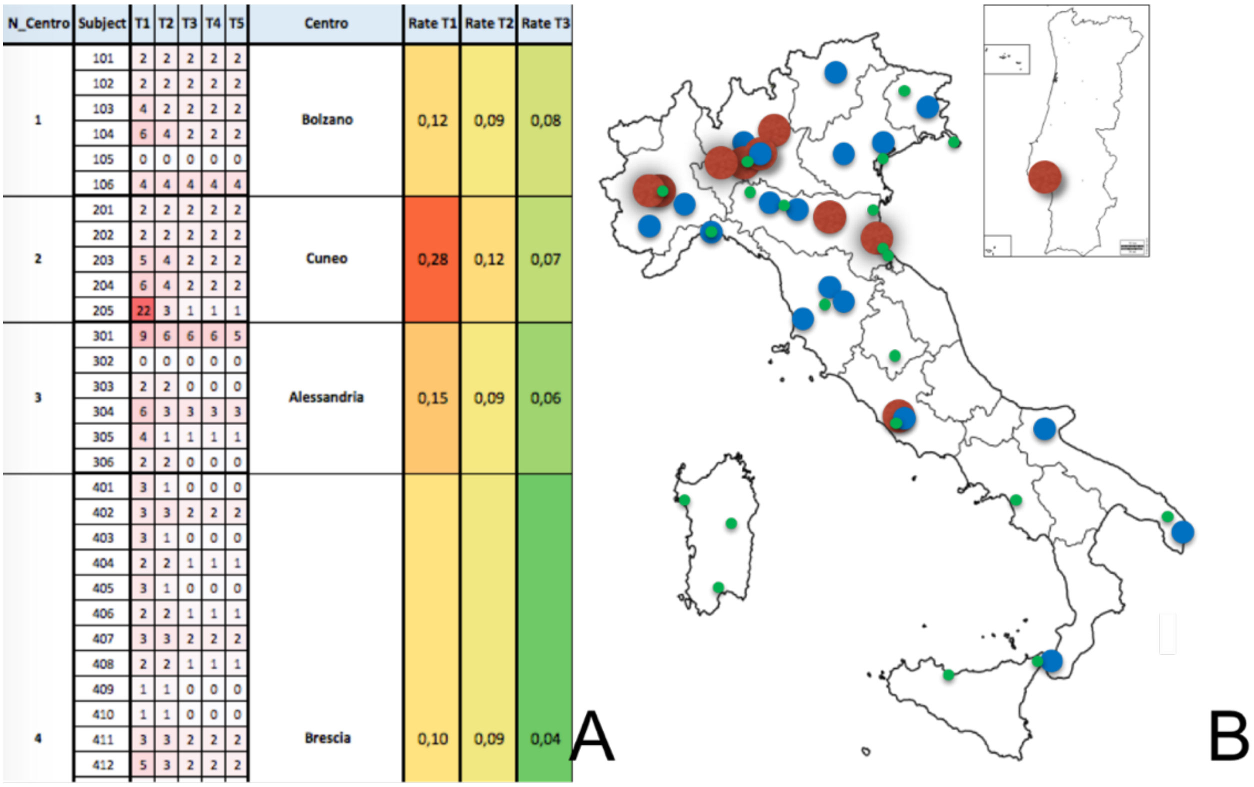


Figure 26: PFS outcome on time assessed for a cohort of subjects with a low (<30% - green line) Ki-67 proliferation marker vs. high (>= 30% - red line). Early 2016 dataset wrongly recorded 22 subjects within group of low Ki-67 (green line). However, adjusted Ki-67 values moved 22 subjects to "no-info" group. P values do not record significant variations between models.

Quality Index2

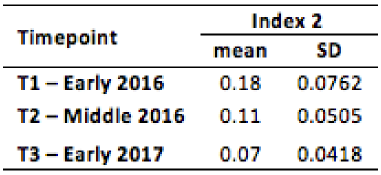
Index2 has been modelled to measure quality performance of centers that were active in study enrollment (figure 27B). Centers were classified in reference of the number of enrolled subjects: red centers enrolled 45% of total subjects (300), blue centers enrolled among 5 and 10 subjects and green centers less than 5. To evaluate improvement in data-entry, index has been assessed on 3 time points from early 2016 (T1) to early 2017 (T3) (figure 27A and ANNEX 1). Therefore, has been possible to detect the capacity by centers to catch up mistakes filled in eCRF. At T1, worst centers expressed an index2 higher than 0.2: Cuneo, Modena, Ravenna, Udine (0.35), Pisa (0.36), Roma Cattolica (0.33), Roma Tor Vergata (0.35), Verona (0.43). On the contrary at T3, best centers expressed an Index equal (or close) to 0: Siena (0.00), Nuoro (0.00), Cesena (0.02), San Giovanni Rotondo (0.03), Lisbon (Portugal), Monza, Tricase (0.03) and Roma La Sapienza (0.03). Moreover, data quality improvement has been assessed. Cuneo, Udine, Pisa, Roma Cattolica and Verona caught up with highest part of mistakes. In general, Index 2 slightly decreases from 0.18 at T1 to 0.07 confirming that centers recovered many mistakes with eCRFs (table 18). Figure 28 represents Index_2 global trend for all centers and its standard deviation.



N_Centro	Subject	T1	T2	T3	T4	T5	Centro	Rate T1	Rate T2	Rate T3
1	101	2	2	2	2	2	Bolzano	0,12	0,09	0,08
	102	2	2	2	2	2				
	103	4	2	2	2	2				
	104	6	4	2	2	2				
	105	0	0	0	0	0				
	106	4	4	4	4	4				
2	201	2	2	2	2	2	Cuneo	0,28	0,12	0,07
	202	2	2	2	2	2				
	203	5	4	2	2	2				
	204	6	4	2	2	2				
	205	22	3	1	1	1				
3	301	9	6	6	6	5	Alessandria	0,15	0,09	0,06
	302	0	0	0	0	0				
	303	2	2	0	0	0				
	304	6	3	3	3	3				
	305	4	1	1	1	1				
	306	2	2	0	0	0				
4	401	3	1	0	0	0	Brescia	0,10	0,09	0,04
	402	3	3	2	2	2				
	403	3	1	0	0	0				
	404	2	2	1	1	1				
	405	3	1	0	0	0				
	406	2	2	1	1	1				
	407	3	3	2	2	2				
	408	2	2	1	1	1				
	409	1	1	0	0	0				
	410	1	1	0	0	0				
	411	3	3	2	2	2				
	412	5	3	2	2	2				

Figure 27: assessed Index 2 on three time-point from early 2016 (T1) to early 2017 (T3) to evaluate centers' performance in recovering data-entry mistakes (A) – see ANNEX 1. Centers that were active in enrollment for MCL0208 clinical trial (B).

Table 18: Index_2 global trend at three timepoint.



Timepoint	Index 2	
	mean	SD
T1 – Early 2016	0.18	0.0762
T2 – Middle 2016	0.11	0.0505
T3 – Early 2017	0.07	0.0418

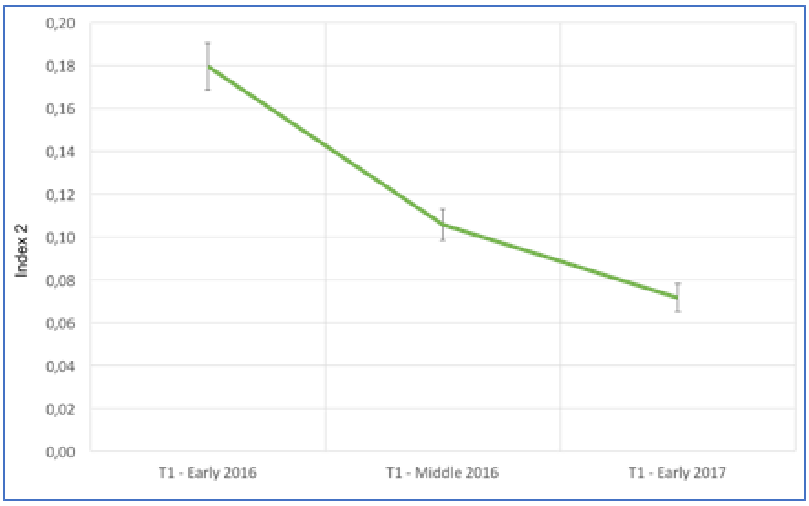


Figure 28: Index_2 global trend for all centers at three timepoint.

I2ECR extractions: Automatic MIPI calculation.

I2ECR allowed automatic calculations of prognostic factors (table 5). In this case, figure 29 shows the representation of both MIPI standard values and MIPI biologic on total patients basing on Hoster (Hoster et al. 2016). Patients are ordered in function of MIPI Standard Value (in blue). MIPI biologic value for patients has been represented (in red). MIPIb curve has been fit with a polynomial of 3rd degree for a better visualization. Comparing MIPISt to MIPIb-Fit, there is a slight shift.

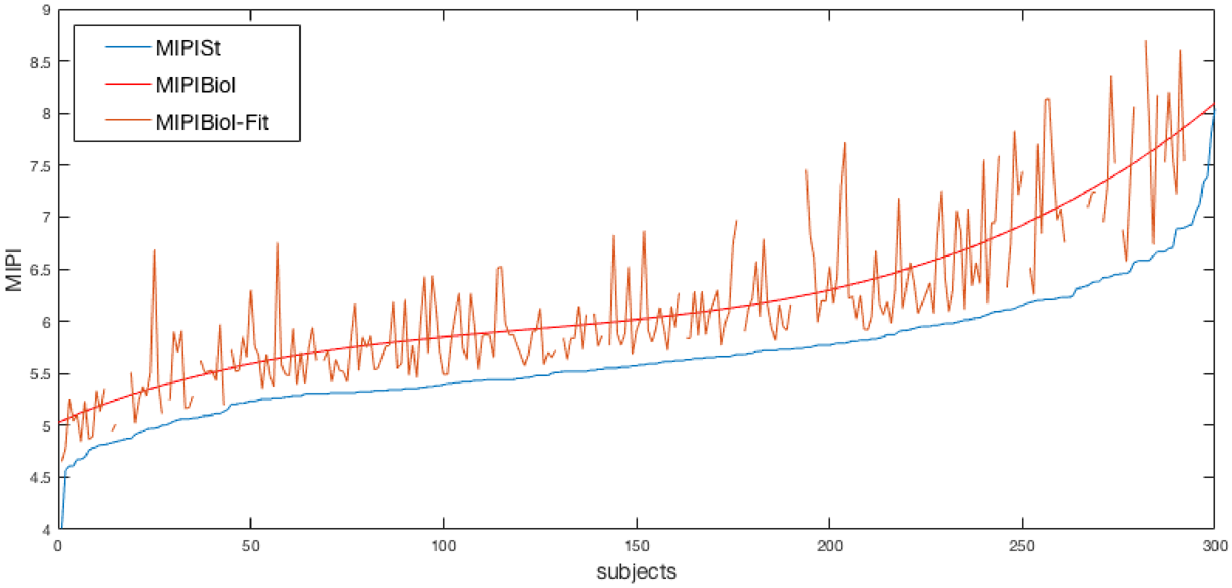
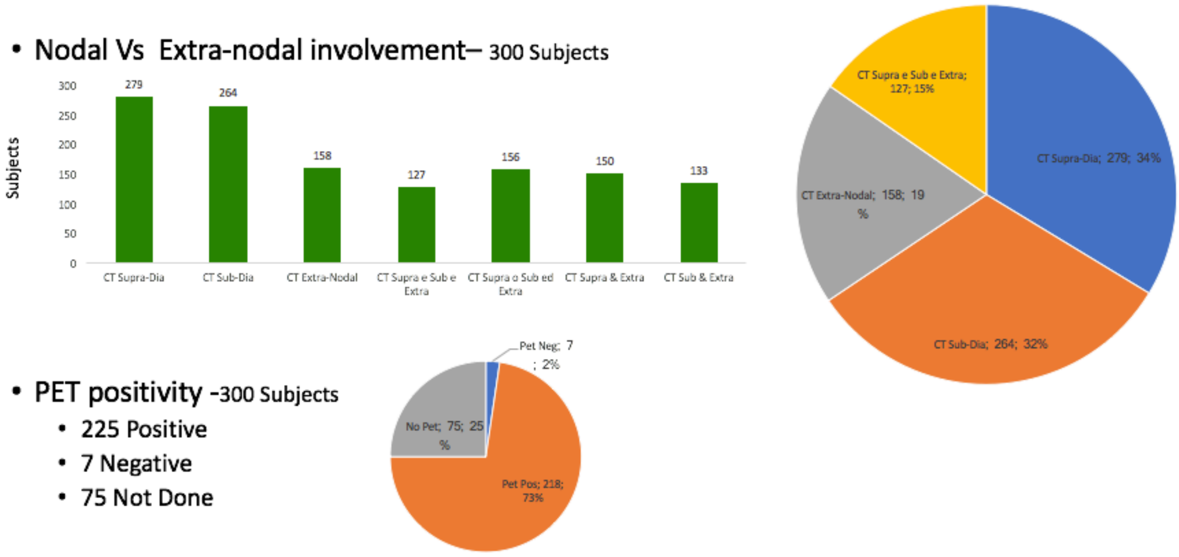


Figure 29: MIPI Standard VS MIPI Biologic for MCL0208 clinical trial

Staging of disease

I2ECR allowed the extraction of clinical data on disease staging. Figure 30 depicts involvement status of anatomical areas detected both by CT (Computational Tomography) and PET (Positron Emission Tomography) scan. Green histogram and big pie both show that 279 subjects had nodal supra-diaphragmatic involvements and 264 had sub-diaphragmatic involvements. 127 subjects had both nodal and extra-nodal lesions. Little pie diagram describes that 73% of enrolled subjects (218 subjects) had at least a lesion detected by PET scan against 2% (7). 75 subjects did not execute PET scan (not mandatory for MCL0208 clinical study).



- Nodal Vs Extra-nodal involvement– 300 Subjects

- PET positivity -300 Subjects
  - 225 Positive
  - 7 Negative
  - 75 Not Done

Figure 30: involvement status of anatomical areas detected both by CT (Computational Tomography) and PET scan.

Furthermore, a more precise extraction of anatomical nodes involvements detected by CT scan has been possible. Figure 31 is a quantitative representation of subjects' distribution concerning both supra (left) or sub (right) diaphragmatic lesions. In this clinical study 210 as well as 208 observations suffered of positive CT scan on Axillary, whereas 256 as well as 258 observations had a negative scan on Renal-hilus. 14 observations were affected by a bulky (>5 [cm]) involvement in para-lomboartic anatomic section.

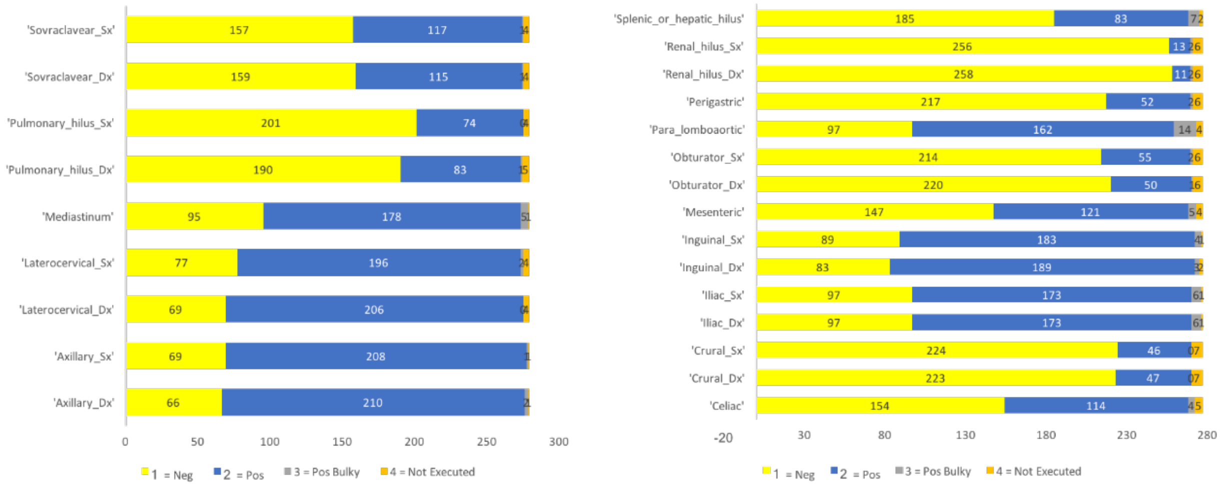


Figure 31: involvement status of anatomical areas detected both by CT (Computational Tomography) and PET scan.

Mutations

In this subsection, subjects grouped in risk classes according to MIPI prognostic factor are graphed in relation to mutational profiles. Figure 32 depicts this quantitative distribution. According to the MIPI, 182 patients were classified as low risk, 73 intermediate risk subjects and 43 as high-risk subjects. WT ("No Mut" in figure 32) and mutated observations describe different behaviors. For WT, percentage of subjects on total of subjects of respective class constantly decreases from 35% to 15%. However, for ATM, CCND1, WHSC1 and TP53 this value rises. Patients with ATM mutations are more representative in risk class 3 (40%), twice than risk class 1, where 20% of observations are mutated. CCND1, WHSC1 and TP53 assumes similar trend, despite a minor mutational expression. Percentage of subjects with both CCDN1 and WHSC1 lesions range among 5% of risk class 1 up to 10% of risk class 3.

Both staging of disease and mutations extractions through the I2ECR project refer to the information processing application of data warehousing theory. According to Han, "information processing supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts, or graphs". In I2ECR information processing is allowed by connection between DBMS and Matlab (or PHP for I2ECR web-service) via ODBC/JDBC.

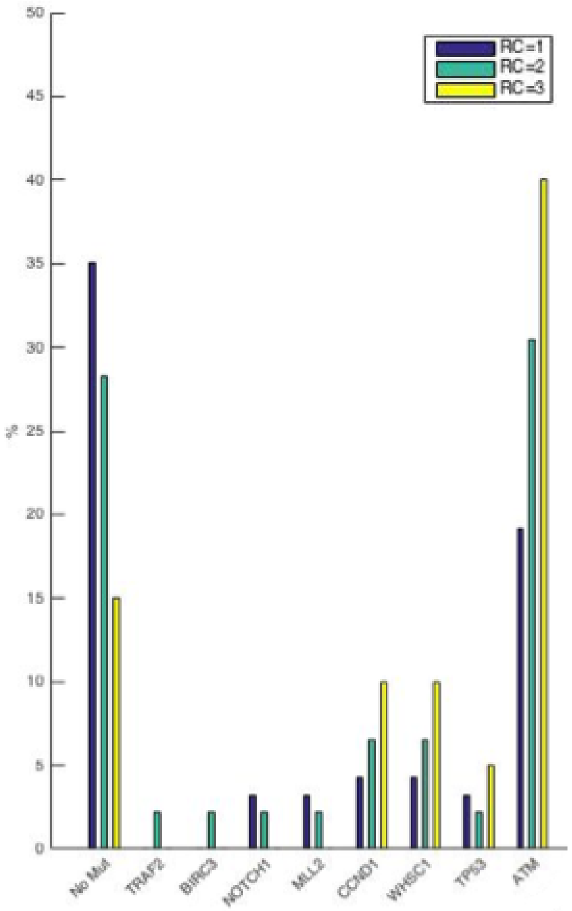


Figure 32: percentage of observations grouped in risk classes for each mutation analyzed.

Clinical Responses

I2ECR allowed analysis on MCL0208 clinical response (table 2 of subsection 3.1.3.1). Table 19 describes clinical responses' distribution at each restaging for different MIPI risk classes. First of all, observations at baseline (299) drop off from 279 at R1, to 262 at R2 and 245 at R3. This is due to expected decrease of patients during treatment. If clinical responses are related to MIPI classification at baseline, subjects with complete response (CR) classified in risk class 1 at baseline dramatically rise from 46 (R1) to 132 (R2), contrarily to subjects with a partial response (PR) that decrease from T1 (130 observations) to T2 (31 observations). This trend is confirmed by subjects classified in both risk class 2 and 3. Subjects with stable disease (SD) slightly increase at R3 if belonging to both risk class 2 (3) and risk class (2) at baseline.

Table 19: clinical responses' distribution at each restaging for different MIPI risk classes.

		1st Restaging = T1					2nd Restaging = T2					3rd Restaging = T3				
N. Sub at Baseline	**MIPI Risk Class**	**N. Sub**	**CR**	**PR**	**PD**	**SD**	**N. Sub**	**CR**	**PR**	**PD**	**SD**	**N. Sub**	**CR**	**PR**	**PD**	**SD**
182	1	178	46	130	2	0	167	132	31	4	0	159	149	8	2	0
74	2	66	15	46	4	1	62	37	23	2	0	54	45	6	0	3
43	3	35	16	18	1	0	33	27	5	0	1	32	26	4	0	2
299	**TOT**	**279**	**77**	**194**	**7**	**1**	**262**	**196**	**59**	**6**	**1**	**245**	**220**	**18**	**2**	**5**

Moreover, I2ECR allowed qualitative cross controls on clinical responses through restaging (from rule n. 8 of table 8 included in section 3.1.4). Table 20 detects that 8 subjects clinically regressed from a CR at T1 to PR (or SD in case of observation 3501) at T2. 3 of these subjects (809, 902 and 1302) furtherly have a CR at T3.

Table 20: qualitative analysis on clinical responses among three restaging.

Subject	Clinical Response – R1	Clinical Response – R2	Clinical Response – R3
519	**CR**	**PR**	**PD**
809	**CR**	**PR**	**CR**
902	**CR**	**PR**	**CR**
1302	**CR**	**PR**	**CR**
1310	**CR**	**PR**	**PR**
2404	**CR**	**PR**	**PR**
3501	**CR**	**SD**	**SD**
3704	**CR**	**PR**	**PR**

Figures 30, 31 and 32 purpose different level of aggregation of data, whereas table 19 shows an example of extraction with high-level granularity manner. First of all, I2ECR allowed to investigate diffusion of disease among enrolled subjects, from both a macro (figure 30) and a micro (figure 31) point of view. Clinically, to evaluate outcomes on patients with lymph node involvement detected via CT of anatomical site A (e.g. Axillary) than anatomical site B (e.g. Spleen) shall be scientifically relevant. Again, the combination of clinical data (e.g. class risk distribution of patients at diagnosis) with mutational analysis at baseline (figure 32) emphasizes the translational characterization of this clinical study. Technically, multidimensionality of data analysis in oncology field is not allowed by a "statistics" databases management (Han, 2012). Data warehousing overcomes this limitation. Conceptually, dealing with data representation, if a data sheet represents data in 2 dimensions (figure 8), a data warehouse introduces to n-dimensional data management.

Multidimensional way to analyze data from a data warehouse is correctly associated to OLAP (On-Line Analytical Process) than informational processing systems (Han, Kamber, and Pei 2012). However, an idea of multidimensional data storage (that refers to Data Cube modeling theory) is given in Figure 33. Figure 33 describes data represented in table 19 with a 3-dimensional perspective. In this case, patients enrolled in MCL0208 clinical trial are aggregated in reference to 3 attributes: RCMIPISt from Clinical_data_Baseline table, id_Clinical_Response from both Clinical_Data_Restagins and Clinical_Response tables and N_timepoint from Clinical_Data_Restagins. Figure 33A shows a result on using aggregate functions on data: in this case function SUM along all 3 dimensions: class risk, Clinical Response and Timepoint. Figure 33B instead depicts the "slicing" operations on same data retrieved from a data ware-house: data analysts (in our case the clinical stakeholders) can easily analyze a clinical behavior through different levels of abstraction without implement complex and time-consuming formulas on a 2-dimensional statistical data-sheet.

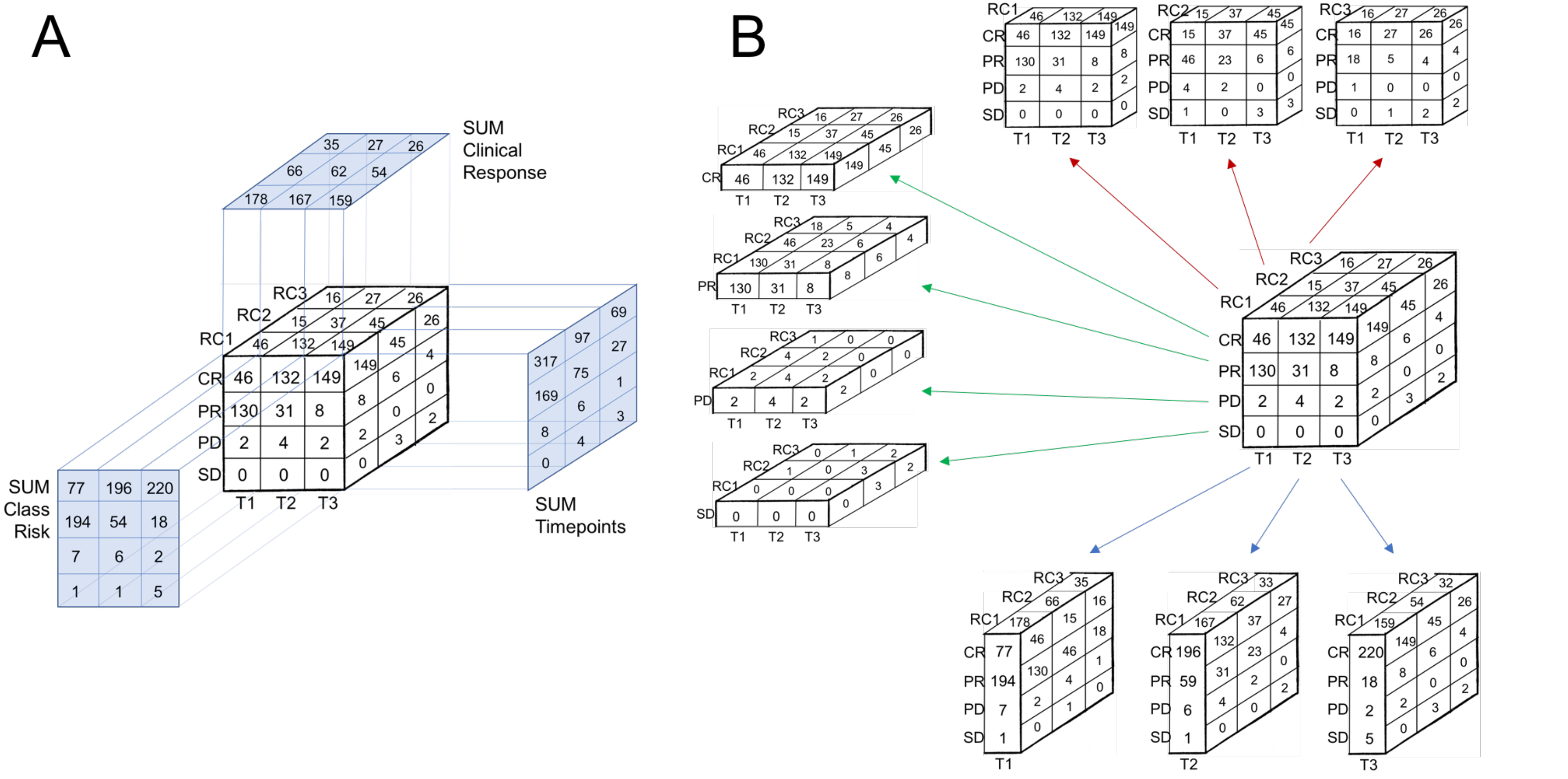


Figure 33: Figure A shows a result on using aggregate functions on data: in this case function SUM along all 3 dimensions: class risk, Clinical Response and Timepoint. Figure B depicts the “slicing” operations on same data retrieved from a data ware-house

Toxicities and Study Interruption

I2ECR allowed to analyze toxicities recorded for subjects by centers. Toxicities are expressed by CTCAE international standard. Table 21 lists hematological toxicities from 296 patients from T1 (post R-CHOP treatment). 85% of observed subjects accused hematological toxicities of grade >=3. Among these, 232 observed subjects suffered of granulocytes toxicities, 238 of PLTs and 233 of WBC. 68% of observed subjects recorded "No Hematological" toxicities (199).

Table 21: Maximum toxicities by patients during the treatment phase. Percentage are based on 296 patients with at least the R-CHOP- 1 recorded.

Toxicity (CTCAE)	No Toxicities		Grade 1-2		Grade >= 3	
	N	%	N	%	N	%
Hematological	30	10	14	5	250	85
Granulocytes	54	18	8	3	232	79
HB	45	15	126	43	123	42
PLTs	45	15	11	4	238	81
WBC	49	17	12	4	233	79
No Hematological	32	11	63	21	199	68

Finally, I2ECR allowed cross controls actions between recorded grade 5 toxicities (categorized as mortal adverse events in reference of table 3 of sub-section 3.1.3.1.). and "study interruption" cases. Patient 5205 with a mortal adverse event due to a pulmonary and infective bacterial toxicity has not been recorded with a death (9) study interruption class by center (table 22).

Table 22: qualitative analysis on toxicity vs. study interruption data-entry

Subject	Toxicity Type	CTCAE	Study Interruption
707	**Febrile neutropenia**	**5**	**9 (death)**
4503	**Infective viral**	**5**	**9 (death)**
5205	**Pulmonary and Infective bacterial**	**5**	**0**

4.3.3 Data Analysis from I2ECR

Feature Selection project

Feature selection (or data reduction) aim is to remove noise effect on data (Han, Kamber, and Pei 2012) improving the performance of mining in terms of result comprehensibility (Fahrudin et al. 2017). Application of Feature Selection techniques to oncology field is novel (Ravi et al. 2017), (Shi 2016), (Fahrudin, Syarif, and Barakbah 2017). Feature selection techniques can reduce a dataset in terms of dimensions or sample-size, can be parametric or not (Han, Kamber, and Pei 2012), supervised or unsupervised (Rosati, Balestra, and Molinari 2014). In I2ECR, an example of feature selection supervised on patients' risk class variables (MIPI based on Hoster et al.) has been proposed. This means that the selected variables are the most important to discriminate patients according to their classes.

In more detail, has been demonstrated that data-processing, if designed on a well-done pipeline, may both reduce the noise effect given by missing values (i) and extract clinical variables that significantly stratify patients in terms of clinical outcomes (ii). Feature selection can improve this process, acting a significant reduction of data redundancy avoiding the loose of information.

First of all, for a survival analysis point of view, feature selection may be considered as a discovery tool to investigate on new outcomes research as well as is either a supervised or an unsupervised technique. Once the both subset R1 and R2 were discretized in step C of subsection 3.1.6, feature selection (FS) algorithm was applied on. Table 23 shows FS by QRA algorithm on both subsets R1 and R2. For subset R1, QRA selected Age, WBC, B2M, Protein, IgG, $qPCR_{BM}$, $qPCR_{PB}$ features. For subset R1, QRA selected Age, WBC, B2M, Protein, IgG, $qPCR_{BM}$,

qPCR_{PB} features. For subset R2, QRA selected PLTs, B2M, IgG, qPCR_{BM}, qPCR_{PB}, flowPB features.

Table 23: FS by QRA algorithm on both subsets R1 and R2

Subsets			FS via QRA	FS via QRA
N	**R1**	**R2**	**R1**	**R2**
1	Age	-	1	-
2	LDH	-	0	-
3	WBC	-	1	-
4	PLTs	PLTs	0	1
5	Hb	Hb	0	1
6	B2M	B2M	1	1
7	Protein	Protein	1	0
8	Albumin	Albumin	0	0
9	IgG	IgG	1	1
10	IgA	IgA	0	0
11	IgM	IgM	0	0
12	AST	AST	0	0
13	ALT	ALT	0	0
14	GGT	GGT	0	0
15	ALP	ALP	0	0
16	Bilirubin	Bilirubin	0	0
17	qpcrBM	qpcrBM	1	1
18	qpcrPB	qpcrPB	0	1
19	flowPB	flowPB	1	1
20	flowBM	flowBM	0	0
21	BMInfperc	BMInfper	0	0
22	IgHOmo	IgHOmo	0	0

Selected features from R2 have been related to PFS outcome (updated in September 2017) in order to evaluate FS performance on starting MCL0208 baseline features (figure 34). For each variable, subjects were divided into two groups: wild type (green curves) versus abnormal (red curves) values in base of table 11. Among these, Hb and FlowBM features significantly allowed patients stratification (p=0.0058 and p=0.0133).

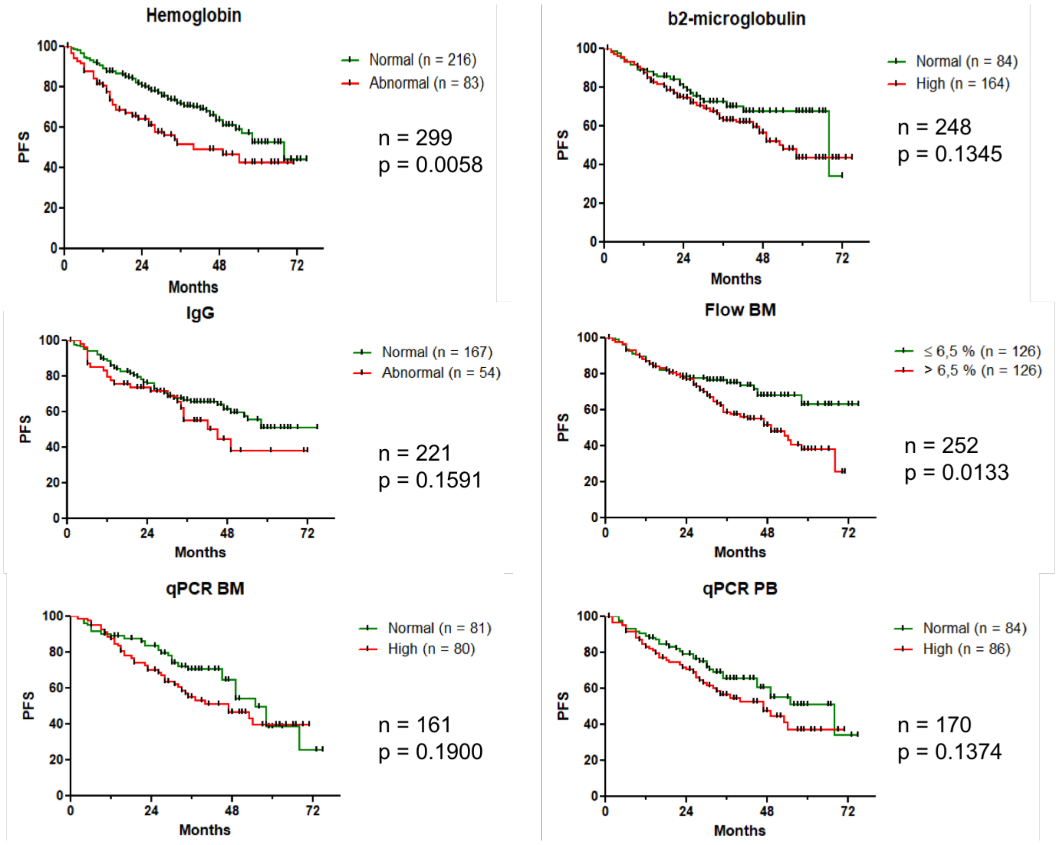


Figure 34: selected FS related to PFS. For each variable, subjects were divided into two groups: wild type (green curves) versus abnormal (red curves) values. Among these, Hb and FlowBM features significantly allowed patients stratification (p=0.0058 and p=0.0133).

However, focusing on Missing Values management, validation of this study consisted in assessment of classification performances, using the initial dataset and after each step of the applied methodology (table 24). The percentage of observations that have been correctly, incorrectly or not classified by the K-nearest neighbor is presented. Observing the second column of table 24, missing values imputation (step IIB) allowed to rise the percentage of correct classification of about 20% with respect to previous step. Moreover, even if the variables discretization (step IIC) slightly reduced the performances with respect to step IIB, the feature selection method (step III) produced a further improvement of the classification accuracy, meaning that the discarded variables represented a source of noise for identification of the patients' risk class. On the contrary, analyzing the number of incorrectly and not classified patients (last two columns of Table 24), these percentages were considerably reduced after MVs imputations, meaning that the missing information was necessary for a correct identification of the patient risk class. Furthermore, feature selection did not produce a significant deterioration of the performances.

In Table 25 the confusion matrix obtained at the end of the data quality improvement process is presented. Each percentage is calculated with respect to the total number of subjects belonging to a specific class risk for mantle cell lymphoma disease. As it emerges from the table, the highest accuracy has been obtained for the low risk class (92.83%), that is the largest group of subjects. Although the K nearest neighbor slightly suffers the effects of the class imbalance, this class should influence the classification accuracy. The lowest performance was returned for the intermediate risk class (65.8%). From the MIPI point of view, this class is assigned to patients obtaining a score between 5.7 and 6.2 (Hoster et al. 2008), which is a very tight interval. This can be due to a misclassification of border observations that can affect also the classifier accuracy.

Table 24: Validation results

	Patients correctly classified	Patients incorrectly classified	Patients not classified
Initial dataset	57.6%	31.9%	10.5%
Step I	57.2%	31.6%	11.2%
Step IIA	59.5%	32.2%	8.2%
Step IIB	79.3%	14.8%	5.9%
Step IIC	77.6%	16.8%	5.6%
Step III	81.9%	13.2%	4.9%

Table 25: Final confusion matrix

		Predicted Class			
		Not Classified	Low Risk	Intermediate Risk	High Risk
Actual Class	Low Risk	2.2%	92.3%	1.6%	3.8%
	Intermediate Risk	2.7%	21.9%	65.8%	9.6%
	High Risk	7.0%	7.0%	9.3%	76.7%

DELPHI – Data ELaboration to Predict Hypothetical assocIations.

DELPHI (Data ELaboration to Predict Hypothetical assocIations) is a project to discover associations among huge amounts of data retrieved from MCL0208 via I2ECR. "Delphi" name is common in several areas (e.g. project management, social sciences[25]). These have in common the goal to purpose novel methodologies to discover information from retrospective data (Delphi is inspired from ancient Greek oracle). In this case, DELPHI is a tool that identifies novel putative associations between clinical and molecular features. In order to statistically compare categorical with continues features, these later have been opportunely categorized. Validation has been led splitting initial dataset in 2 subsets:

- A discovery-set of 195 subjects from 10 to 52 active centers.
- A validation set of 105 subjects from 1 to 9 active centers.

DELPHI identified 231/1860 (12%) associations in the discovery set, of whom 64% (149/231) were confirmed in the validation set (figure 35A). Among these mismatches, main contribution is imputed to associations between clinical and laboratory data with baseline tumor burden by quantitative PCR, and pathology data (figure 35B and 35C).

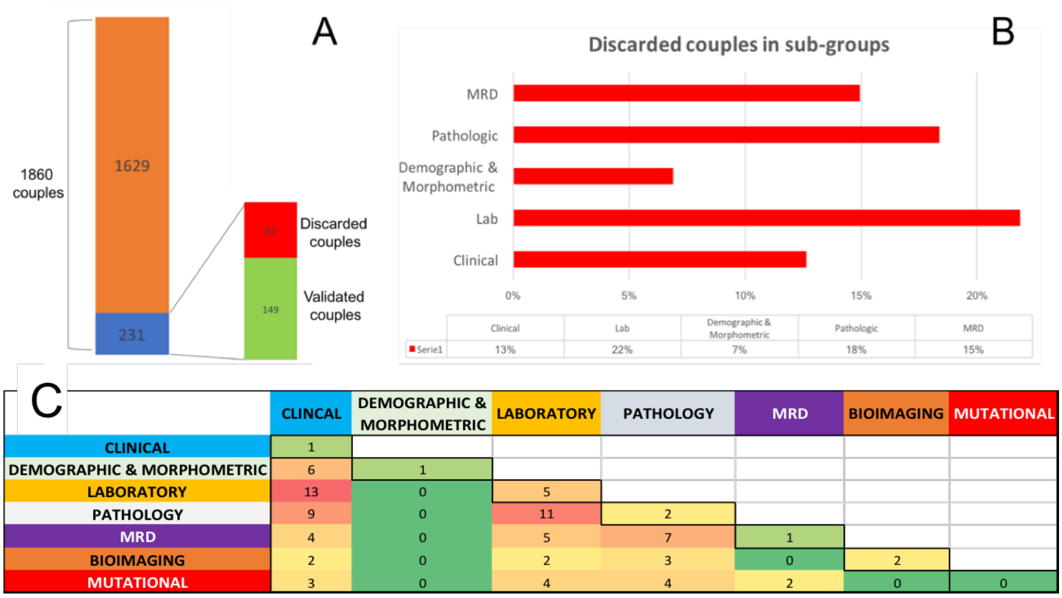
The clinical team classified as "expected" 242/1860 variables matches (13%), 54% of whom were confirmed by DELPHI (figure 35A and 36A). The thickest

[25] https://www.aacu.org/delphi

ribbons have V>0.5: among these, bone marrow tumor invasion (BMInf) by immunohistochemistry with nodal (NLTB) and extra- nodal (ENLTB) tumor burden at CT scan; Lymphocytes with BMInf by cytofluorimetry (Flow). TP53 mutations and blastoid histology with a V=0.39. Moreover, discovery of novel associations is shown in Figure 36B. 54 of 1860 (3%) matches were identified by DELPHI as statistically significant unexpected associations: the thickest ribbons have V>0.35. Among these, MIPI, albumin and LDH with Hemoglobin; BMInf by qPCR with MIPIc. Moreover, TP53 mutations and GammaGT (V=0.34), as well as NOTCH1 mutations with Alkaline- Phosphatase (V=0.35) and MIPIb with Beta2 Microglobulin, B2M (V=0.33).

Initial dataset extracted from I2ECR presented a non-monotone data "missingness" (Horton and Lipsitz 2001) and lead to follow the unbalanced validation strategy. To overcome this technical drawback, sample size of each association has been monitored in reference to association factor for each couple. Hence, sample size analysis (i) and statistical test application (ii) have been employed to seek putative association among included features (figures 35 e 36). Cramer V technique has been used to assess the strength of associations between 2 variables. There are several techniques to assess strength of association between 2 variables (e.g. lift and χ^2 methods - (Han, Kamber, and Pei 2012)). This choice depends by variable type and differently performs if comparison is lead between a categorical variable with more than 2 classes and a categorical value with 2 classes (also known as nominal variable). Despite χ^2, Cramer's V allows statistical adjustment about sample size and unbalanced contingency tables[26]. An example of this is given by accounting associations with either MIPI score (1: low risk; 2: intermediate risk; 3: high risk) or pharmacogenomics polymorphisms (e.g. ABCB1-1236_C>T can assume 0 for WT cases, 1 heterozygotes polymorphisms or 2 homozygotes polymorphisms). Observations with missing values were discarded by the algorithm. In order to consider confounding interactions and multiple comparison issues, every association needs further investigation on independent data series.

[26] http://uregina.ca/~gingrich/ch11a.pdf



	Clinical	Lab	Demographic & Morphometric	Pathologic	MRD
🟥 Serie1	13%	22%	7%	18%	15%

	CLINCAL	DEMOGRAPHIC & MORPHOMETRIC	LABORATORY	PATHOLOGY	MRD	BIOIMAGING	MUTATIONAL
CLINICAL	1						
DEMOGRAPHIC & MORPHOMETRIC	6	1					
LABORATORY	13	0	5				
PATHOLOGY	9	0	11	2			
MRD	4	0	5	7	1		
BIOIMAGING	2	0	2	3	0	2	
MUTATIONAL	3	0	4	4	2	0	0

Figure 35: Figure A. 231/1860 (12%) associations detected in the discovery set of whom 64% were in the validation set. Figure B and C. Validation discarding is caused by a technical reason: different sample size between discovery and validation sets (200 vs. 100 subjects). Red bars diagram (B) shows the distribution of discarded couples grouped in macro groups of features. Table C quantities couples discarding.

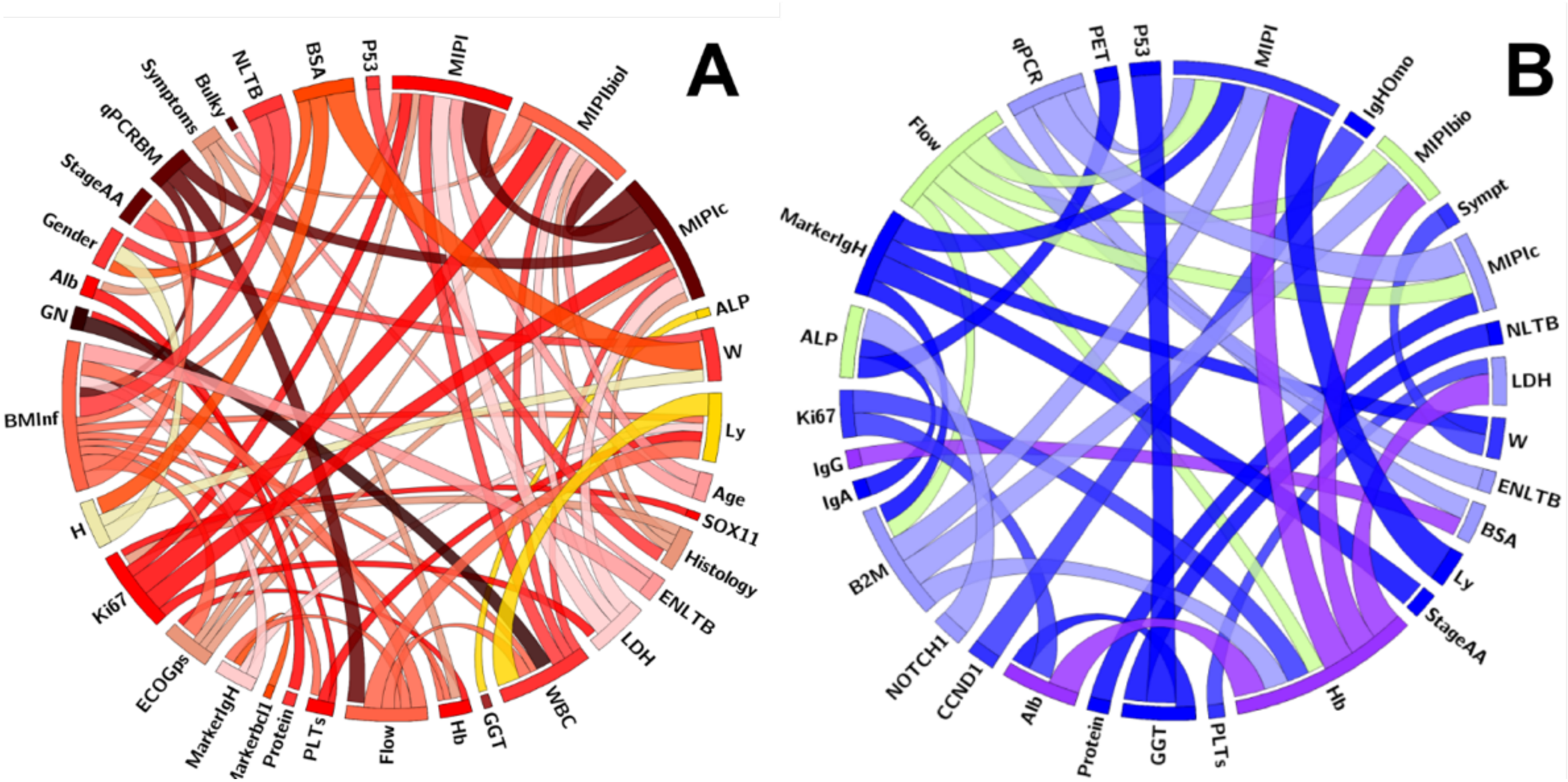


Figure 36: Circles of expected (A) and unexpected (B) associations. Ribbons represent the strength of association between significant couples of features extracted from baseline of MCL0208. Figure A shows the associations expected by clinicians. Figure B shows the unexpected associations.

4.4 I2ECR software

I2ECR SW is a webserver application designed to provide clinical researchers an easy-to-use environment to manage datasets associated to translational clinical trials. In this section, “SET-UP A DW” and “LOAD A NEW DATASET IN A DW” activities are described (figure 37 – 1 and 2).

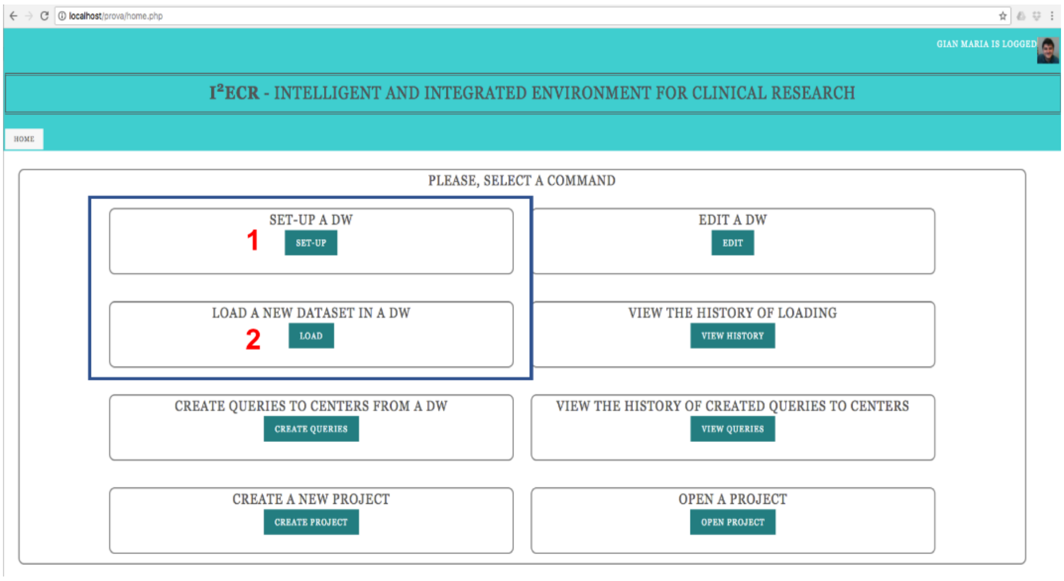
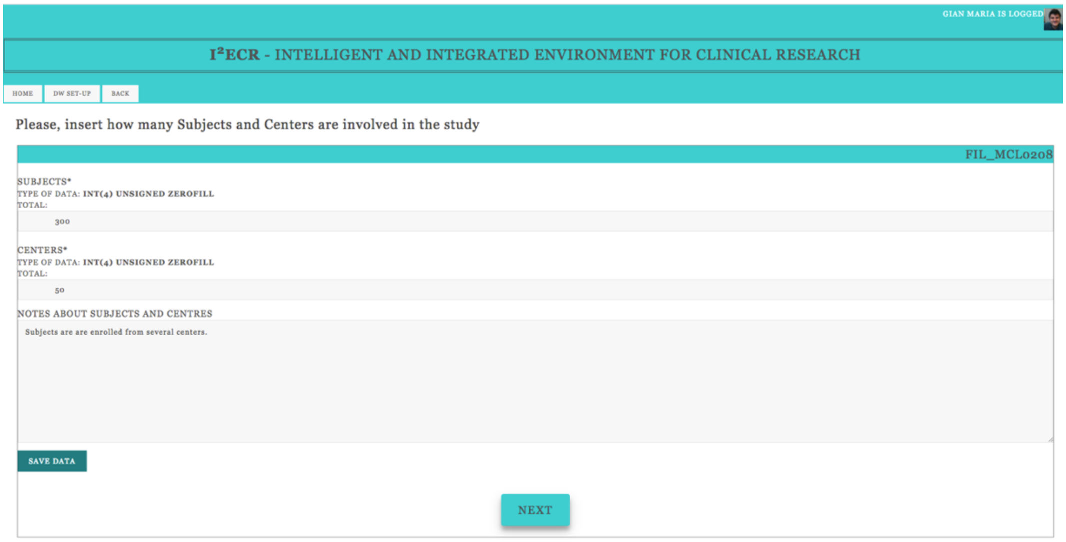


Figure 37: C I2ECR Home Page.

SET-UP A DW

User accesses to “SET-UP A DW” use case from Home page (figure 37). To set-up a DW, user needs for protocol information. First of all, I2ECR needs to associate to DW total number of restaging and follow-ups, the expected number of enrolled subjects and active centers in enrollment (figure 38 – GUIA). Once that a table is associated to right time-point (figure 39 – GUIB), user has to encode control constraints for data cleaning. In particular, the user selects a feature for all set of already features encoded in DW (in our case FIL_MCL0208 DW) and assigns controls about “MISSING DATA ENCODING”, “0 VALUE DATA ENCODING”: e.g. what is a typical missing value associated to LDH by centers? Is it possible that LDH value is equal to 0? Once that user clicks on “Get Selected values”, I2ECR saves data into the server. In figure 40 – GUIC is depicted a GUI for setting validity and normality controls on features. E.g. for WBC user inserts the correct unit of magnitude associated to WBC (10^9/L). However, figure 41 – GUID shows the resuming GUI for section "SET-UP DW


GIAN MARIA IS LOGGED
I²ECR - INTELLIGENT AND INTEGRATED ENVIRONMENT FOR CLINICAL RESEARCH
HOME
DW SET-UP
BACK
FIL_MCL0208
SAVE DATA
NEXT

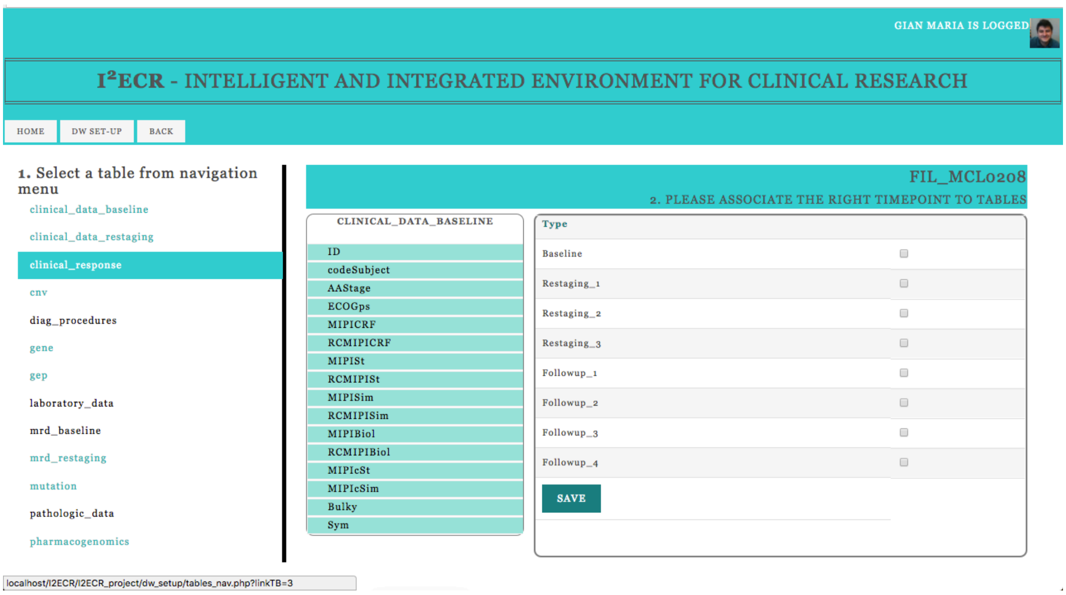
Please, insert how many Subjects and Centers are involved in the study

SUBJECTS*
TYPE OF DATA: INT(4) UNSIGNED ZEROFILL
TOTAL:
300

CENTERS*
TYPE OF DATA: INT(4) UNSIGNED ZEROFILL
TOTAL:
50

NOTES ABOUT SUBJECTS AND CENTRES
Subjects are are enrolled from several centers.

Figure 38: GUI A - I2ECR data ware-house setup. Encoding of Number of subjects and number of centers.


GIAN MARIA IS LOGGED
I²ECR - INTELLIGENT AND INTEGRATED ENVIRONMENT FOR CLINICAL RESEARCH
HOME
DW SET-UP
BACK
FIL_MCL0208
SAVE

1. Select a table from navigation menu

clinical_data_baseline
clinical_data_restaging
clinical_response
cnv
diag_procedures
gene
gep
laboratory_data
mrd_baseline
mrd_restaging
mutation
pathologic_data
pharmacogenomics

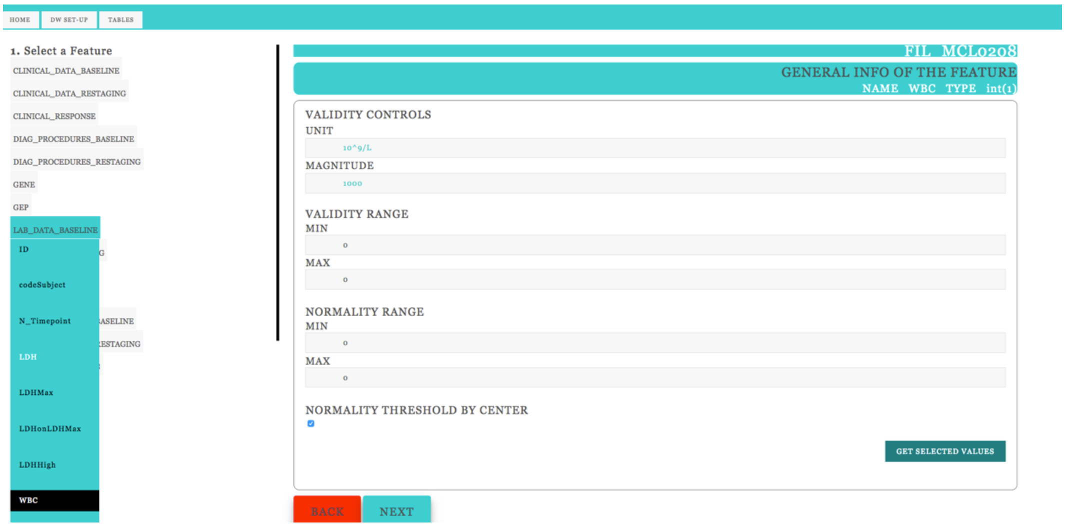
2. PLEASE ASSOCIATE THE RIGHT TIMEPOINT TO TABLES

CLINICAL_DATA_BASELINE
ID
codeSubject
AAStage
ECOGps
MIPICRF
RCMIPICRF
MIPISt
RCMIPISt
MIPISim
RCMIPISim
MIPIBiol
RCMIPIBiol
MIPIcSt
MIPIcSim
Bulky
Sym

Type	
Baseline	□
Restaging_1	□
Restaging_2	□
Restaging_3	□
Followup_1	□
Followup_2	□
Followup_3	□
Followup_4	□

localhost/I2ECR/I2ECR_project/dw_setup/tables_nav.php?linkTB=3

Figure 39: GUI B - I2ECR data ware-house setup. Definition of the exact time-point to each table.



NAME WBC TYPE int(1)

VALIDITY CONTROLS

UNIT: 10^9/L

MAGNITUDE: 1000

VALIDITY RANGE

MIN: 0

MAX: 0

NORMALITY RANGE

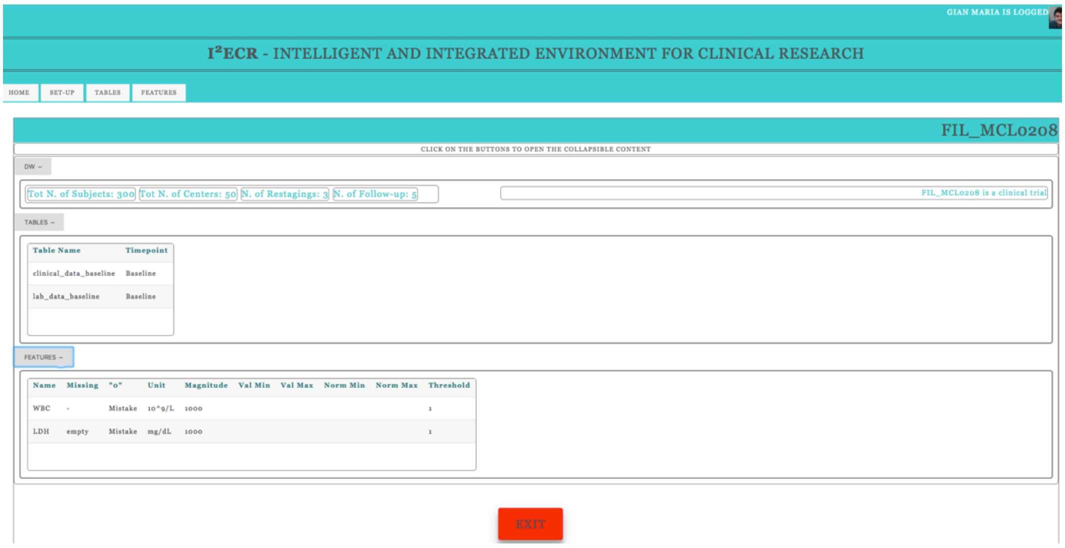
MIN: 0

MAX: 0

NORMALITY THRESHOLD BY CENTER

☑

Figure 40: GUI C - I2ECR data ware-house setup. Encoding of controls for each feature.


GIAN MARIA IS LOGGED
I^{2}ECR - INTELLIGENT AND INTEGRATED ENVIRONMENT FOR CLINICAL RESEARCH
HOME
SET-UP
TABLES
FEATURES
FIL_MCL0208
CLICK ON THE BUTTONS TO OPEN THE COLLAPSIBLE CONTENT
DW –
TABLES –
FEATURES –
EXIT

Tot N. of Subjects: 300 | Tot N. of Centers: 50 | N. of Restagings: 3 | N. of Follow-up: 5

FIL_MCL0208 is a clinical trial

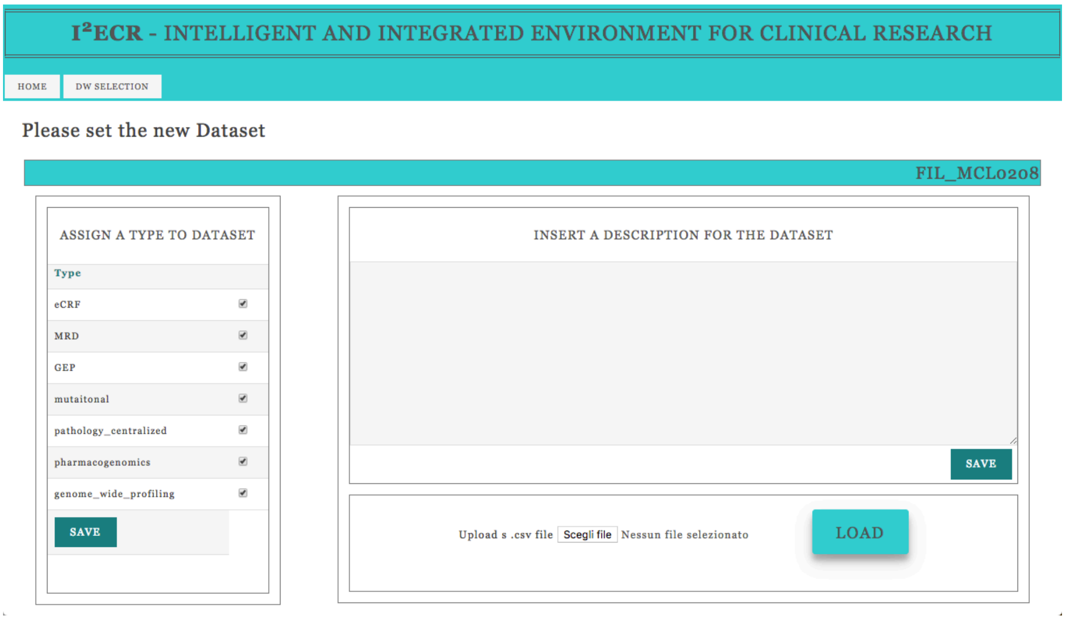
Table Name	Timepoint
clinical_data_baseline	Baseline
lab_data_baseline	Baseline

Name	Missing	"0"	Unit	Magnitude	Val Min	Val Max	Norm Min	Norm Max	Threshold
WBC	-	Mistake	10^9/L	1000					1
LDH	empty	Mistake	mg/dL	1000					1

Figure 41: GUI D - I2ECR data ware-house setup. Resume of all activity done by user in the module.

LOAD A NEW DATASET IN A DW

User accedes to “LOAD A NEW DATASET IN A DW W” use case from Home page (figure 37). A dataset can be loaded if a DW has already been set-up. To load a dataset, user needs for protocol information. First of all, I2ECR needs to associate to dataset total the right type of data included in dataset (if associated from clinical baseline, Outcomes or Mutational studies) - figure 42 – GUIA. At this point, user selects preferred dataset clicking on “choose your file” (figure 43 – GUIB). Once that a dataset has been loaded, I2ECR executes a “subjects’ status analysis” on dataset and indicates no-conventional subjects. User can choose to fix the problem for that subject or not or simply moving it into “Stand-by table” (figure 44 – GIC). Figure 45 – GUID fixes single features detected from dataset, in order to populate the right feature encoded in DW set-up. I2ECR automatically detects number of features from dataset and asks user to assign a feature to associate feature controls to DW population. Finally, figure 46 – GUIE depicts the resume of the activity followed by user.


I²ECR - INTELLIGENT AND INTEGRATED ENVIRONMENT FOR CLINICAL RESEARCH
HOME
DW SELECTION

Please set the new Dataset

FIL_MCL0208

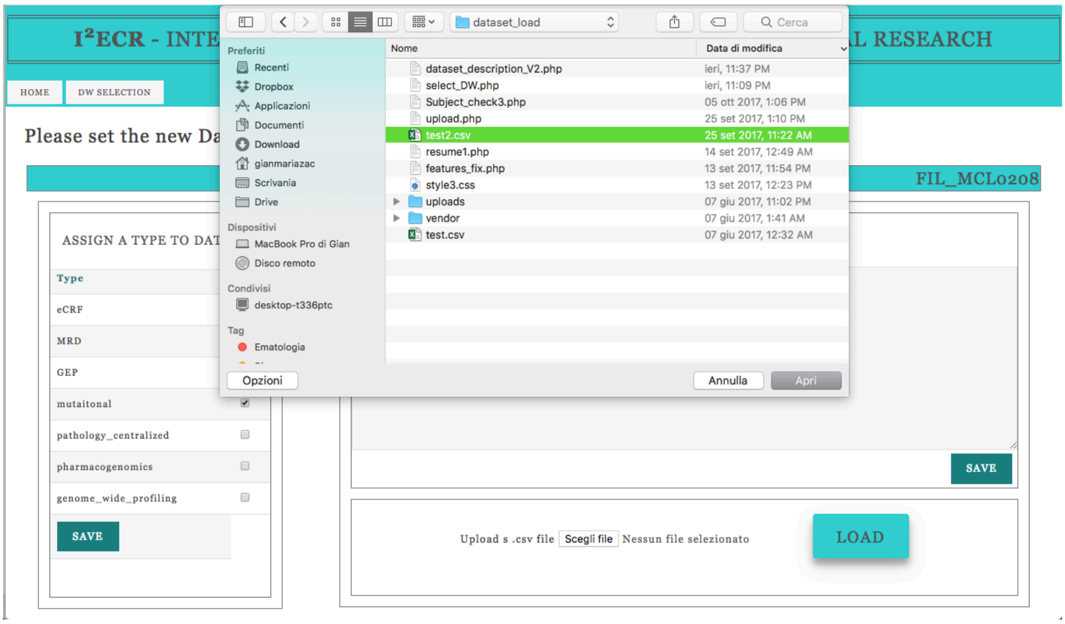
ASSIGN A TYPE TO DATASET

Type	
eCRF	☑
MRD	☑
GEP	☑
mutaitonal	☑
pathology_centralized	☑
pharmacogenomics	☑
genome_wide_profiling	☑

SAVE
INSERT A DESCRIPTION FOR THE DATASET
SAVE

Upload s .csv file Scegli file Nessun file selezionato

Figure 42: GUI A. I2ECR import of a dataset. Assignment of a type of dataset.



I²ECR - INTE… …L RESEARCH

HOME | DW SELECTION

Please set the new Da…

FIL_MCL0208

ASSIGN A TYPE TO DAT…

Type	
eCRF	
MRD	
GEP	
mutaitonal	☑
pathology_centralized	□
pharmacogenomics	□
genome_wide_profiling	□

SAVE

SAVE

Upload s .csv file Scegli file Nessun file selezionato LOAD

dataset_load

Preferiti: Recenti, Dropbox, Applicazioni, Documenti, Download, gianmariazac, Scrivania, Drive

Dispositivi: MacBook Pro di Gian, Disco remoto

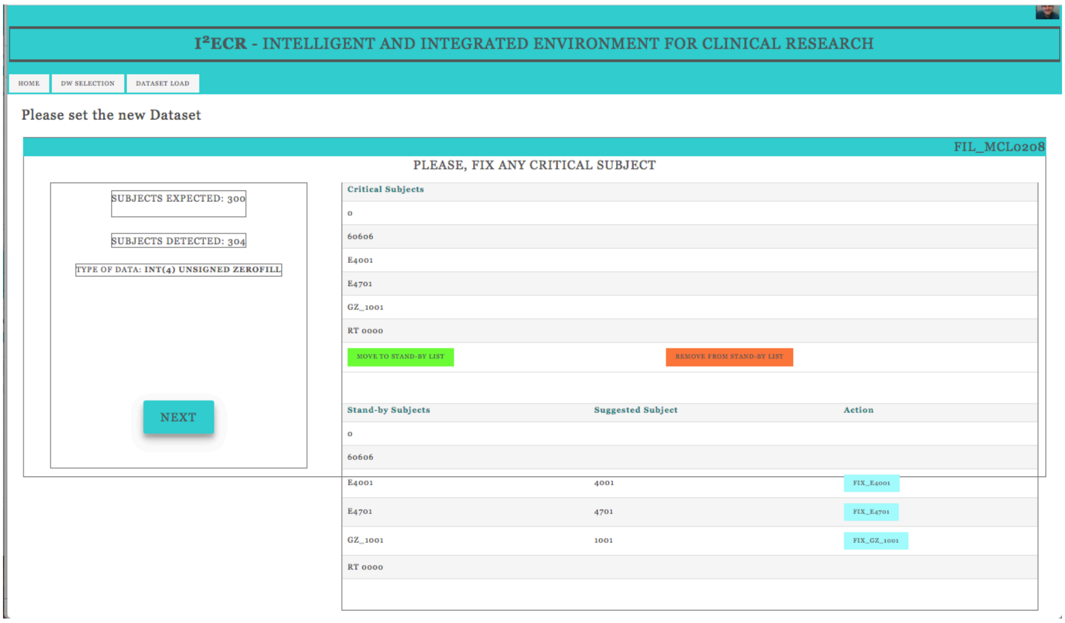
Condivisi: desktop-t336ptc

Tag: Ematologia

Nome	Data di modifica
dataset_description_V2.php	ieri, 11:37 PM
select_DW.php	ieri, 11:09 PM
Subject_check3.php	05 ott 2017, 1:06 PM
upload.php	25 set 2017, 1:10 PM
test2.csv	25 set 2017, 11:22 AM
resume1.php	14 set 2017, 12:49 AM
features_fix.php	13 set 2017, 11:54 PM
style3.css	13 set 2017, 12:23 PM
uploads	07 giu 2017, 11:02 PM
vendor	07 giu 2017, 1:41 AM
test.csv	07 giu 2017, 12:32 AM

Opzioni | Annulla | Apri

Figure 43: GUI B. I2ECR import of a dataset. Loading of a dataset.



I²ECR - INTELLIGENT AND INTEGRATED ENVIRONMENT FOR CLINICAL RESEARCH

HOME | DW SELECTION | DATASET LOAD

Please set the new Dataset

FIL_MCL0208

PLEASE, FIX ANY CRITICAL SUBJECT

SUBJECTS EXPECTED: 300

SUBJECTS DETECTED: 304

TYPE OF DATA: INT(4) UNSIGNED ZEROFILL

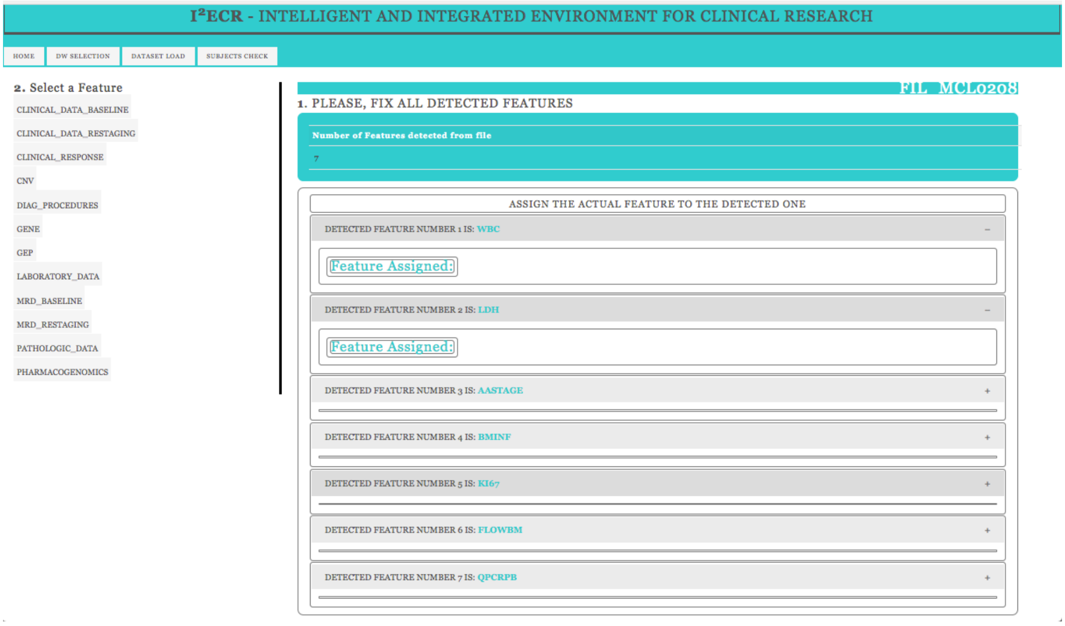
NEXT

Critical Subjects
0
60606
E4001
E4701
GZ_1001
RT 0000

MOVE TO STAND-BY LIST | REMOVE FROM STAND-BY LIST

Stand-by Subjects	Suggested Subject	Action
0		
60606		
E4001	4001	FIX_E4001
E4701	4701	FIX_E4701
GZ_1001	1001	FIX_GZ_1001
RT 0000		

Figure 44: GUI C. I2ECR import of a dataset. Management of incongruent subjects in reference of the requirements on the subject encoding defined in the DW_setup module.



I^{2}ECR - INTELLIGENT AND INTEGRATED ENVIRONMENT FOR CLINICAL RESEARCH

2. Select a Feature

CLINICAL_DATA_BASELINE
CLINICAL_DATA_RESTAGING
CLINICAL_RESPONSE
CNV
DIAG_PROCEDURES
GENE
GEP
LABORATORY_DATA
MRD_BASELINE
MRD_RESTAGING
PATHOLOGIC_DATA
PHARMACOGENOMICS

FIL_MCL0208

1. PLEASE, FIX ALL DETECTED FEATURES

Number of Features detected from file
7

ASSIGN THE ACTUAL FEATURE TO THE DETECTED ONE

DETECTED FEATURE NUMBER 1 IS: WBC

DETECTED FEATURE NUMBER 2 IS: LDH

DETECTED FEATURE NUMBER 3 IS: AASTAGE

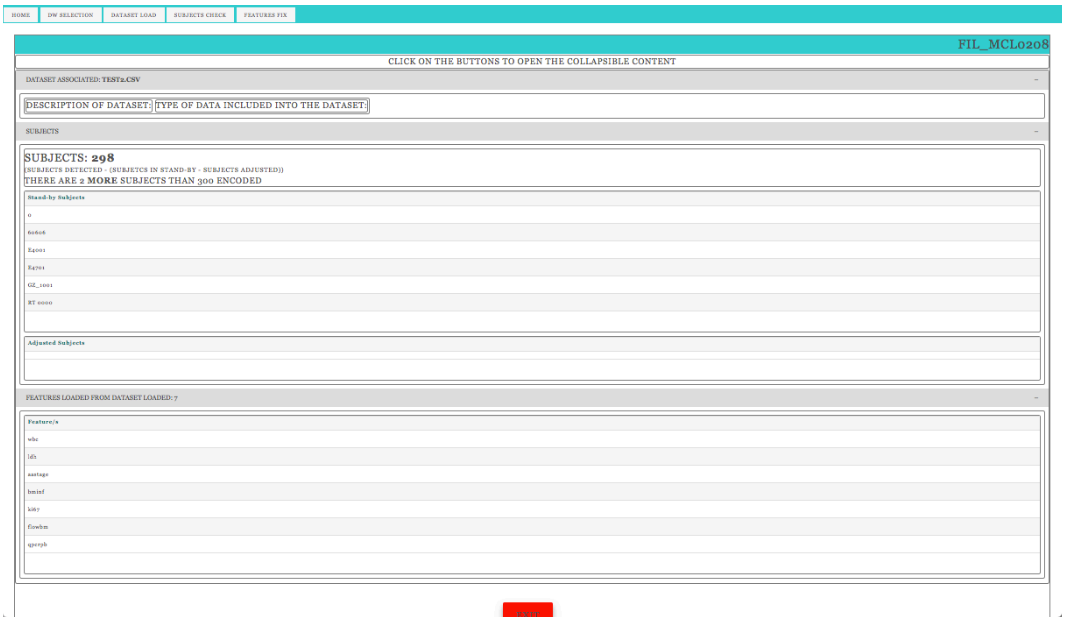
DETECTED FEATURE NUMBER 4 IS: BMINF

DETECTED FEATURE NUMBER 5 IS: KI67

DETECTED FEATURE NUMBER 6 IS: FLOWBM

DETECTED FEATURE NUMBER 7 IS: QPCRPB

Figure 45: GUI D. I2ECR import of a dataset. Assignment of the right feature to each column detected from imported dataset by software.



FIL_MCL0208

CLICK ON THE BUTTONS TO OPEN THE COLLAPSIBLE CONTENT

DATASET ASSOCIATED: **TEST2.CSV**

DESCRIPTION OF DATASET: TYPE OF DATA INCLUDED INTO THE DATASET:

SUBJECTS

SUBJECTS: **298**
(SUBJECTS DETECTED - (SUBJETCS IN STAND-BY - SUBJECTS ADJUSTED))
THERE ARE 2 **MORE** SUBJECTS THAN 300 ENCODED

Stand-by Subjects
0
60606
E4001
E4701
GZ_1001
RT 0000

Adjusted Subjects

FEATURES LOADED FROM DATASET LOADED: 7

Feature/s
wbc
ldh
aastage
bminf
ki67
flowbm
qpcrpb

Figure 46: GUI E. I2ECR import of a dataset. Resume of all activity done by user in the module.

Chapter 5

Conclusions

5.1 Data ware-housing in I2ECR

Clinical research is changing traditional scopes (Mirnezami, Nicholson, and Darzi 2012) moving from the classical application of epidemiology principles to high-tech data-driven projects (Harris et al. 2009). Data- driven projects need platforms tailored on data, in order to optimize resources and time in setting-up data quality strategies.

Data ware-housing theory may help clinical researchers to achieve outcomes, because it allows to design data-platforms oriented to subject centrality and data analysis (Zapletal et al. 2010; Han, Kamber, and Pei 2012). International projects as I2B2 (Murphy et al. 2006) and Harmony[27] are interesting examples of this technological innovation. Unfortunately, those concepts are not commonly applied to "local" clinical research projects. FIL_MCL0208 data ware-house is suited on a hematological clinical trial (Zaccaria, Ferrero et al., 2017). Hence, the structure of the implemented DW depends on the data structure of the MCL0208 study. In order to propose I2ECR as a broad platform including a wide set of clinical and genomic data for every hematological malignancy, in case of a new application, the design of further specific Entity-Relationships (E-R) model is required. In fact, a new E-R model has been designed on a further clinical trial on Multiple Myeloma (Eudract Code 2014-000782-53), sponsored by the University of Turin.

5.1.1 Informed consent, privacy and data integrity.

At the time of enrollment to a clinical trial, a patient must sign an informed consent. "The informed consent is a process by which a subject voluntarily confirms his or her willingness to participate in a clinical trial after having been informed of

[27] https://www.harmony-alliance.eu/

all aspects of the trial that are relevant to the subject's decision to participate" (ICH 1996). Moreover, signing the informed consent, the subject allows sponsors to both access to his/her demographic data and manage his/her clinical data. Hence, the sponsor shall maintain a security system that prevents unauthorized access to the data, guaranteeing to use unambiguous subject identification strategies that may avoid to clearly identify each enrolled subject. Technically, subjects' demographic data and either clinical or genomic data must be physically stored in different databases. Within those databases, subjects must be anonymized.

I2ECR does not includes personal information of patients enrolled in the MCL0208 clinical study. Clinical and genomic data has been collected with the previous authorization by the scientific board. Moreover, according to Han et al, "a data warehouse refers to a data repository that is separately maintained from an organization's operational databases" (Han, Kamber, and Pei 2012). In this case, the operational databases are whom used for the eCRF platform.

In reference to data integrity, I2ECR addresses ALCOA requirements as follows (Woollen 2010):

- **A**ttributability: log information about who entered data in eCRFs is not objective of data ware-housing because eCRFs already include this information. Data ware-housing does not correspond to OLAP (On-Line Analytical Process) technology (Han, Kamber, and Pei 2012).
  However, within I2ECR, the traceability on data editing is followed in observance of E6 Good Clinical Practice Guidance which affirms that "if data are transformed during processing, it should always be possible to compare the original data and observations with the processed data" (ICH 1996).
- **L**egibility: data are readable any time.
- **C**ontemporaneity: this feature is not allowed by data warehousing (eCRFs already included that information).
- **O**riginality: data are compliant to either clinical or laboratory row data.
- **A**ccuracy: data are correct, exact and free from errors.

About the physical infrastructure, I2ECR databases are mounted on a server (OS Microsoft) included in a Virtual Private Network (VPN) physically placed inside the Hematology Unit of University of Turin. User access to VPN is managed by a web-based software (MySQL) mounted on a dedicated server (OS Linux).

5.1.2 Candidate involvement.

The strong inter-disciplinarily nature of the I2ECR project has required to point out the involvement of actors to each of its phase. First of all, the candidate daily participated to each of activities in first person working in Hematology unit of University of Turin and strongly bridging with Biolab group of Politecnico di Torino. The project achievement has been allowed thanks to the share of information with several stakeholders belonging to both clinical and biomedical engineering fields.

- Problem analysis and requirements collection (Figure 13): I2ECR project derives from the clinical need to optimize data-management of a III phase multicenter clinical trial. Data were-housing has been proposed by Biolab group as a technical solution. The collection of requirements was performed by the candidate.
- DW design, implementation and population: those activities have been entirely taken up by the candidate in collaboration with Biolab group (moreover, the winning of a grant allowed to involve a young biomedical engineer who has been trained by the candidate.).
- Data cleaning and standardization: this phase has been supervised by clinical team. Once that clinical requirements have been addressed, the candidate developed a data quality strategy in collaboration with Biolab team.
- Data Analysis. The feature selection project has been proposed and supervised by Biolab team. The candidate worked on the analysis in collaboration with a colleague. However, DELPHI project has been entirely developed by the candidate with the clinical team. All projects have been shared with groups.
- SW design and implementation. The I2ECR front-end SW has been developed (and it is ongoing) under the supervision of Biolab team: SW design and technical implementation (e.g. the choice to use the PHP programming language) have been addressed by the candidate in reference of the clinical usability of the SW.

5.2 Next steps for I2ECR

5.2.1 DW modelling: integration with hematology bio-bank

Since many years, the laboratory of the Hematologic Division is the centralized lab for the storage of peripheral blood (PB) and bone marrow (BM) samples in the context of multicentric prospective clinical trials or specific, academic research projects. As an example, in the last five years, the lab of the Hematologic Division stored >800 diagnostic samples of non-Hodgkin lymphoma, more than 650 diagnostic samples of multiple myeloma and about 250 diagnostic samples of myeloproliferative disorders. First of all, to implement a hematologic biobank, several points must be considered:

- take into account processes that rule biobank sample management.
- Best technological solutions to record of each sample (e.g. barcode, QR code).
- To apply a quality control on robustness of the platform and correctness of data.
- To safely track data (anonymization in respect of privacy).

I2ECR allows to integrate clinical trials DW as FIL_MCL0208 to software dedicated to the biobank management to allow analysis of correlations between clinical and biologic data. This may be a promising challenge to boost on translational research on hematologic diseases in cohort studies. Technically this is possible via DBMS (Data Base Management System). Some commercial biobank management systems as EasyTrack2D®[28] are developed with “open-source” technologies (MySQL® and JavaScript) and integration with I2ECR may result easy and rapid (figure 47).

Integrating I2ECR to hematologic biobank, comprehensive of patients' samples for research and advanced diagnostic analyses, the idea of a more "personalized medicine" could be translated into the clinical practice.

[28] http://easytrack2d.it/

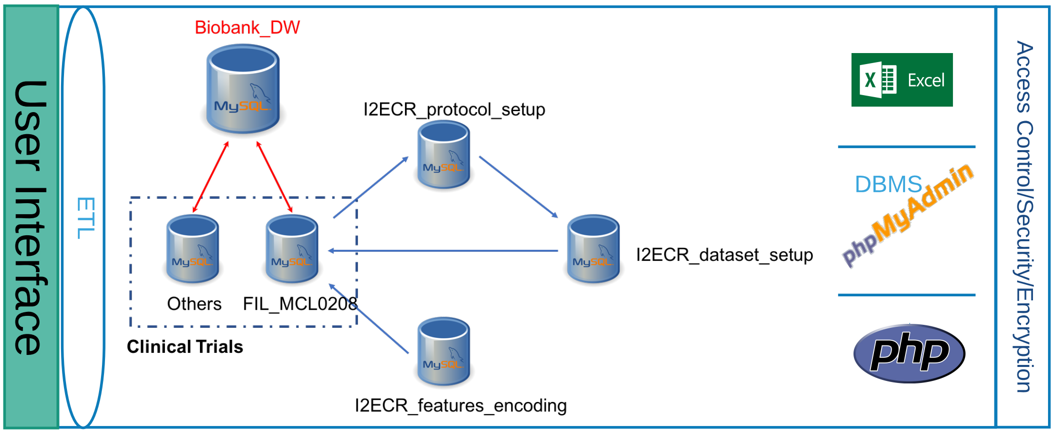


Figure 47: I2ECR implementation. This architecture takes in account integration with external biobank management system.

5.2.2 DW reporting: semi-automatic cohort's selection and data visualization.

I2ECR project founds on strong cooperation between biomedical engineers and clinicians. Both data-processing and data analysis are computed with Matlab® querying data to a DW via DBMS. Therefore, Principal Investigators (PIs) must request to biomedical engineers to extract a dataset from I2ECR, providing requirements on expected datasets (figure 13). I2ECR software (SW) main goal is to semi-automatize this process allowing clinical researchers to directly retrieve data from I2ECR.

Currently, I2ECR SW exploits a connection between a web-server (PHP web server provided from XAMPP) to a DBMS to fix quality controls on data stored in DWs (i) and to export reports to provide to stakeholders (ii). To do that, the web server requires DW metadata to DBMS. Below are listed some queries of metadata referred to FIL.MCL0208 DW:

```
// Query of the List of DBs
$query = "SELECT SCHEMA_NAME FROM information_schema.SCHEMATA WHERE SCHEMA_NAME
LIKE '%FIL%'";
$databases = mysqli_query($dbconn, $query);
$rs = mysqli_fetch_assoc($databases);

// Query of the List of Tables
$query = "SELECT * FROM INFORMATION_SCHEMA.TABLES WHERE TABLE_SCHEMA ='$db_name'
AND (TABLE_NAME NOT LIKE '%toxicity%') AND (TABLE_NAME NOT LIKE 'Cell_Type') AND
(TABLE_NAME NOT LIKE 'Center') AND (TABLE_NAME NOT LIKE 'Drug') AND (TABLE_NAME NOT
LIKE '%Nodal%') AND (TABLE_NAME NOT LIKE '%Biopsy%');
```

```
$tables = mysqli_query($dbconn, $query);
$rs = mysqli_fetch_assoc($tables);

// Query of the List of Attributes
$arr[$i]=$rs['TABLE_NAME'];
$query_attr = "SELECT * FROM INFORMATION_SCHEMA.COLUMNS WHERE TABLE_SCHEMA
='$db_name' AND TABLE_NAME = '$arr[$i]'";
$attributes = mysqli_query($dbconn, $query_attr);
$rs_attr = mysqli_fetch_assoc($attributes);
```

Semi-automatic cohort selection is the natural development step for I2ECR SW. The idea is to refers to successful commercial experiences as Trinetx®[29], Trialx®[30] and Cohort Explorer[31] by Oracle®. Herein a list of cohort extracted manually via I2ECR:

- Subjects with extra-nodal localization on CT.
- Subjects with ATM and TP53 mutations but with CXCR4 not mutated.
- Subjects with positive MRD at baseline, but with negative MRD at restaging 1.

Moreover, front-end tools to provide customized data shall be useful for stakeholders (Han, Kamber, and Pei 2012). Commercial solutions for data-processing and visualization are available in market (e.g. Knowledge base web by OHDSI®, IBM Big SQL®, HDP® by Hortonworks[32]). However, open-access tools are best solutions for public institutions with restricted budgets (Hadhoop by Apache®). Free JavaScript libraries for data visualization are several. Among these Google Charts, NVD3ad chartlist.js provide several .js functions to implement API.

5.2.3 Data Mining: classification methods to impute missing values and for knowledge discovery

In order to discover hidden patterns behind data, data-mining scope includes classification theories. Classification is "a form of data analysis that extracts models describing important data classes" (Han, Kamber, and Pei 2012). Data classification

[29] https://www.trinetx.com/

[30] http://trialx.com/

[31] http://www.oracle.com/us/industries/health-sciences/hs-cohort-explorer-ds-1672120.pdf

[32] https://it.hortonworks.com/about-us/

consists of a learning phase on data followed by a classification step. Learning may be supervised if classes to which tuples belong are known or unsupervised if not. In case of unsupervised learning classification is also defined clustering. Therefore, cluster analysis is the process to partition a set of observations in subsets. According to Han et al., classification methods can be dived in 3 main categories:

- Basic Classification Methods: Information Gain, naïve Bayesian.
- Advanced Classification Methods – K-Nearest Neighbor, Neural Networks: back-propagation and feed-forward propagation, Multi-layer perceptron. Genetic Algorithms and Rough set approaches.
- Clustering methods – Partitioning: K-means, Hierarchical: decision trees, Density-based and grid-based methods, Neural Networks: Self-organized maps.

Moreover, missing values issue strongly affects clinical dataset implicating a strong decrease of dataset accuracy (Zaccaria, Rosati, et al. 2017). Besides data-mining problems, classification methods can be used to lead with that issue, representing an innovative alternative to statistical-based methods, such as conditioned and unconditioned imputation (Horton and Lipsitz 2001). In recent years, the use of ML techniques was explored for MVs imputation (García-Laencina, Sancho-Gómez, and Figueiras-Vidal 2010). A comparison between statistical and classification methods in dealing with missing values and evaluating the performance capability to recover "missingness" in a translational dataset can be exploited. Imputation of missing values must overcome dependency by specific dataset, hence performance of classifiers implemented for imputation must result independent by oncology diseases.

5.3 I2ECR: from Precise Medicine to Big-data?

Precision medicine in onco-hematologic field needs high professional skills to integrate 'omics' to clinical knowledge (Servant et al. 2014; Bellazzi et al. 2011). The main goal is to boost on high-technological innovations to discover unknown phenomenon that causes the affection of diseases and to develop new pharmacology solutions (Shi-kai et al. 2015). Clinical research is changing traditional scopes (Mirnezami, Nicholson, and Darzi 2012) moving from the classical application of epidemiology principles to high-tech data-driven projects (Harris et al. 2009).

Clinical trials and cohort retrospective studies are designed to produce new knowledge about a certain disease (Gholap et al. 2015) with the aim to collect several data from huge cohorts of patients. Moreover, national and international institutions are dramatically investing in pan-communitarian projects to reach wide sample sizes more representative of population as possible to boost on precision medicine (Meric-Bernstam et al. 2013; Siebert et al. 2015; Murphy et al. 2006).

Electronic Case Report Forms (eCRFs) are developing in even more comprehensive platforms capable to collect clinical and biological data (Harris et al. 2009). Electronic Health Records (EHR) have great potential in data collection (Downing et al. 2009; Joyner and Paneth 2015), and if well-integrated with eCRF platforms can be a big tool for hospital institutions that wants to invest on clinical research. Moreover, post-NGS development in molecular biology aims to seek for new biomarkers (i) and clinical outcomes (ii) with high-throughput technologies (Pirooznia et al. 2008). Again, clinical trials projects involving bioinformatics concepts are growing (Servant et al. 2014). Therefore, we are looking to a complex mosaic of solutions that needs to be ruled and standardized. Biomedical engineers can contribute in merging all those disciplines exploiting their professional background (Bellazzi et al. 2011) to develop tools for analysis on clinical-data (Chen and Asch 2017).

Unfortunately, eCRFs are not integrated to single EHRs entities (Figure 3) and every hospital is technically isolated from each other. Lack of integration is more critical if we consider molecular biology and genomics, so the risk is that translational medicine may become a challenge with "inflated expectations" (Chen and Asch 2017; Obermeyer and Emanuel 2016).

This manuscript describes a project that crosses between medical sciences and biomedical engineering. The key point is "medical data" and their broad scope concerned to. In last 20 years, data management in clinical research became a resident discipline and needs for standardized rules (ICH 1996; ICH Harmonised Tripartite Guideline 1996).

I2ECR is an Integrated and Intelligent Environment for Clinical Research where clinical and omics data stand together for clinical use and for generation of new clinical knowledge. I2ECR is adapted to MCL0208 phase III trial, which is a translational trial with several clinical prognostic factors associated to treatment data, biological assessment of disease (MRD - Minimal Residue Disease) and ancillary studies as Pharmacogenomics, Pathology, Mutational Analysis and GEP (Gene Expression Profile). I2ECR main objectives are:

- to propose an integration project on clinical and molecular data.
- to be a dynamic repository of data congruency quality rules.
- to provide to clinical stake-holders a platform from where they can easily design statistical and data mining analysis.

To achieve those goals, I2ECR covers several disciplines from clinical to biomedical engineering scopes (figure 48). This project retrieves data from **translational clinical trials** end reshapes their organization and management applying **data warehousing** concepts. Cleaned and standardized data are set to **query and data reporting** for clinical investigations by **programming** tools. The environment allowed to setup 2 data-analysis projects for clinical investigations. However, data analysis field covers more sophisticated **data-mining** techniques that may be implemented in parallel projects.

I2ECR project doesn't stop, because can be adapted to further clinical trials of different phases and focused on different diseases. Finally, **Bio-banking** management systems will be integrated to I2ECR boosting on potential impact of this tool in terms of translational clinical research.

The ambitious idea is to setup a Big-Data project (Viceconti, Hunter, and Hose 2015) to allow clinical stakeholders to compute meta-analysis within both clinical and molecular domains (Figure 1).

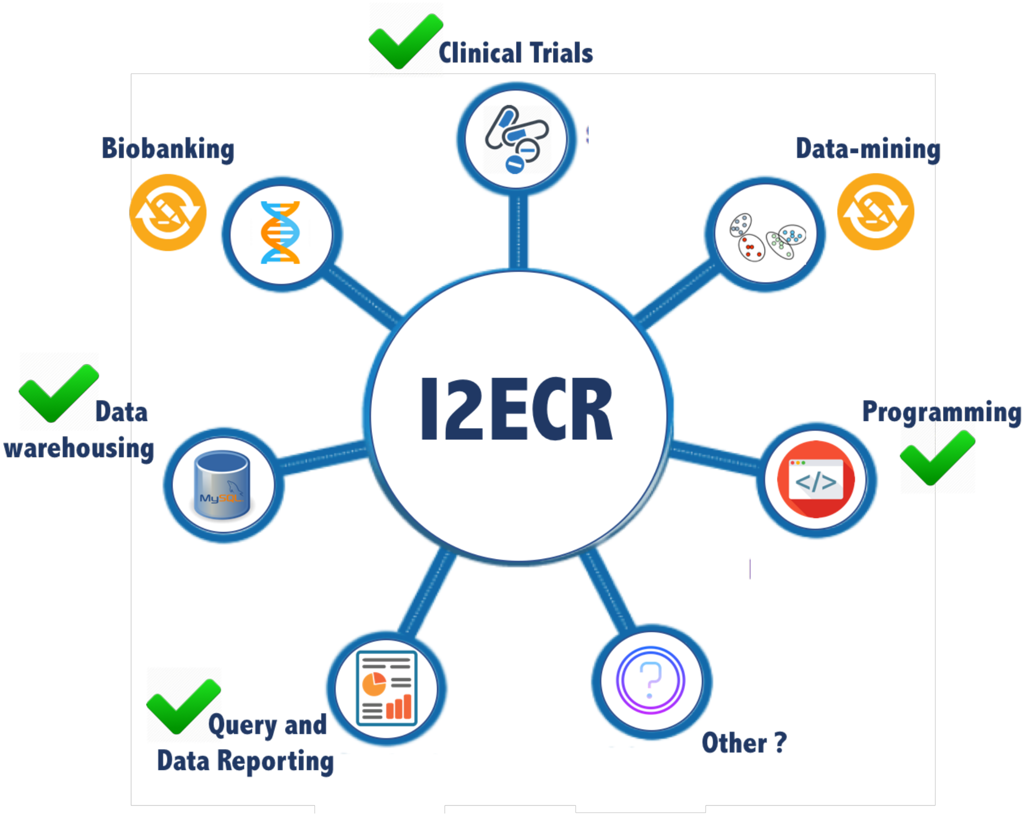


Figure 48: I2ECR final schema.

References

Aalst, W. M. P. Van Der, and Kees Max van Hee. 2004. *Workflow Management: Models, Methods, and Systems*. https://doi.org/10.1.1.95.9284.

Adams, Samantha A., and Carolyn Petersen. 2016. "Precision Medicine: Opportunities, Possibilities, and Challenges for Patients and Providers." *Journal of the American Medical Informatics Association* 23 (4):787–90. https://doi.org/10.1093/jamia/ocv215.

Akkaoui, Zineb El, Esteban Zimànyi, Jose-Norberto Mazón, Juan Trujillo, Zhuolun Zhang, Sufen Wang, Zineb El Akkaoui, et al. 2011. "A Model-Driven Framework for ETL Process Development." *Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP (DOLAP'11)*, 45–52. https://doi.org/10.1145/2064676.2064685.

Alamyar, Eltaf, Véronique Giudicelli, Shuo Li, Patrice Duroux, and Marie Paule Lefranc. 2012. "IMGT/Highv-Quest: The IMGT® Web Portal for Immunoglobulin (IG) or Antibody and T Cell Receptor (TR) Analysis from NGS High Throughput and Deep Sequencing." *Immunome Research* 8 (1):1–15. https://doi.org/10.1038/nmat3328.

Alexander, Lorraine K, Kristen Ricchetti-masterson, and Karin B Yeatts. n.d. "ERIC Notebook," 1–6.

Andersen, Jeppe Ragnar, Inger Byrjalsen, Asger Bihlet, Faidra Kalakou, Hans Christian Hoeck, Gitte Hansen, Henrik Bo Hansen, Morten Asser Karsdal, and Bente Juel Riis. 2015. "Impact of Source Data Verification on Data Quality in Clinical Trials: An Empirical Post Hoc Analysis of Three Phase 3 Randomized Clinical Trials." *British Journal of Clinical Pharmacology* 79 (4):660–68. https://doi.org/10.1111/bcp.12531.

Andersen, Niels S., Lone B. Pedersen, Anna Laurell, Erkki Elonen, Arne Kolstad, Anne Marie Boesen, Lars M. Pedersen, et al. 2009. "Pre-Emptive Treatment with Rituximab of Molecular Relapse after Autologous Stem Cell Transplantation in Mantle Cell Lymphoma." *Journal of Clinical Oncology* 27 (26):4365–70. https://doi.org/10.1200/JCO.2008.21.3116.

Andreu-Perez, J., C.C.Y. Poon, R.D. Merrifield, S.T.C. Wong, and G.Z. Yang. 2015. “Big Data for Health.” *Juornal of Biomedical Health Information* 19 (4):1193–1208.

Atzeni, Ceri, Paraboschi. n.d. *Basi Di Dati.*

Beck, Tim, Sirisha Gollapudi, S??ren Brunak, Norbert Graf, Heinz U. Lemke, Debasis Dash, Iain Buchan, Carlos D??iaz, Ferran Sanz, and Anthony J. Brookes. 2012. “Knowledge Engineering for Health: A New Discipline Required to Bridge the ‘ICT Gap’ between Research and Healthcare.” *Human Mutation* 33 (5):797–802. https://doi.org/10.1002/humu.22066.

Bellazzi, R., M. Diomidous, I. N. Sarkar, K. Takabayashi, a. Ziegler, and a. T. McCray. 2011. “Data Analysis and Data Mining: Current Issues in Biomedical Informatics.” *Methods of Information in Medicine* 50 (6):536–44. https://doi.org/10.3414/ME11-06-0002.

Bergsma, Wicher. 2013. “A Bias-Correction for Cram??r’s V and Tschuprow’s T.” *Journal of the Korean Statistical Society* 42 (3):323–28. https://doi.org/10.1016/j.jkss.2012.10.002.

Bruce D. Cheson , Sandra J. Horning , Bertr Coiffier , Margaret A. Shipp , Richard I. Fisher , Joseph M. Connors , T. Andrew Lister , Julie Vose , Antonio Grillo-López , Anton Hagenbeek , Fernando Cabanillas , Donald Klippensten , Wolfgang Hiddemann , Ron, George P. Canellos. 1999. “Report of an International Workshop to Standardize Response Criteria for Non-Hodgkin’s Lymphomas.” *Journal of Clinical Oncology* 17 (4):1244–1244.

Chan, YH. 2003. “Series of 16 Aticles on Basic Statistics for Doctors.” *Singamore Med J* 44:498–503.

Chen, Jonathan H, and Steven M Asch. 2017. “Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations.” *Machine Learning and Prediction in Medicine N Engl J Med* 37626:2507–9. https://doi.org/10.1056/NEJMp1702071.

Cortelazzo, Sergio, Maurizio Martelli, Marco Ladetto, Simone Ferrero, Giovannino Ciccone, Andrea Evangelista, Michael Mian, et al. 2015. “HIGH DOSE SEQUENTIALCHEMOTHERAPY WITH RITUXIMAB AND ASCT AS FIRST LINE THERAPY IN ADULT MCL PATIENTS: CLINICAL AND MOLECULARRESPONSE OF THE MCL0208 TRIAL, A FIL STUDY.” In .

Dancey, Janet. 2012. “Genomics, Personalized Medicine and Cancer Practice.”

Clinical Biochemistry 45 (6). Elsevier B.V.:379–81. https://doi.org/10.1016/j.clinbiochem.2012.03.003.

Daniele Bochiccio, Stefano Mostarda. 2015. *HTML5 Con CSS E Javascipt*.

De, Sourabh. 2011. "Hybrid Approaches to Clinical Trial Monitoring: Practical Alternatives to 100% Source Data Verification." *Perspectives in Clinical Research* 2 (3):100–104. https://doi.org/10.4103/2229-3485.83226.

Dernoncourt, Franck, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. "De-Identification of Patient Notes with Recurrent Neural Networks." *Journal of the American Medical Informatics Association* 24 (3):596–606. https://doi.org/10.1093/jamia/ocw156.

Diehr Diane E.Harris, Tamara B.Duxbury, AndrewSiscovick, DavidRossi, Michelle, PaulaBild. 1998. "Body Mass Index and Mortality in Nonsmoking Older Adults: The Cardiovascular Health Study." *American Journal of Public Health* 88 (4):623–29. http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=pbh&AN=1137998&site=ehost-live&custid=s4121186.

Downing, Gregory J, Scott N Boyle, Kristin M Brinner, and Jerome A Osheroff. 2009. "Information Management to Enable Personalized Medicine: Stakeholder Roles in Building Clinical Decision Support." *BMC Medical Informatics and Decision Making* 9:44. https://doi.org/http://dx.doi.org/10.1186/1472-6947-9-44.

Drandi, Daniela, Jimenez, Luigia Monitillo, Barbero Barbero, Marina Ruggeri, Barbara Mantoan, Elisa Genuardi, et al. 2016. "First Comparison between Multicolor Flow Cytometry and Droplet Digital PCR for Tumor Burden Quantification at Baseline in Mantle Cell Lymphoma."

Dreyling, M., S. Ferrero, and O. Hermine. 2014. "How to Manage Mantle Cell Lymphoma." *Leukemia* 28 (11). Nature Publishing Group:2117–30. https://doi.org/10.1038/leu.2014.171.

Dreyling, M., C. Geisler, O. Hermine, H. C. Kluin-Nelemans, S. Le Gouill, S. Rule, O. Shpilberg, J. Walewski, and M. Ladetto. 2014. "Newly Diagnosed and Relapsed Mantle Cell Lymphoma: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up." *Annals of Oncology* 25 (August):iii83-iii92. https://doi.org/10.1093/annonc/mdu264.

Eubank, Michael H., David M. Hyman, Amritha D. Kanakamedala, Stuart M. Gardos, Jonathan M. Wills, and Peter D. Stetson. 2016. "Automated Eligibility Screening and Monitoring for Genotype-Driven Precision Oncology Trials."

Journal of the American Medical Informatics Association 23 (4):777–81. https://doi.org/10.1093/jamia/ocw020.

F. Kauffmann, A. Cambon-Thomsen. 2008. “Tracing Biological Collections.” *JAMA : The Journal of the American Medical Association* 299 (19):2316–18. https://doi.org/10.1001/jama.299.19.2316.

Fahrudin, T.M., I. Syarif, and A.R. Barakbah. 2017. “Feature Selection Algorithm Using Information Gain Based Clustering for Supporting the Treatment Process of Breast Cancer.” *2016 International Conference on Informatics and Computing, ICIC 2016*, no. Icic. https://doi.org/10.1109/IAC.2016.7905680.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. “The KDD Process for Extracting Useful Knowledge from Volumes of Data.” *Communications of the ACM* 39 (11):27–34. https://doi.org/10.1145/240455.240464.

Ferrero, Simone, Daniela Drandi, Barbara Mantoan, Paola Ghione, Paola Omedè, and Marco Ladetto. 2011. “Minimal Residual Disease Detection in Lymphoma and Multiple Myeloma: Impact on Therapeutic Paradigms.” *Hematology Oncology*. https://doi.org/10.1002/hon.989.

Fiscon. 2014. “Alzheimer ’ S Disease Patients Classification through EEG Signals Processing,” 0–7.

Fletcher, Robert H.; Fletcher, Suzanne W.; Fletcher, Grant S. 2003. “Introduction.” In *Clinical Epidemiology: The Essentials, 5th Edition*, 1–27. https://doi.org/10.1016/S0422-9894(08)71343-6.

Francis S. Collins, M.D., Ph.D., and Harold Varmus, M.D. 2010. “A New Initiative on Precision Medicine.” *Perspective* 363 (1):793–95. https://doi.org/10.1056/NEJMp1002530.

García-Laencina, Pedro J., José-Luis Sancho-Gómez, and Aníbal R. Figueiras-Vidal. 2010. “Pattern Classification with Missing Data: A Review.” *Neural Computing and Applications* 19 (2):263–82. https://doi.org/10.1007/s00521-009-0295-6.

Ghielmini, M., U. Vitolo, E. Kimby, S. Montoto, J. Walewski, M. Pfreundschuh, M. Federico, et al. 2013. “ESMO Guidelines Consensus Conference on Malignant Lymphoma 2011 Part 1: Diffuse Large B-Cell Lymphoma (DLBCL), Follicular Lymphoma (FL) and Chronic Lymphocytic Leukemia (CLL).” *Annals of Oncology* 24 (3):561–76. https://doi.org/10.1093/annonc/mds517.

Gholap, Jay, Vandana P. Janeja, Yelena Yesha, Raghu Chintalapati, Harsh Marwaha, and Kunal Modi. 2015. "Collaborative Data Mining for Clinical Trial Analytics." *Proceedings - 2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015*, 1063–69. https://doi.org/10.1109/BIBM.2015.7359829.

Gökbuget, Nicola, Michael Kneba, Thorsten Raff, Heiko Trautmann, Claus-rainer Bartram, Rainer Fietkau, Mathias Freund, et al. 2014. "Display a Poor Prognosis and Are Candidates for Stem Cell Transplantation and Targeted Therapies Adult Patients with Acute Lymphoblastic Leukemia and Molecular Failure Display a Poor Prognosis and Are Candidates for Stem Cell Transplantation and Targeted " 120 (9):1868–76. https://doi.org/10.1182/blood-2011-09-377713.

Halkidi, Maria, Michalis Vazirgiannis, and Yannis Batistakis. 2000. "Quality Scheme Assessment in the Clustering Process." *Principles of Data Mining and Knowledge Discovery*, 265–76. https://doi.org/10.1007/3-540-45372-5_26.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. *Data Mining: Concepts and Techniques. San Francisco, CA, ltd: Morgan Kaufmann*. https://doi.org/10.1016/B978-0-12-381479-1.00001-0.

Harris, Paul A., Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, and Jose G. Conde. 2009. "Research Electronic Data Capture (REDCap)-A Metadata-Driven Methodology and Workflow Process for Providing Translational Research Informatics Support." *Journal of Biomedical Informatics* 42 (2). Elsevier Inc.:377–81. https://doi.org/10.1016/j.jbi.2008.08.010.

Hewett, Rattikorn, and Phongphun Kijsanayothin. 2008. "Tumor Classification Ranking from Microarray Data." *BMC Genomics* 9 Suppl 2:S21. https://doi.org/10.1186/1471-2164-9-S2-S21.

Horton, Nicholas J, and Stuart R Lipsitz. 2001. "Multiple Imputation in Practice." *The American Statistician* 55 (3):244–54. https://doi.org/10.1198/000313001317098266.

Hoster, Eva, Martin Dreyling, Wolfram Klapper, Christian Gisselbrecht, Achiel Van Hoof, Hanneke C. Kluin-Nelemans, Michael Pfreundschuh, et al. 2008. "A New Prognostic Index (MIPI) for Patients with Advanced-Stage Mantle Cell Lymphoma." *Blood* 111 (2):558–65. https://doi.org/10.1182/blood-2007-06-095331.

Hoster, Eva, Andreas Rosenwald, Françoise Berger, Heinz Wolfram Bernd, Sylvia Hartmann, Christoph Loddenkemper, Thomas F.E. Barth, et al. 2016.

"Prognostic Value of Ki-67 Index, Cytology, and Growth Pattern in Mantle-Cell Lymphoma: Results from Randomized Trials of the European Mantle Cell Lymphoma Network." *Journal of Clinical Oncology* 34 (12):1386–94. https://doi.org/10.1200/JCO.2015.63.8387.

ICH. 1996. "Guidance for Industry: E6 Good Clinical Practice." *US Department of Health and Human Services*, no. April:63. https://doi.org/10.1056/NEJMp1012246.

ICH Harmonised Tripartite Guideline. 1996. "Guideline for Good Clinical Practice E6(R1)." *ICH Harmonised Tripartite Guideline* 1996 (4):i-53. https://doi.org/10.1056/NEJMp1012246.

Inmon, W.H. William H. 2005. *Building the Data Warehouse*.

Jares, Pedro, Dolors Colomer, and Elias Campo. 2012. "Review Series Molecular Pathogenesis of Mantle Cell Lymphoma." *Jci* 122 (10). https://doi.org/10.1172/JCI61272.3416.

Joyner, Michael J, and Nigel Paneth. 2015. "Seven Questions for PersonalizedMedicine." *Jama* 55905:2015–16. https://doi.org/10.1001/jama.2015.7725.Conflict.

Jun, Gyuchan Thomas, James Ward, Z O E Morris, and John Clarkson. 2009. "Health Care Process Modelling : Which Method When ?" 21 (3):214–24.

Kerber. 1992. "ChiMerge: Discretization of Numeric Attributes." In *AAAI*, 123–28. https://dl.acm.org/citation.cfm?id=1867154.

Kim, Dokyoon, Ruowang Li, Anastasia Lucas, Shefali S Verma, Scott M Dudek, and Marylyn D Ritchie. 2016. "Using Knowledge-Driven Genomic Interactions for Multi-Omics Data Analysis: Metadimensional Models for Predicting Clinical Outcomes in Ovarian Carcinoma." *Journal of the American Medical Informatics Association* 24 (December 2016):ocw165. https://doi.org/10.1093/jamia/ocw165.

Krzywinski, M. et al. 2009. "Circos: An Information Aesthetic for Comparative Genomics." *Genome Res* 19 (604):1639–45. https://doi.org/10.1101/gr.092759.109.19.

Luo, Jie, Xing Rong Guo, Xiang Jun Tang, Xu Yong Sun, Zhuo Shun Yang, Yong Zhang, Long Jun Dai, and Garth L. Warnock. 2014. "Intravital Biobank and Personalized Cancer Therapy: The Correlation with Omics." *International Journal of Cancer* 135 (7):1511–16. https://doi.org/10.1002/ijc.28632.

Mallick, Rajiv, Brajesh Kumar Lal, and Claire Daugherty. 2017. "Relationship between Patient-Reported Symptoms, Limitations in Daily Activities, and Psychological Impact in Varicose Veins." *Journal of Vascular Surgery: Venous and Lymphatic Disorders* 5 (2). Society for Vascular Surgery:224–37. https://doi.org/10.1016/j.jvsv.2016.11.004.

Mendelsohn, John, Ulrik Ringborg, and Richard Schilsky. 2015. "Innovative Clinical Trials for Development of Personalized Cancer Medicine." *Molecular Oncology* 9 (5):933–34. https://doi.org/10.1016/j.molonc.2015.02.013.

Meric-Bernstam, Funda, Carol Farhangfar, John Mendelsohn, and Gordon B. Mills. 2013. "Building a Personalized Medicine Infrastructure at a Major Cancer Center." *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 31 (15):1849–57. https://doi.org/10.1200/JCO.2012.45.3043.

Mirnezami, Reza, Jeremy Nicholson, and Ara Darzi. 2012. "Preparing for Precision Medicine." *New England Journal of Medicine* 366 (6):489–91. https://doi.org/10.1056/NEJMp1114866.

Mosteller. 1987. "Simplified Calculation of Body-Surface Area." *New England Journal of Medicine* 317 (1098). https://doi.org/10.1056/NEJM198710223171717.

Murphy, S N, M E Mendis, D A Berkowitz, I Kohane, and H C Chueh. 2006. "Integration of Clinical and Genetic Data in the i2b2 Architecture." *AMIA Annual Symposium Proceedings* 2006 (2):2006.

Niazkhani, Zahra, Habibollah Pirnejad, Heleen van der Sijs, and Jos Aarts. 2011. "Evaluating the Medication Process in the Context of CPOE Use: The Significance of Working around the System." *International Journal of Medical Informatics* 80 (7). Elsevier Ireland Ltd:490–506. https://doi.org/10.1016/j.ijmedinf.2011.03.009.

Obermeyer, Ziad, and Ezekiel J Emanuel. 2016. "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine." *The New England Journal of Medicine* 375 (13):1216–19. https://doi.org/10.1056/NEJMp1606181.

Pirooznia, Mehdi, Ping Gong, Jack Y Yang, Mary Qu Yang, Edward J Perkins, and Youping Deng. 2008. "ILOOP--a Web Application for Two-Channel Microarray Interwoven Loop Design." *BMC Genomics* 9 Suppl 2:S11. https://doi.org/10.1186/1471-2164-9-S2-S11.

Pott, C., Macintyre, E., Delfau-Larue, M., Ribrag, V., Unterhalt, M., Kneba, M., Hiddemann, W., Dreyling, M., Hermine, O., & Hoster, E. 2014. "MRD

Eradication Should Be the Therapeutic Goal in Mantle Cell Lymphoma and May Enable Tailored Treatment Approaches: Results of the Intergroup Trials of the European MCL Network." *Blood* 124 (21):147. http://www.bloodjournal.org/content/124/21/147.

Pott, Christiane, Eva Hoster, Kheira Beldjord, Vahid Asnafi, Anne Plonquet, Reiner Siebert, Evelyne Callet-bauchu, et al. 2010. "HAL Archives Ouvertes – France Author Manuscript Molecular Remission Is an Independent Predictor of Clinical Outcome in Patients with Mantle Cell Lymphoma after Combined Immunochemotherapy : A European MCL Intergroup Study HAL-AO Author Manuscript." *Blood* 115 (16):3215–24. https://doi.org/10.1182/blood-2009-06-230250.Molecular.

Ravi, Daniele, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang Zhong Yang. 2017. "Deep Learning for Health Informatics." *IEEE Journal of Biomedical and Health Informatics* 21 (1):4–21. https://doi.org/10.1109/JBHI.2016.2636665.

Rosati, Samanta, Gabriella Balestra, and Filippo Molinari. 2014. "Feature Extraction by Quick Reduct Algorythm: Assessing the Neurovascular Pattern of Migraine Sufferers from NIRS Signals." In *Machine Learning in Healthcare Informatics*, edited by Springer Berlin Heidelberg.

S. Rosati, K. Meiburger, G. Balestra, U.R. Archarya, F. Molinari. 2016. "Carotid Wall Measurement and Assessment Based on Pixel-Based and Local Texture Descritors." *Journal of Mechanics Medical Biology*.

Servant, Nicolas, Julien Roméjon, Pierre Gestraud, Philippe La Rosa, Georges Lucotte, Séverine Lair, Virginie Bernard, et al. 2014. "Bioinformatics for Precision Medicine in Oncology: Principles and Application to the SHIVA Clinical Trial." *Frontiers in Genetics* 5 (MAY):1–16. https://doi.org/10.3389/fgene.2014.00152.

Shen, Qiang and, and Alexios Chouchoulas. 2000. "Modular Approach to Generating Fuzzy Rules with Reduced Attributes for the Monitoring of Complex Systems." *Eng. Appl. Artif. Intell.* 13 (3):263–78.

Shi-kai, Y A N, L I U Run-hui, J I N Hui-zi, L I U Xin-ru, Y E Ji, and Shan Lei. 2015. "' Omics ' in Pharmaceutical Research : Overview , Applications , Challenges , and Future Perspectives" 13 (2011):3–21.

Shi, Peng. 2016. "Automated Quantitative Image Analysis of Hematoxylin-Eosin Staining Slides in Lymphoma Based on Hierarchical Kmeans Clustering," 99–104. https://doi.org/10.1109/ITME.2016.190.

Siebert, Uwe, Beate Jahn, Ursula Rochau, Petra Schnell-Inderst, Agnes Kisser, Theresa Hunger, Gaby Sroczynski, et al. 2015. “Oncotyrol - Center for Personalized Cancer Medicine: Methods and Applications of Health Technology Assessment and Outcomes Research.” *Zeitschrift Fur Evidenz, Fortbildung Und Qualitat Im Gesundheitswesen* 109 (4–5):330–40. https://doi.org/10.1016/j.zefq.2015.06.012.

Strunk, Dirk, Eva Rohde, Gerhard Lanzer, and Werner Linkesch. 2005. “Transplantation and Cellular Engineering.” *Transfusion* 45 (March):315–26. https://doi.org/10.1111/j.1537-2995.2006.00675.x.

Swerdlow, Steven H., and Michael E. Williams. 2002. “From Centrocytic to Mantle Cell Lymphoma: A Clinicopathologic and Molecular Review of 3 Decades.” *Human Pathology* 33 (1):7–20. https://doi.org/10.1053/hupa.2002.30221.

Tiemann, Markus, Carsten Schrader, Wolfram Klapper, Martin H. Dreyling, Elias Campo, Andrew Norton, Francoise Berger, et al. 2005. “Histopathology, Cell Proliferation Indices and Clinical Outcome in 304 Patients with Mantle Cell Lymphoma (MCL): A Clinicopathological Study from the European MCL Network.” *British Journal of Haematology* 131 (1):29–38. https://doi.org/10.1111/j.1365-2141.2005.05716.x.

Vegliante, Maria Carmela, Jara Palomero, Patricia Pérez-Galán, Gaël Roué, Giancarlo Castellano, Alba Navarro, Guillem Clot, et al. 2013. “SOX11 Regulates PAX5 Expression and Blocks Terminal B-Cell Differentiation in Aggressive Mantle Cell Lymphoma.” *Blood* 121 (12):2175–85. https://doi.org/10.1182/blood-2012-06-438937.

Viceconti, Marco, Peter Hunter, and D Hose. 2015. “Big Data, Big Knowledge: Big Data for Personalised Healthcare.” *IEEE Journal of Biomedical and Health Informatics* 2194 (c):1–1. https://doi.org/10.1109/JBHI.2015.2406883.

Vose, Julie M. 2015. “Mantle Cell Lymphoma: 2015 Update on Diagnosis, Risk-Stratification, and Clinical Management.” *American Journal of Hematology* 90 (8):739–45. https://doi.org/10.1002/ajh.24094.

Wang, Guohua, Yadong Wang, Weixing Feng, Xin Wang, Jack Y Yang, Yuming Zhao, Yue Wang, and Yunlong Liu. 2008. “Transcription Factor and microRNA Regulation in Androgen-Dependent and -Independent Prostate Cancer Cells.” *BMC Genomics* 9 Suppl 2:S22. https://doi.org/10.1186/1471-2164-9-S2-S22.

Warner, Jeremy L., Matthew J. Rioth, Kenneth D. Mandl, Joshua C. Mandel, David A. Kreda, Isaac S. Kohane, Daniel Carbone, et al. 2016. “SMART Precision Cancer Medicine: A FHIR-Based App to Provide Genomic Information at the

Point of Care." *Journal of the American Medical Informatics Association* 23 (4):701–10. https://doi.org/10.1093/jamia/ocw015.

Woollen, Stan W. 2010. "Data Quality and the Origin of ALCOA." *The Compass*. http://www.southernsqa.org/newsletters/Summer10.DataQuality.pdf.

Zaccaria, Gian Maria, Simone Ferrero, Andrea Evangelista, Samanta Rosati, Cristina Castagneri, Marco Ghislieri, Barbero Daniela, et al. 2017. "Delphi, a Data Warehouse to Discover Associations between Variables in Clinical Trials: Application to the Fondazione Italiana Linfomi (FIL) MCL0208 Phase III Trial." *Blood* 130 (Suppl 1):3451 LP-3451. http://www.bloodjournal.org/content/130/Suppl_1/3451.abstract.

Zaccaria, Gian Maria, Samanta Rosati, Cristina Castagneri, Simone Ferrero, Marco Ladetto, Mario Boccadoro, and Gabriella Balestra. 2017. "Data Quality Improvement of a Multicenter Clinical Trial Dataset." *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, (accepted for publication).

Zapletal, Eric, Nicolas Rodon, Natalia Grabar, and Patrice Degoulet. 2010. "Methodology of Integration of a Clinical Data Warehouse with a Clinical Information System: The HEGP Case." *Studies in Health Technology and Informatics* 160 (PART 1):193–97. https://doi.org/10.3233/978-1-60750-588-4-193.