UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA

DIP. INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO IN BIOINGEGNERIA E BIOINFORMATICA

CICLO XXIX - 2013/2016

LONGITUDINAL DATA ANALYTICS FOR CLINICAL DECISION SUPPORT IN TYPE 2 DIABETES

PhD THESIS BY

ARIANNA DAGLIATI

Advisor:

Lucia Sacchi

PhD Program Chair:

Riccardo Bellazzi

Table of contents

List of figures

CHAPTER 1

1 Introduction

Type 2 Diabetes Mellitus (T2DM) is assuming epidemic proportions, which will progressively worsen as the population ages. Managing T2DM is a complex task, such complexity being embodied in long clinical histories, lasting longer than 10 years and characterized by substantial variability in the type and frequency of clinical events that are manifested across the population and within a single patient history. In addition, the pathology itself entails a number of complications and comorbidities. These issues suggest the difficulty in managing T2DM chronic patients (World Health Organization 2016).

A major source of complexity in the management of T2DM patients arises from events such as hospital admissions, follow-up clinic visits, laboratory tests, and therapy changes., During these events, patients are often treated by many different health professionals such as general practitioners (GPs), physicians in specialist centers and those working in hospitals for acute events, as well as pharmacists Moreover, These events are stored in different data repositories using different formats and occurring in temporal sequences that represent the patient careflow (Quaglini et al. 2001; Quaglini et al. 2000).

Although these data are distributed in sources such as the Electronic Health Record (EHR) and, Administrative Data Warehouses (DW), new data management technologies are able to gather and merge them, and consequently enable researchers and other to access a huge quantity of complex data for the interpretation and exploitation of these data for a management of chronic diseases. The application of longitudinal analysis and careflow discovery to these data enable the recognition of hidden temporal patterns, population stratification and cohorts' identification, and phenotypes definition.

Temporal data analysis and careflow mining techniques can automatically detect the most frequent patterns and careflows from routinely collected administrative and/or clinical data. Once identified, the enacted careflows might be used for comparison with clinical protocols to check their adherence to best practices, but they can be also exploited to identify different sub-groups of individuals in large cohorts of patients. This means that these temporal data mining techniques can be used as a type of electronic phenotyping., which has been defined as the detection of computable phenotypes through query to EHRs and clinical data repository using specific data elements and logical expressions (Rachel & Michelle 2014).

Currently, the automated identification of complex careflows from clinical data represents a major unmet clinical need and what is currently missing is a holistic framework able to address the following tasks:

- To gather, integrate and handle heterogeneous temporal multivariate data;

- To leverage on longitudinal data to: (i) develop and apply advanced temporal data mining methods to perform risk stratification, and (ii) develop new predictive models to deliver calibrated risk prediction;
- Deliver the discoveries of these models through efficient visual analytics for decision support.

The research questions that motivate this work are focused on the problem of discovering novel yet relevant knowledge in longitudinal clinical data:

- How does one perform meaningful analytics on such data and derive the right knowledge for novel insights on the disease?
- How does one deliver insights for decision support to improve disease management and care delivery?

Clinical guidelines and health care protocols are well-established tools used to improve and standardize health care services. Nevertheless, in the absence of effective technology-based solutions to automatically extract frequent patterns and careflows, it is often impossible to measure their implementation. Patients' management processes can be improved through an overall system that integrates longitudinal heterogeneous data, and implements temporal data mining methods that illustrate the evolution of the disease and the individual and population variability. The detection of temporal patterns makes possible to reconstruct clinical pathways and forecast the complications that might arise during the process of care, to identify interesting clusters of patients with similar care histories and re-assess their risk profiles accordingly. The identification of healthcare pathways through methods derived from temporal and careflow mining research can be used for Decision Support.

These facts suggest the need to investigate novel methods for improving the clinical decision support in T2DM and the utility of creating an analytics methodological framework. This is the overall goal of this dissertation, which was successfully retained completing these three specific aims:

- To implement a system that integrates a large amount of unstructured and structured data from heterogeneous sources;
- To extend longitudinal analytic approaches to enable recognition of trending patterns and enhance temporal electronic phenotypes description;
- To create an expansion of existing methods for clinical decision support that is based on a more complete and easily understood description of patient health status.

At a high-level this dissertation is organized around three main facets that described below; a detailed description of the structure is provided in Section 1.5.

- Methods for collecting data from various sources and organizing and integrating them with clinical knowledge and patients' information, to enhance health related decision/action and bring into clinical practice a new paradigm for a more coordinated care.
- Development of longitudinal analytics methods, and their implementation as modular software tools, to allow users to both analyze their patient population and studying individual patients to better specify their risk profiles.

- Scientific findings and algorithms integration in a Decision Support Tool for T2DM, structured as a Dashboard. It is shown how the implemented system provides clinicians with a new approach for the follow up of a chronic population, moving towards a strategy more focused on the continuous follow up and prevention of worsening than in treatment of acute events.

1.1 Background

In the first section the basic physiopathology of DM is reviewed. The focus of the paragraphs is T2DM, which is a more prevalent disease and poses significant challenges in Clinical Decision Support in a long term clinical context. The second section of the background introduce the European Union project that founded this research program.

1.1.1 Diabetes Mellitus and its Complications

The global prevalence of Diabetes Mellitus has risen dramatically over the past two decades, with more than 400 million adults currently living with diabetes (World Health Organization 2016). Based on current trends, the International Diabetes Federation estimates that by 2030 there will be about 550 million people suffering from DM, making it a leading cause of morbidity and mortality for the foreseeable future (Mathers & Loncar 2006; Shaw et al. 2010; Whiting et al. 2011).

Diabetes Mellitus (DM) is a chronic metabolic syndrome which encompasses a range of metabolic disorders that share the phenotype of hyperglycemia over a prolonged time (Alberti & Zimmet 1998). In healthy adults, blood glucose concentrations are normally maintained within a relatively narrow range, approximately of 70-110 mg/dL in the fasting state (Nelson & Cox 2013). Glucose homoeostasis reflects a balance between hepatic glucose production and peripheral glucose uptake. Maintenance of the glucose levels within the physiological range is accomplished by a complex regulatory system based on two types of dynamic mechanisms: the hormonal system, which consists of a balance between insulin and counter regulatory hormones (glucagon, cortisol, epinephrine), and the neural mechanism, which consists of the activation of messages issued from glucose sensors of various organs (Cryer 2008). Insulin is the most important regulator of this metabolic equilibrium.

Insulin speeds up glucose uptake in the adipose and muscle tissue, where the glucose transportation across the cellular membrane is largely mediated by insulin-sensitive transport proteins. Insulin enables the conversion of glucose to storage compounds, via glycogenesis to produce glycogen and via lipogenesis for the formation of triacylglycerol (Triplitt 2012). The release of insulin is regulated by the level of glucose in the blood (Nelson & Cox 2013). In patients suffering from diabetes, this regulatory system for glycaemia mediated by insulin is deviant.

Distinct types of DM are caused by a complex interplay of genetic and environmental factors. Depending on the etiology of DM, hyperglycemia may present as a result of reduced insulin production, decreased glucose utilization and/or increased glucose production (Fauci et al. 2008). DM is classified on the basis of the pathogenic process that leads to hyperglycemia as type 1 and type 2. Although type 1 diabetes is an important clinical problem with a substantial burden on the children afflicted with this disease, this work focuses on type 2.

Diabetes chronic complications. Metabolic dysregulation associated with DM causes secondary pathophysiologic alterations in multiple organ systems. The complications associated to chronic DM can be categorized in macrovascular and microvascular. The mechanism(s) by which it leads to such diverse cellular and organ dysfunctions is still unknown, though several theories have been proposed (Fauci et al. 2008). The association between increased levels of glycated hemoglobin, which is the reference parameter to measure average plasma glucose concentration over prolonged periods of time, and a higher risk of developing complications has been shown by landmark studies such as the United Kingdom Prospective Diabetes Study (UKPDS) and the Kumamoto study of T2DM patients (Fauci et al. 2008). The findings of these studies emphasized the importance of intensive glycemic control in all forms of DM and of early diagnosis and strict blood pressure control in T2DM to prevent the adverse effects of the complications of diabetes.

Cardiovascular disease A marked increase in macro-cardiovascular disease including peripheral artery disease, coronary heart disease, congestive heart failure, myocardial infarction, stroke, and sudden death has been associated to DM (Kannel et al. 1974; Kannel & McGee 1979). Cardiovascular disease is the major cause of morbidity and mortality for individuals with DM and the largest contributor to the direct and indirect costs of DM (American Diabetes Association 2014). The increase in cardiovascular disease appears to relate to the synergism of hyperglycemia with other cardiovascular risk factors (Fauci et al. 2008): the common conditions coexisting with T2DM (e.g. hypertension, dyslipidemia, obesity) are clear risk factors for cardiovascular disease. Cardiovascular complications in DM can be reduced by improved glycemic control; however, the glycemic goal for diabetic patients remains undefined and cardiovascular disease outcomes are less clearly impacted by hyperglycemia levels or intensity of glycemic control than microvascular complications (Fauci et al. 2008).

Diabetic Retinopathy is a microvascular complication of both T1DM and T2DM, with prevalence strictly related to the duration of diabetes (American Diabetes Association 2014). Diabetic retinopathy is the leading cause of non-reversible blindness in the adult population in North America and Europe. Diabetic retinopathy is the result of microvascular retinal changes and is classified into two stages: non-proliferative and proliferative. Non-proliferative diabetic retinopathy usually appears late in the first decade or early in the second decade of the disease and is marked by vascular micro-aneurysms and can be asymptomatic for a long time. Non-proliferative retinopathy can evolve into proliferative retinopathy when newly formed vessels appear near the optic nerve and/or macula and rupture easily, leading to vitreous hemorrhage, fibrosis, and ultimately retinal detachment (Fauci et al. 2008).

Diabetic Nephropathy occurs in 20-40% of patients with DM and is the single leading cause of end-stage renal disease (American Diabetes Association 2014). Although renal disease has several known risk factors (like hyperglycemia and hypertension), the genetic component behind development and progression of nephropathy may play a significant role.

Nephropathy is a marker of increased cardiovascular morbidity and mortality (Alberti & Zimmet 1998). The mechanisms by which chronic hyperglycemia leads to nephropathy is still currently investigated (Penno et al. 2013). Nephropathy onset involves hemodynamic alterations in the renal microcirculation related to glomerular hyper filtration or hyper perfusion, increased capillary pressure and structural changes in the glomerulus (Fauci et al. 2008). These alterations ultimately result in impaired renal functionality. Persistent albuminuria in the range of 30–299 mg/dL in a 24

h collection has been shown to be an early stage indicator of diabetic nephropathy in T1DM and a marker for development of nephropathy in T2DM (American Diabetes Association 2014).

Diabetic Neuropathy occur in about 50% of individuals with long-standing T1DM and T2DM and are heterogeneous pathologies with diverse clinical manifestations (American Diabetes Association 2014). The most prevalent neuropathies are (i) chronic sensorimotor diabetic peripheral neuropathy, which leads to neuropathic pain, loss of limb sensitivity and loss of muscular strength, and (ii) autonomic neuropathy, which leads to development of cardiovascular symptoms, gastroparesis and bladder-emptying abnormalities (Fauci et al. 2008). These conditions are thought to result from diabetic microvascular injury involving small blood vessels that supply nerves in addition to macro vascular conditions that can culminate in nerve damage. The presence of cardiovascular disease, elevated triglycerides, and hypertension is often associated with diabetic peripheral neuropathy. In presence of diabetic neuropathy, specific treatment for the underlying nerve damage is currently not available, other than glycemic control, which may modestly slow progression (American Diabetes Association 2014). Approximately 15% of individuals with T2DM develop a foot ulcer and a significant subset ultimately undergoes amputation (Fauci et al. 2008). The reasons for the increased incidence of these disorders involve the interaction of several pathogenic factors: peripheral sensory neuropathy (which interferes with normal protective mechanisms and causes abnormal weight bearing while walking and subsequent formation of callus or ulceration), motor neuropathy (which results in abnormal foot muscle mechanisms and structural changes in the foot) and peripheral arterial disease and poor wound healing (which impede the resolution of minor breaks in the skin, favoring infections).

Diabetes is also associated with an increased incidence of infections (especially dermatological, dental and urinary) and skin diseases. Chronic cerebrovascular disturbances and liver disease have also been related with diabetes.

Type 2 Diabetes Mellitus (T2DM), also referred to as non-insulin dependent diabetes or adult-onset diabetes, is the most common form of DM, accounting for nearly the 90% of cases of DM (WHO 2006). It refers to a heterogeneous group of disorders characterized by variable degrees of insulin resistance (diminished tissue responses to insulin) and insulin deficiency (inadequate insulin secretion for glucose load) (Triplitt 2012). The majority of patients with this form are obese, and obesity itself causes or aggravates insulin resistance.

In the early stages of the disorder, glucose tolerance remains near-normal, despite insulin resistance, because the β-cells compensate by increasing insulin output. As insulin resistance and compensatory hyper-insulinemia progress, the pancreatic islets in certain individuals are unable to sustain the hyper-insulinemic state. Phases characterized by elevations in blood glucose levels are important to consider as they can lead to the onset of T2DM. These phases follow: ***Impaired Fasting Glucose (IFG)*** (American Diabetes Association 2014), with fasting plasma glucose level between 110 mg/dL and 126 mg/dL, and ***Impaired Glucose Tolerance (IGT)*** (American Diabetes Association 2014), defined as a postprandial glycaemia in the 140–199 mg/dL range. IFG and IGT represent intermediate stages between normal glucose tolerance and final diabetes and identify subjects with increased risk of developing T2DM, who form an important target group for interventions aimed at preventing diabetes (Tuomilehto et al. 2001). A further decline in insulin secretion and an increase in hepatic glucose production lead to overt diabetes with fasting hyperglycemia. Ultimately, β-cells failure ensues.

T2DM results from the interaction between a genetic predisposition and behavioral and environmental risk factors (Tuomilehto et al. 2001). The risk of developing T2DM increases with age, obesity, and lack of physical activity. It occurs more frequently in individuals with hypertension or dyslipidemia and its frequency varies in different ethnic subgroups. It is often associated with strong familial, likely genetic, predisposition; however, the genetics of this form of diabetes are complex and not yet clearly defined. This form of diabetes is frequently undiagnosed for many years because the hyperglycemia is often not severe enough to provoke noticeable symptoms of diabetes. T2DM patients are at increased risk of developing microvascular (mainly stroke and acute coronary syndromes) and microvascular (mainly retinopathy, neuropathy, nephropathy, and limb ischemia) complications; diagnosis is often made from associated complications or incidentally through an abnormal blood or urine glucose test. At least initially, and often throughout their lifetime, these individuals do not need insulin treatment to survive. Insulin sensitivity may be increased by weight reduction, increased physical activity, and/or pharmacological treatment of hyperglycemia (e.g. with metformin) but cannot be restored to normal (American Diabetes Association 2014).

From the medical and health care point of view the evolution of T2DM is complex. Figure 1 shows a schematic timeline of the different stages in the disease history. This work is focused on the study of the disease after the diagnosis, as highlighted in the figure. Different actors intervene and the status of patients evolves over time as characterized by different healthcare professionals who are in charge of their treatment, by different facilities they access, and by the complications of the disease they experience.

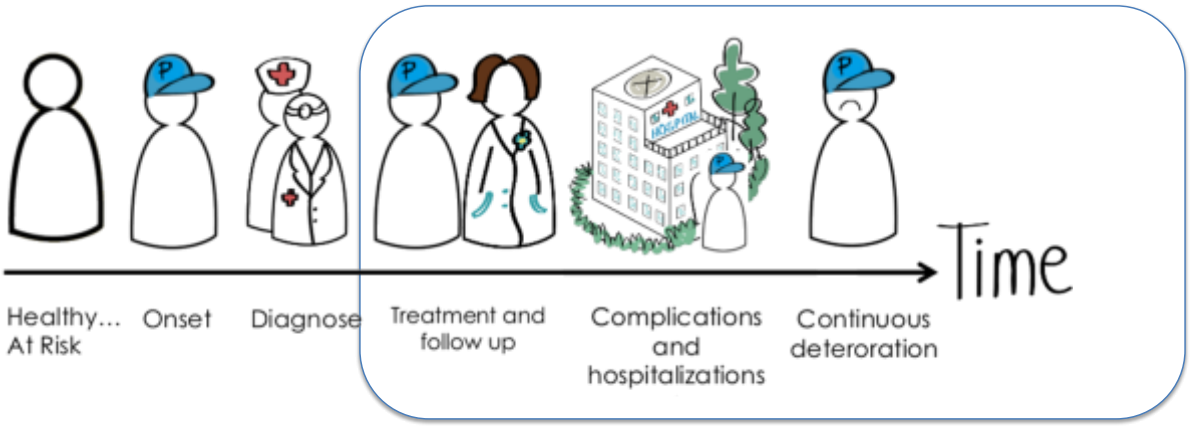


Figure 1 The evolution of diabetic disease

T2DM is characterized by (i) high prevalence, (ii) complex pathophysiology, (iii) the availability of large amount of, structured and unstructured, data, and (iv) the difficulty to extract knowledge from these data in a temporal context. The next section describes a European initiative to understand factors that influence the development and progression of the T2DM diseases.

1.1.2 The context of this research program: the MOSAIC Project.

A large part of the presented research has been performed within the MOSAIC project. MOSAIC (Models and simulation techniques for discovering diabetes influence factors) is an EU-funded

project carried out within the 7th Framework Program [http://www.mosaicproject.eu/]. The project was conducted by a consortium of European partners, including: Medtronic Ibérica S.A., Università degli Studi di Padova, Universidad Politécnica de Madrid, Università degli Studi di Pavia, IRCCS Fondazione Salvatore Maugeri (FSM), Lund University, Folkhalsan Research Center, National Technical University of Athens. Partners are shown in Figure 2.

Figure 2 Mosaic Partners

The MOSAIC project was devoted to the development of tools for the diagnosis and management of T2DM based on Big Data, mathematical models, and algorithms; enabling the improvement in the characterization of patients suffering from this metabolic disorder and to help evaluate the risk of developing T2DM and related complications were among the main objectives of the project.

Within the project, multiple clinical databases were made available to the consortium as a result of the activities conducted by its members in previous projects and studies. On this basis, one of the main tasks of MOSAIC was to improve data mining techniques to better understand the mechanisms underlying the evolution of diabetes through the analysis of individual patient histories, temporal events and behavioral factors.

The activities of two of the scientific partners (University of Padova and University of Pavia) were mainly focused on developing predictive models and innovative methods to:

- Determine external factors influencing the onset and progression of diabetes, in order to improve diagnosis and prediction.

- Define novel methods and procedures for the continuous stratification of the population at risk of diabetes and of developing related complications.
- Improve existing modelling and prediction tools in diabetes management.

Mosaic scientific and technological partners contribute to the project goals on the basis of different sources of data and focusing on different aspects and stages of the disease. University of Pavia efforts were devoted to analyze clinical data, gathered from hospitals EHRs and administrative systems, and to build models to depict the events diabetic patients undergoes after the diagnosis, during the progressive deterioration of his/her health status.

The overall goal of the MOSAIC project was to integrate the developed models in a decision support system (DSS) and to enhance decision-making in clinical care. The ***MOSAIC system*** made available health technology devices and tools to improve the characterization of diabetic states and deliver a clinical decision support system to diabetologists and healthcare managers, facilitating the interpretation and visualization of the data and enabling a comprehensive understanding of the information by the healthcare professionals.

In the last phase of the project, the University of Pavia collaborated with the Universidad Politécninca de Madrid to develop a DSS intended for use in many healthcare settings (e.g. hospitals, clinical centers and health care agencies). The implementation of this DSS required analyzing several aspects, like the unmet needs that final users expected to be covered by the proposed solutions and how to efficiently turn the developed models into DSS tools for T2DM management.

Several issues were addressed with a holistic approach able to contextualize the design of a system that turns computerized modelling techniques into IT tools that support decision makers in T2DM. A set of user Interfaces and Interactions flows, defined as the tools that support the creation of visual models for user interactions with software and trace their responses behaviors (Brambilla & Fraternali 2015) in order to maximize Usability and User Experience (UX) were designed. The design of preliminary prototypes and mock-ups followed the state-of-the-art methodologies for usability and UX, and included an exhaustive review by UX experts through focus groups.

1.2 Specific Aims

The central scientific research objective of this dissertation is to implement state of the art and develop novel Temporal Data Mining Methods able (i) to recognize subtle changes in time (patterns, careflows) and suggest that a patient condition is worsening and (ii) to identify tailored sub cohorts of patients who meet specific conditions that are relevant to clinical actions.

To reach the scientific objectives of the research, the following specific aims were addressed.

Aim1. Gathering and Integration of data from heterogeneous sources

This aim focused on the need to collect and merge multivariate temporal data efficiently from heterogeneous sources into a common data model, which will be the basis of the temporal analysis

and will issue the challenge of handling with patients and population variability through three sub-aims:

- To identify sources of data that are relevant for key clinical questions and decisions in monitoring and treating T2DM patients;
- To develop a data gathering strategy);
- To create a common data model to guide the integration of data and the execution of the transformation and preprocessing actions necessary for assessing data quality.

Aim 2. Data analytics

This aim focused on the design and implementation of temporal based methods to view the evolution of disease both at patient and population level, and to dynamically stratify the population on the basis of the risk of developing complications and worsening metabolic control. To meet the multifaceted aspects of the T2DM disease, this aim includes two sub-aims:

Aim 2a - Longitudinal data analytics focused on

- Implementation of temporal data mining methods to abstract higher-level concepts from heterogeneous data, and give them a qualitative, common, representation in time in order to jointly analyze them;
- Development of innovative careflow mining methods to handle longitudinal data and to recognize patterns in time, and to use these results for the identification of well-tailored sub cohorts of patients (temporal phenotypes);
- Study the influence of exposure factors (like drugs purchasing patterns) in the evolution of the disease of specific subjects tracing their behavioral temporal patterns.

Aim 2b - Risk models for complications and metabolic control variations focused on

- Development of calibrated multivariate risk prediction models for T2DM associated complications;
- Implementation of state of the art multivariate longitudinal models to predict and assess metabolic control evolution.

Aim 3. Models integration for Clinical Decision Support.

This aim focused on the integration of methods developed within Aim 2 in a Clinical Decision Support System able to support clinicians in managing both single patients and large cohorts through a synthesis of distributed information. The developed system is able:

- To collect a copy of clinical data and merge them with other sources for providing research results into clinical practice;
- To broadly encompass the clinical decision chain, leveraging on several strategies, from visual analytics to computer based models integration.

Figure 3 represents the executive framework of this work and it illustrates how the presented aims were related and how each aim was composed of several technological and methodological concepts that guided the entire research activities. The gathering and integration of heterogeneous

data sources (Aim 1) represented the basis for the group of knowledge discovery activities (Aim 2). Aim 2 was twofold, as the implemented methods served to monitor and manage the entire cohort of patients through the identification of groups of subjects following the same disease trajectories (Careflow Mining and Longitudinal Models), but also to discover the path that a single patient followed in terms of his/her risk of developing complications (Predictive Models) and exposure patterns (Visual Analytics). To deliver clinical decision support (Aim 3) the software tools implemented the analytical methods and models and applied them to the data as collected in a structured format, as represented by the common data model.

The executive framework schema is presented at the beginning of each methodological chapter, where Aims relations are in specifically illustrated, and it is discussed how the accomplishment reached within one Aim influenced the realization of the others.

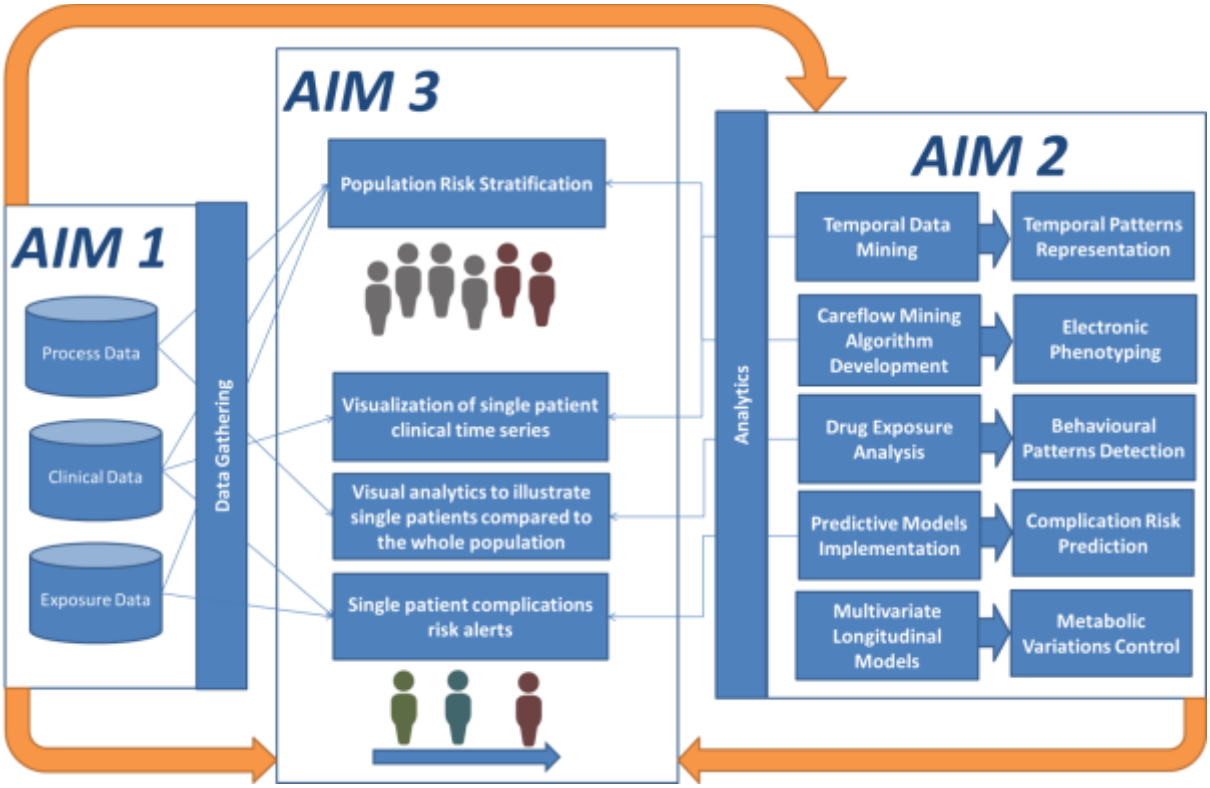


Figure 3 Executive Framework. Aims, their interconnections (blue arrows) and dependencies (orange arrows).

1.3 Overview of research

The concept of ***precision medicine***, refers to the objective of providing the best available care for each individual (Bahcall 2015; Ashley 2015; Collins & Varmus 2015). A concept central to precision medicine is that stratifying patients into subsets with a common biological basis of disease is most likely to improve medical management and, consequently, the quality of care (Disease, Committee on A Framework for Developing a New Taxonomy of Disease, Board of Life Sciences, Division of Earth and Life Sciences 2011). Clinically meaningful disease sub- classifications can be recognized as distinct *phenotypes*. The next step is the consideration that a phenotype might encompass not only disease abnormalities and complications, but also the response of a certain

patient to a specific treatment or even to the way the treatment is delivered. For this reason, an essential requirement to achieving precision medicine is the systematic application of methodologies able to identify clinical deviations in morphology, physiology but also in medical behaviors.(Robinson 2012; Frey et al. 2014).

The adoption of electronic health record (EHR) in the past years made available a unique source of clinical information for research. The EHR can be used to extract and interpret clinical information, to automatically support clinical research and improve quality of care. Specifically, ***EHR-based phenotyping*** uses data captured in the delivery of healthcare to identify individuals or cohorts with conditions or events relevant to clinical studies. (Newton et al. 2013; Richesson, Hammond, et al. 2013; Hripcsak & Albers 2013). Some distinguishing novelties in respect to this approach can be represented by (i) considering the temporal nature of the data, explicitly including not only clinical information from EHR, but merging them with process information from administrative databases; (ii) exploit the methods used to retrieve electronic phenotypes not only for research purposes in clinical studies but also to support health care in its daily activities through decision support.

The novel, essential directions for researchers in the ***medical informatics*** field have been recently redefined to be effectively employed in clinical practice (Tenenbaum et al. 2016). Specifically, the conceptual approach of the well-known "data, information, and knowledge" continuum has been reconsidered as the so called "Learning Healthcare System Cycle", where healthcare practice and research should be part of a unique and synergic process, as shown in Figure 4, adapted from (Tenenbaum et al. 2016). The cycle core is precision medicine and it starts considering the clinical history of the patients, including their current status, their previous history, and possible future scenarios. The first main novelty of this approach is to emphasize the synergy of clinical practice and research in the generation of data and knowledge.

The role of informatics, which enables each transition of the cycle, is to provide the right tools to turn data into information, and information into knowledge, understanding data relations, retrieving and understating patterns. The other novel step in this cycle is the deployment of knowledge to guide individual behavior and to inform patient care.

(Tenenbaum et al. 2016) identified several key areas that should be the foundation for progress in informatics research. In the following they are discussed and reorganized to illustrate the purposes of this research and to be matched with its aims.

Care Informs Research (from Data to Knowledge).

Starting from clinical practice, data should be collected from multiple sources (EHR, Administrative DW) possibly integrated with other outcomes, like environmental exposure data, and jointly analyzed. While some of these data (e.g. diagnosis codes collected in EHRs both for clinical and billing purposes) are increasingly being collected in the same warehouse within the same organization, one of the main issues to face is the functional heterogeneity of these data and associated information. In this context, informatics applications should be focused on specific key areas.

> ***Key Area 1 - Data integration, standardization and exchange***: facilitate data gathering procedures, terminologies standardization and information exchange.

Key Area 2 – New Phenotype definition: define novel phenotypes, which are computationally manageable and well simulate disease behaviors in space and time.

Research Informs Care (from Knowledge to Action).

Clinical decision making requires the consolidation of precision medicine knowledge and the development of decision support system.

Key Area 3 – Consolidation of clinical decisions: enrich patient data with actionable information that can be exploited by care givers at the point of care. In this context, informatics applications should: (i) build a comprehensive knowledge base containing information about disease subtype, diagnosis, therapies and (ii) develop tools that support custom workflows, novel analytics, data visualization and data aggregation.

Figure 4 Learning Health Care System Cycle. Healthcare practice and Research as part of a unique and synergic process centered on Precision Medicine.

In the following is illustrated how the research Aims are mapped into the Key Areas defined in the Learning Health Care System. Figure 5 represents the Conceptual Framework of this research program. It shows how the implemented framework and the methods adopted within this work meets current medical informatics challenges and are an innovative contribution to support T2DM management. More specifically it shows how Electronic Phenotyping can be integrated into a CDSS on the basis of the Learning Health System model.

In Figure 5 ***Aim 1*** related tasks are represented in blue, and are all the research activities associated to gathering data from healthcare delivery actions. ***Aim 2*** related tasks are represented in red, and are all the research efforts spent to develop analytical methods for knowledge discovery from longitudinal data. ***Aim 3*** related tasks are represented in green, and are all the informatics activities for the implementation of a CDSS, which, once integrated in the clinical settings, can transfers research finding into medical and management activities.

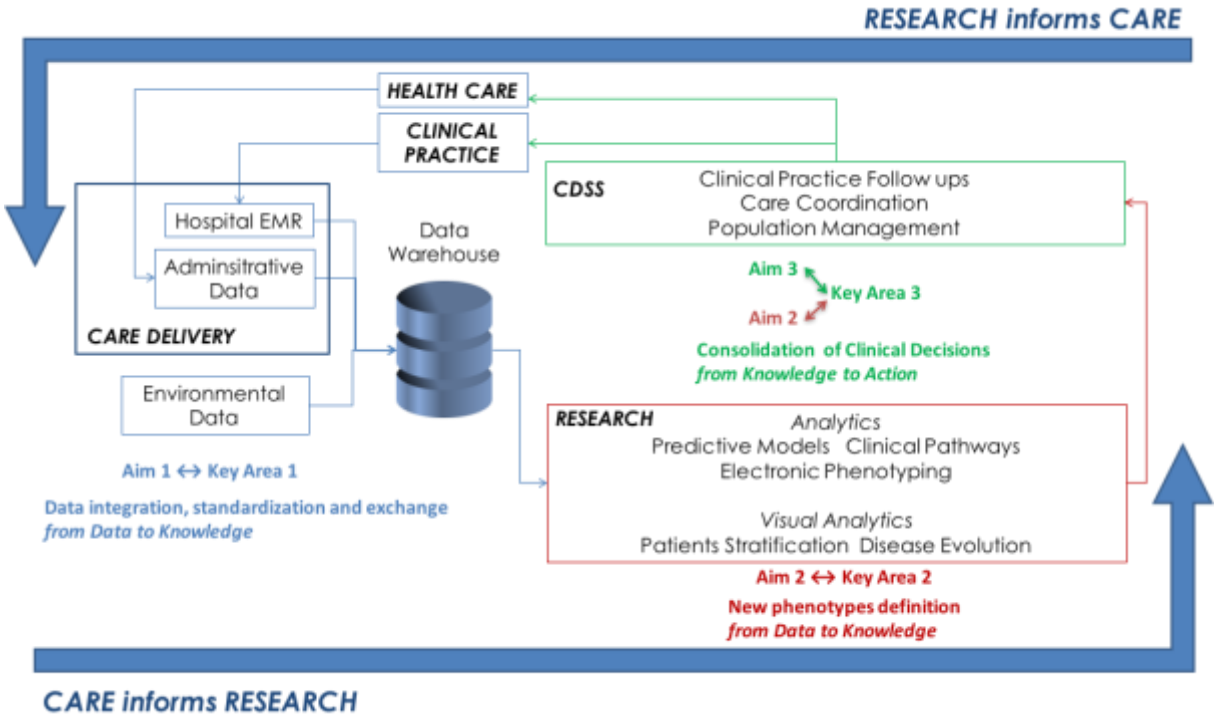


Figure 5 Conceptual framework designed and adopted for addressing the goals of the research program. It illustrates how is possible to leverage on the Learning Health Care Cylce to build a CDSS based on Electronic Phenotyping.

Key Area 1 states the importance of data integration, standardization and exchange to support the translation of the information gathered during clinical practice into relevant knowledge that supports new scientific evidences. To accomplish ***Aim 1,*** data were gathered from heterogeneous sources and contexts such as hospital EHRs and local health care agency information systems). The data warehouse structure has been implemented to support clinical practice, as it is the basis for the decision support system.

Key Area 2 focuses on defining new phenotypes which are computationally manageable and able to simulate disease behaviors. This research embarked upon analytical methods devoted to mine temporal exposure patterns able to represents T2DM disease evolution through longitudinal and sparse data. A new careflow mining algorithm for performing electronic phenotyping was developed and multivariate predictive models for the risk of T2DM complications were implemented.

Key Area 3 encompass two main topics. The first one is related to the needs of informatics applications able to deliver comprehensive knowledge base containing information about disease subtype, diagnosis, therapies. It can be defined as the joining link between ***Aim 2*** and ***Aim 3***. Research findings were, in fact, delivered into clinical practice through the implementation of several visual analytics strategies that enabled patient stratification based on the results of the developed temporal models. The second statement is about the development of tools that support custom workflows, novel analytics, data visualization and data aggregation. It is strictly connected to ***Aim 3***. It regards all the activities that allowed closing the Learning Health Care System cycle while exporting the knowledge derived from the research activities into real clinical actions. The developed mining algorithms and the results of the longitudinal data analyses and the predictive

models were integrated into a Clinical Decision Support System (CDSS) to make effective the concept of "Research informs Care". The developed system implements several functionalities and use cases devoted to the consolidation of clinical decision during follow up, and for population management.

1.4 Novel Contributions of this work

Specific candidate's contributions, together with the main achievements and novelties within medical informatics are introduced in the following. The core methodological efforts were focused on the development and implementation of innovative temporal and careflow mining algorithms, their application to the available data set and integration in the final system. To successfully accomplish this objective, it was necessary to study and formalize the clinical setting, to efficiently retrieve and organize the available data and, finally, to combine the clinical and methodological acquired knowledge into the decision support system.

Technical Contributions to Medical Informatics.

Within the strategy designed to gather and organize data from different clinical and administrative institutes, the candidate specifically worked on (i) the definition of the ontology to represent the acquired information, the ontology implementation, and (ii) the design and realization of Extraction, Transformations and Loading (ETL) procedures to fill the data warehouse. Data gathering and integration activities were performed through state of the art technologies that will be detailed in the following of this work.

The discussed novelties largely depended on to the background of the studied health care system and hospitals. The scientific findings of this research program are based on an Italian population, though within the MOSAIC project the data model was designed to represent data streams coming from three European Hospitals (Salvatore Maugeri of Pavia - Italy, La Fe in Valencia - Spain, Hippokration in Athens, Greece), and also to integrate data from retrospective studies (Botnia Finnish Study (A.-J. et al. 2010), VIVA (Gabriel-Sánchez et al. 2009) and Opt2mize (Aronson et al. 2014) Spanish studies). These novelties are related to (i) the secondary reuse of data usually collected for reimbursement purposes and never accessed before by medical doctors (e.g. patients' drug purchases within the territorial area managed by the local health administration) and (ii) the retrospective collection of multivariate longitudinal data over more than 10 years of a T2DM population and, consequently, (iii) the definition of a data model able to represent the most important features of an European T2DM chronic population (i.e. diseases profiles, environmental and behavioral factors).

Methodological Contributions to Data Analytics and Temporal Data Mining.

Contributions to data analytics and temporal data mining represent the core of this research program efforts, which were focused on the investigation of several methodological techniques: (i) the implementations of a novel careflow mining algorithm to tease out complex health care patterns, (ii) the mining of drug purchasing data to assess drug consumption and (iii) the implementation of risk predictions models that are based on Bayesian methods and include metabolic phenotypic and lifestyle factors. The candidate designed and implemented these analysis

approaches, validated them on the available data set and deal with the technical predisposition for their integration into the final CDSS tool.

The main novelties are focused on the inclusion of the temporal dimension in electronic phenotyping and on the exploitation of these finding into a CDSS that reproduces the Learning Health Care System Cycle. From the clinical point of view, the obtained results illustrate why the use of longitudinal data is essential for the evaluation and management of T2DM. From a technical and methodological point of view, this work provides an example of an innovative longitudinal data analytic system based on mixture of process and clinical data. The developed careflow mining algorithm tackles issues of process mining approaches (e.g. data sparsity and variability), and readapts sequential and temporal data mining techniques to tackle specific clinical problems in chronic diseases (e.g. patients affected by multiple complication, complex therapies and different care providers). The algorithm provides novel insights by detecting hidden patterns, which are used to profile patients through temporal phenotypes. The developed multivariate risk models enhance currently used risk profiling tools, which are mainly cross sectional, based on population (not patient) profiles and exclusively based on clinical data. The research followed three innovation directions: the description of variable evolution over continuous time, the exploitation of hierarchical models to predict single patient's state variation over time, and the inclusion of pollution information, derived from remote sensing data, in the models.

Clinical and Methodological Contributions to Clinical Decision Making in T2DM.

Regarding the development of the CDSS tool, the candidate actively participated in the design of the system, especially focusing on the best solutions to integrate the developed algorithms and models in the tool, both from clinical and usability perspectives. The candidate, with the support of the project consortium, managed the evaluation of the tool and performed the related statistical analysis. Within this phase, the candidate collaborated with the developers' team and usability experts.

The realization of the CDSS faced several of the challenges previously identified (i.e. custom workflows, novel analytics, data visualization). The system reveals technical innovations in the exploitation of longitudinal clinical, process and environmental data in novel time-based models, and clinical innovations regarding how these analytic approaches have been combined with visual analytics to introduce a new approach for T2DM management in clinical practice. This research program findings contribute to several aspects of the tool novelties. (i) T2DM patients careflow are mined through secondary data reuse. The analytics approach not only takes into account the nature of temporal events but also their interactions along the whole patients' histories. Within the studied clinical context, this feature was not considered before and represent a novel approach to describe followed careflows, clinical actions and patients' behaviors. (ii) The tool represents heterogeneous information, processed through a suite of temporal data mining approaches. T2DM management is facilitated, as the final user is provided with a tool that helps to identify potential clinical misconducts/errors, suboptimal treatments or the need for further diagnostic procedures. (iii) CDSS user's actions to extract new insights in patients' histories are conveyed by a process that integrates methodological novelties into a more standard drill down approach. The user/tool interaction process is structured to guide temporal data exploration. Visual analytics solutions, together with medical knowledge, facilitate the detection of risk profiles.

1.5 Structure of the thesis

The MOSAIC system has been developed to make available new models and tools for improving T2DM patients' care and management. The dissertation presents the methods and implementation efforts to realize a novel system for the management of chronic populations. The outline of this work retraces the components of a CDSS. Figure 6 shows the system key components as described through the chapters of the manuscript. In the introduction of each chapter, motivations, assumptions and objectives at the basis of the technical and methodological choices are detailed.

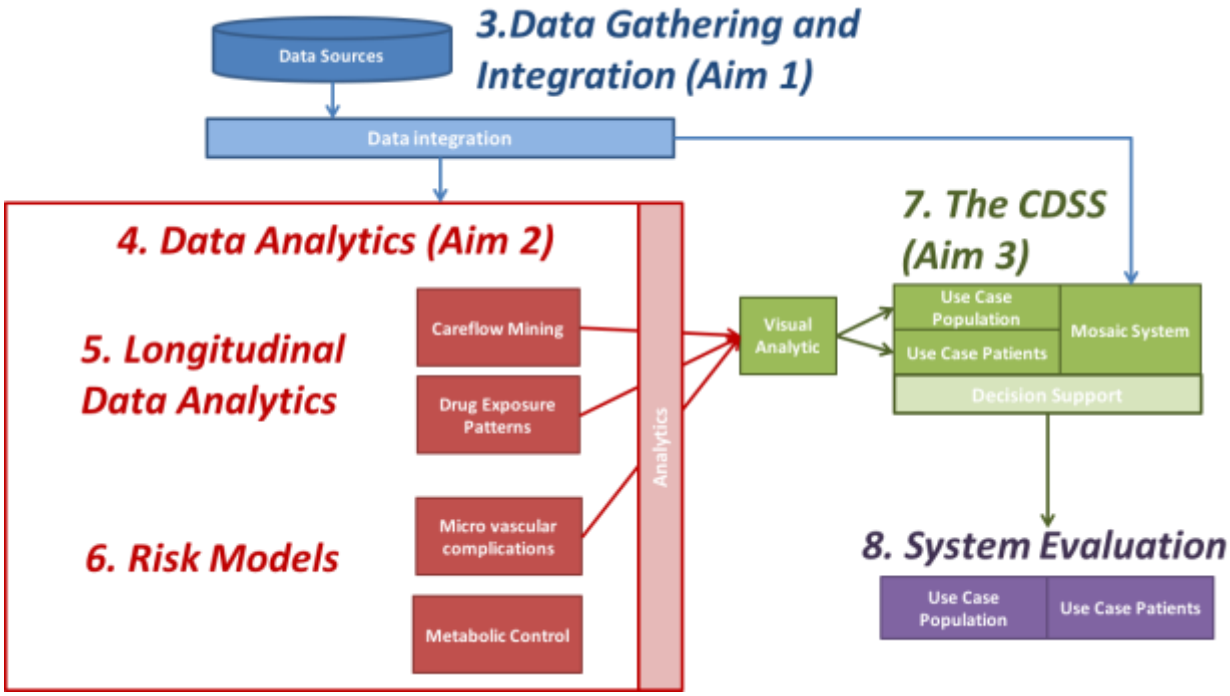


Figure 6 Structural layout illustrates the main system components with references to the chapters where they are described.

This dissertation is structured as follows.

Chapter 2. The Literature Review chapter explores current challenges triggered by the recent availability of large amount of clinical data. From the potentials that novel analytics might provide to support clinical practice through innovative Decision Support tools, the chapter describes the scientific methods needed to extract new relevant knowledge to be used for precision medicine (Temporal and Careflow Mining, Electronic Phenotyping).

Chapter 3. The Data Gathering and Integration (Aim 1) chapter describes the implementation of the data layer, based on the i2b2 framework. The presented solution is aimed at collecting data from heterogeneous sources (hospital EHRs, billing data streams, environmental data) and building a data model to represent T2DM disease evolution events. The developed data infrastructure aggregates duplicate of data produced during multiple healthcare activities and provides the possibility of a secondary reuse of them into clinic, supporting a sidecar approach.

Chapter 4. Chapters Data Analytics (Aim 2), Longitudinal Data Analytics (Aim 2a) and Risk Models for T2DM Complications and Metabolic Control Variations (Aim 2b) illustrate the core of the methodological efforts to find new insights for the characterization of the T2DM disease. The objective is to detect and understand new mechanisms able to explain the progression of the disease through the analysis of individual patient histories, temporal events and behavioral factors.

Chapter 5. The chapter Longitudinal Data Analytics (Aim 2a) describes the development of a new careflow mining algorithm that leverages on the temporal sequence of patients' events to add a novel dimension in electronic phenotyping. The second paragraph of the chapter describes how data gathered from administrative sources about drug purchasing can be exploited to trace patients' behavioral patterns.

Chapters 6. The chapter on Risk Models for T2DM Complications and Metabolic Control Variations (Aim 2b) illustrates the development and validation of predictive models for the risk of developing microvascular complications tailored on the studied Italian population. Moreover, it presents a collection of state of the art longitudinal analysis approaches, based on Bayesian methods and on the integration of remote sensing data, which are used to investigate metabolic variations in time and space.

Chapter 7. The chapter on The clinical decision support system (Aim3) shows how the data model and the analysis methods are finalized to realize a CDSS for reaching a better control of T2DM patient clinical condition over time, guiding personalized interventions and enhance precision medicine. The functionalities made available by the developed algorithms and analysis methods are implemented through suitable visual analytics solutions. To support T2DM professionals and healthcare managers at any stage of the disease care process, the system adopts two solutions, which are focused on user's specific need but also aimed at promoting a more coordinated care of T2DM. One solution is focused on supporting medical doctors' decisions during patients' follow-ups, the other one develops an innovative approach to assist critical managerial decisions in healthcare.

Chapter 8. The chapter on System Evaluation describes the results of the system solutions evaluation once a prototype was introduced into clinical practice and the comments of potential users from focus group discussions.

CHAPTER 2

2 Literature Review

The following paragraphs, summarized in Figure 7, provide a background on the research areas related to the activities carried out during this doctoral project.

Figure 7 Literature Review Contents

2.1 Challenges in the use of Data from Heterogeneous sources for Decision Support.

Due to the exploitation of heterogeneous data from many different sources, their integration in a common data model and their use to address the overall goal of this research program, this work can be seen in the wide context of "Big Data movement" (Gandomi & Haider 2015). Thus, it is worth introducing some aspect the context of Big Data, as it applies to the research program described in this dissertation. A broadly recognized definition of the Big Data states that it is data "whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it" (Harper 2014). This

definition encloses aspects of like the multifactorial nature of these data, and the technological challenges that they imply. Challenges and potentials of Big Data analytics in healthcare have been described in numerous works in the past years (Murdoch & Detsky 2013; Zillner et al. 2014; Halamka 2014; Etheredge 2014; Krumholz 2014).

It is widely known that the term "Big Data" is related to presence of four properties, which are Volume ("large" amounts of data), Variety (diverse formats in which the data are collected) and Velocity (requirements of processing data at a pace fast enough to support decision-making) (Bellazzi 2014; Bellazzi et al. 2015; Peek et al. 2014). Other two peculiarities that are often highlighted include Veracity, which is related to the uncertain nature connected to data of large volumes and very often crudely collected without pre-processing or with poor quality control and Variability, which is associated to the time variance of the data and considers the irregularities in data flows (Katal et al. 2013). What defines Big Data is the simultaneous presence of two or more of these aspects and the consequent requirements to reassess technological architectures used to manage the data, the methods aimed to analyze them and the systems able to leverage on them to support complex activities, like decision support.

Within this research program it was necessary to pre-process, integrate in a suitable framework, and analyze, heterogeneous data streams form at least three different sources (clinical, administrative and environmental). This explains the first characteristic of the collected data set: Variety. Data were collected retrospectively in a 10 years' time window. They were collected by different organizations, through different systems, which possibly changed over time, and they were also sensitive to changes due to the introduction of novel therapies or clinical protocols. This adds the peculiarity of Variability. The quality of data was not always satisfactory and posed several issues that it was necessary to tackle to extract meaningful information and support clinical decisions. These limitations derive from the fact that data collection in a clinical context often takes secondary importance to the delivery of patients' care. These data characteristics faithfully represent the research background, adding Veracity to the data. These facts assimilate this work to the Big Data context.

In (Raghupathi & Raghupathi 2014) the authors discuss current analytics promises in healthcare across various scenarios, from clinical operation to research, to public health. They emphasize the importance of evidence-based medicine and patient profile analytics. Several presented scenarios precisely meet this research efforts, such as (i) the investigation of patient characteristics and clinical outcomes to detect the most cost effective actions to offer the tools needed to improve provider behavior; (ii) advanced analytics implementation for patient profiling, through segmentation and predictive modeling, in order to identify individuals who would benefit from specific preventive actions or life- style changes; (iii) broad scale disease profiling to identify predictive events and support prevention plans; and (iv) creation of novel information streams by aggregating and synthesizing structured and unstructured data (e.g. patient clinical records and administrative data) to provide data and services to third parties.

The interest in the collection and analytics of large and heterogeneous data sources has its roots in industry, nowadays this interest had spread, and it represents one of the current fields where the synergic efforts of research and business could be very effective. This is particularly true as regards Decision Support (Kaltoft et al. 2014; Kohn et al. 2014; Lupse et al. 2014). In (Kohn et al. 2014) the authors define two main fields where researchers should address their efforts to produce

valuable results in data driven decision support: (i) the secondary use of data to create new evidence and glean important insights to make better clinical decision and (ii) the detection of novel correlation from asynchronous events to allow clinicians to promptly identify potential complications, timely adjust treatments or help analyze similar manifestations in clinical diagnoses, as also discussed in (Zhang et al. 2016).

The changes, brought by the availability of new data sources from the health care delivery model, led to growing interest into the development of new data-driven methods for managing quality of care while, which are able to handle these new sources of data, extract knowledge from them and exploit the acquired knowledge for optimizing clinical decisions. To pledge better renewed decision making, and consequent successful clinical outcomes, health care should be ready for the application of real time intelligent risk detection decision support systems using predictive analytic techniques. One of the main focus in using these new information resources for Decision Support has been directed into predictive modeling (Yun Chen & Hui Yang 2014; Moghimi et al. 2013; Wang et al. 2015; Srinivasan Suresh 2014). Other data driven evidences, like similarity measures, should be integrated with physician inputs and experiences to provide the best chance for accurate disease diagnosis through comparative research and, for example, learn about patients affected by multiple chronic conditions.

While multiple chronic conditions patients certainly represent a complex issue for the scientific community to tackle, other works are focused on how novel analytics methods can impact in the management of single specific diseases, like cardiovascular diseases (Rumsfeld et al. 2016), or can be exploited for anesthesia settings (Simpao et al. 2015), neatly depicting the current status of the potential use of big data in complex clinical settings, especially if characterized by disease and treatment heterogeneity. Literature reviews targeted on TD2 are yet quite scarce (Cleveringa et al. 2013). Big data analytics (Rumsfeld et al. 2016) has been recognized to improve outcomes ranging from individual patient care (e.g. for tailored therapy assessments), to leading efficient use of resources in healthcare application or in public health systems. Some of the most well-known current initiatives and networks to support big data research include NIH's BD2K Initiative (Bourne et al. 2015), eMERGE (McCarty et al. 2011) and PCORNet (Fleurence et al. 2014). Some examples of applications to metabolic diseases include the use of PCORNet (McGlynn et al. 2014) to create a common data model for patients affected by specific diagnosis including T2DM, or of eMERGE to secondary data analysis for personalized medicine and phenotypes definitions (Hall et al. 2014; Yazdanpanah et al. 2013). There are also attempts in creating conceptual frameworks for information network specifically in Diabetes (Riazi et al. 2016).

One of the state-of-the-art open source tools available to collect multidimensional data coming from different sources, and aggregate them in a format suitable for temporal analysis is the ***Informatics for Integrating Biology and the Bedside (i2b2)*** Data Warehouse [https://www.i2b2.org/]. I2b2 is one of the seven centers funded by the U.S. National Institute of Health (NIH) Roadmap for Biomedical Computing [http://www.ncbcs.org]. The mission of i2b2 is to provide clinical investigators with a software infrastructure able to integrate clinical records and research data (Murphy et al. 2010). I2b2 Data Warehouse gives the possibility to integrate, visualize and query data in an informative way, considering both their temporal nature and complexity. The i2b2 clinical data analytics platform is currently used in 140 centers and it has been adapted to allow federated querying in multi-institutional networks to get patients clinical data and

use them to support research activities.

Interesting examples on how the i2b2 system has been used to integrate administrative and clinical data in a research framework are presented in (Post, Kurc, Cholleti, et al. 2013), where the authors propose a platform for detecting phenotypes of patients' characteristics to support healthcare data analysis and predictive modeling also embedding temporal data representations, and in (Segagni et al. 2011) where i2b2 is exploited to integrate information from a biobank with clinical data to support translational research in oncology.

Since it was developed, i2b2 framework involved other complementary projects. The interoperability project Substitutable Medical Applications and Reusable Technologies (SMART) was devoted to develop a platform that allows medical applications to be written once and then run across different healthcare IT systems. (Mandl et al. 2012). SMART was lately updated to take advantage of the clinical data models and the application-programming interface described in a new, openly licensed Health Level Seven (HL7) draft standard called Fast Health Interoperability Resources (FHIR). The new platform is called SMART on FHIR (Mandel et al. 2016), and it has been recently exploited to build an interface that serves patient data from i2b2 repositories (Wagholikar et al. 2016). The SMART on FIHIR example is very interesting as it shows how, in addition to collect clinical data and exploit them for research, i2b2 can serve as common data framework to create structured annotation and spread the acquired knowledge back into clinic. I2b2 can support the ***sidecar approach***, which allows to continue using existing clinical system (EHR) as-is while using a secondary database (the i2b2 instance). Typically, data are transferred from clinical database to an application layer, for example a decision support system, where mining algorithm can be applied and calculations performed. In the "Big Data to serve CDSS" context, one of the most interesting features of i2b2 is its capability to support a sidecar approach (Wagholikar et al. 2016). The sidecar approach allows the software to aggregate a copy of the patient data from the EHR and provides query services in parallel to clinical actions for research. This means that, while patients are managed through the EHR during daily clinical practice, the i2b2 sidecar can concurrently host a copy of the data for secondary use, enabling secure data access and interoperability.

2.2 Clinical Decision Support: data integration solutions and Systems

The above described network initiatives, together with advanced ways to gather and merge information from unstructured and heterogeneous sources - mainly Electronic Health Records (Peters & Buntrock 2014; Peters & Khan 2014; Patrick J O'Connor et al. 2011), but also social media or environmental data - have the potential to be the basis for new generation Clinical Decision Support System (CDSS).

CDSSs have been traditionally defined as softwares designed to aid clinical decisions making, which enable the combination of individual patient characteristics with computerized clinical knowledge and specific recommendations, in order to be presented to clinicians for specific decisions (Sim et al. 2001). While it is recognized that developing and deploying CDSSs is essential, especially in

disciplines that require complex decision-making, such as chronic disease care, nowadays CDSSs use in routine clinical practice is still limited (Belard et al. 2016). Possible causes of this limitation might be identified in the technical complexity associated with efficiently integrating large data sets from diverse sources and in the methodological efforts to provide meaningful knowledge in the whole process of care.

To provide successful decisions support, CDSSs should respond to basic requirements including (i) rich contents in terms of knowledge, references and data evidences, (ii) the capability of processing huge amount of data with fast response times and (iii) implementations enough intuitive and appealing to catch users attention and not obstruct clinical actions. These characteristics translate into the fundamental CDSS components: data repositories, rule engines and user interfaces.

As stated before, a CDSS ***data repository*** might be very complex, due to contents being derived from differently structured or unstructured (e.g. free text annotations) and information models that must formalize each event patient undergo and tackle its representation (including provenience of data, type of action and outcome associated with the action) in multiple clinical scenario. Researchers need to access the increasing amount of available information in an integrated way while more outcomes become available to clinicians. The time between knowledge discovery and implementation into clinical practice might decrease thanks to this bidirectional flow of information (Blake et al. 2011; Sarkar et al. 2011). To make this effective, CDSSs must adopt ***common data framework*** to create structured annotation, convert datasets into clinically relevant knowledge and guarantee data security and protection of patient rights to improve diagnostic accuracy and achieving precision medicine (Castaneda et al. 2015; Hovenga & Grain 2013). Translational research offers several examples of platforms providing specific integrated databases upon which researchers can design models, like Hugenet (Canestaro et al. 2014), which aggregates genomic, environmental and public health data. Though one of the most important innovation of the past years is represented by i2b2 (Informatics for Integrating Biology and the Bedside). i2b2 (Kohane et al. 2012) implements a scalable informatics framework that enable researchers to use existing clinical data for discovery research, serving as common data model to create structured annotation and spread the acquired knowledge back into clinic.

The second, and central, component of CDSSs is the ***Engine***, which has the fundamental role of represent and analyze data, and transform them into knowledge.

On the basis of the functional relationships that the delivered support has with data, information or knowledge, CDSSs can be divided into specific categories. They can explicitly represent knowledge and its formalization, for example in the form of guidelines (Bouaud et al. 2013; Ebrahiminia et al. 2006; Shalom et al. 2016; Schoen et al. 2015). In this case the common way to define CDSS is as ***Evidence Based*** or Evidence Adaptive, in which the knowledge that match patients' data are based is derived from research literature evidence and practice-based sources (Sim et al. 2001; Chan et al. 2009). CDSSs can suggest specific actions thanks to ***Data-driven*** models and techniques, like risk prediction models (Huang et al. 2007; Russell & Rosenzweig 2007) or data mining methods (Yu et al. 2008; Kammoun & Ayed 2014).

The third component of any CDSS is the ***user interface***, which displays the knowledge from the data repository to the final user through the rules engine, to the final user, and might include

warnings, possible outcomes or the view of historical events and actions. ***Dashboards*** implement a specific user interface approach and have been defined as CDSSs capable of querying multiple databases to merge information and provide a visual summarized representation of key performance indicators, in a "car's dashboard" format. (Mick 2011; Batley et al. 2011; Sprague et al. 2013; WILBANKS & LANGFORD 2014). In this context it is worth to cite also ***infobuttons***, which can be considered the precursors of modern Dashbords. Infobuttons were introduced (Cimino 1997) as context-specific links between clinical information systems and other source of information, which are intended to anticipate and guide clinician information needs. Since their introduction, they were optimized to improve suboptimal interfaces (Cimino et al. 2007) and, more recently, to link various resources and select those most meaningful for a given context (Cimino et al. 2013).

The visual properties of Dashboards allow providing summaries of a big volume of data through easy to read color coded graphical format (like traffic light) in order to deliver an intuitive assessment of complex clinical conditions or management performances (Simms et al. 2013; Frith et al. 2010). Dashboards are versatile tools that can be deployed both to enhance adherence to evidence based practice guideline or to support data-driven decision making, though it is easy to imagine them as the natural implementation of Visual Analytics based CDSS. Dashboard systems in clinical decision support are relatively new. The performed literature review has shown that one of the fields where Dashboards have been extensively used is ***pharmacotherapy*** and medication surveillance (Mould & Dubinsky 2015; Barrett et al. 2008; Waitman et al. 2011; Linder et al. 2010; Dombrowsky et al. 2011; Dixon et al. 2013) mainly for dose calculation. In the therapeutic drug monitoring field, the most recent Dashboard systems implement ***Bayesian approaches*** to find appropriate dose regimens and adjustment for specific patients representing the relationship between exposure and response to monoclonal antibodies (Mould & Dubinsky 2015), to enable the individualization of drug dosing to achieve predefined clinical targets on the basis of Bayesian stochastic adaptive control (Hope et al. 2013) and to prospectively assess the predictive performance of a dosing method and determine the expected time in a correct therapeutic range (Wright & Duffull 2013).

CDSSs can provide longitudinal data representations through ***Visual Analytics*** that integrate analytical reasoning with interactive visual interfaces. CDSSs based on visual analytics might be the vector to translate into meaningful knowledge representation the findings of Big Data analytics, either data mining or predictive models, and to support cognitive informatics, for example in the perception and recognition of temporal patterns. One of the peculiarities of visual analytics is a great flexibility and the capacity to be adapted to heterogeneous scenarios, which might involve longitudinal data in motion, and implement the solution to visualize the temporal relations among information. When big data are specifically exploited for CDS, visual analytics enables hypothesis generation and facilitates real-time clinical decisions (Vaitsis et al. 2014; Ola & Sedig 2014; Simpao et al. 2015). Visual analytics can be a very powerful tool if used in combination with longitudinal models to analyze long time series (Gálvez et al. 2014; Mane et al. 2012) and to enhance pattern visualization to focus attention in monitoring clinical actions (Simpao et al. 2014; Palacios 2016), or to detect and illustrate patients' behaviors in time and space to identify health-risk scenarios (Juarez et al. 2015).

There are also several examples of CDSSs where visual analytics methods combine evidence-based

and data-driven approaches to improve clinical performances for example retrieving drug interactions (Simpao et al. 2014; Resetar et al. 2005), by using advanced analytics to query clinical records for meaningful information and then integrating these information into knowledge based rules within EHRs (Slonim et al. 2012) or by gathering EHR data and entering them into models able to perform risk stratification (Gotz et al. 2012). There are attempts to use visual analytics into field of epidemiology to understand the interaction among time dependent variables (Chui et al. 2011).

Within ***T2DM management***, computer based models that allow estimating long-term outcomes and identifying the most efficient management strategies, comparing different populations in various clinical settings, are not new (Palmer et al. 2004). Nevertheless, chronic outpatient CDSSs have often had an inconsistent effect on key aspects of diabetes care, due to low use rates in several settings. (Patrick J. O'Connor et al. 2011). A peculiar aspect of the existing CDSSs designed for T2DM is that they are strictly focused on very specific issues and are not able to follow the evolution of the disease allowing a holistic vision of disease management. There are CDS systems developed to enhance personalized treatment and medication recommendation (Donsa et al. 2016; Sáenz et al. 2012; Tan et al. 2010; Toussi et al. 2008; Liu et al. 2013; Ampudia-Blasco et al. 2015). If designed to improve glycemic control (Lim et al. 2011; Lipton et al. 2010; Neubauer et al. 2015; Rodbard & Vigersky 2011; Augstein et al. 2010), they don't take into consideration disease complications, otherwise they are focused on specific complications like diabetic foot (Peleg et al. 2008; Israel et al. 2014), retinopathy (Reza & Eswaran 2011; Kumar & Madheswaran 2012; Mitsch et al. 2016), nephropathy (Cho et al. 2008) or cardiovascular risk (Cleveringa et al. 2008). Some CDSSs are designed for specific settings, like primary care (Häussler et al. 2007; Barlow & Krassas 2013; Ziemer et al. 2006; Heselmans et al. 2013), but there are also serval examples of CDSSs that promote shared care and collaborative decision making (Bødker & Granlien 2008; den Ouden et al. 2015; Holbrook et al. 2009; Liu et al. 2012; Welch et al. 2015; O'Reilly et al. 2012; Parker et al. 2014).

The last set of CDSSs, developed to enhance T2DM management, responds to several of the previous discussed challenges and have the potential to deliver clinical support that is concurrently standardized through evidence-based (Liu et al. 2012) but also highly personalized, as recommendations are tailored not only to a given patient's clinical state or behavior (den Ouden et al. 2015; Holbrook et al. 2009) but also to clinical settings (Welch et al. 2015). Nevertheless, none of these systems takes into account the longitudinal nature of clinical data neither implements any temporal data driven method, which could represent events in time as evidence of the patient's clinical state evolving and finally provide easy to access, homogenous, yet tailored recommendations, to different health care providers in any moment of patients careflows.

To promote a more accessible and clinically informative way to depict the increasingly complex information available for chronic patients, one of the essential success keys for effective monitoring is the detection of changes in activities, patients' status and performances over time, by means of ***longitudinal data*** (Chaudhry & Feest 2011). A powerful example of data representation in time is given by the Gapminder Foundation [http://www.gapminder.org], which is devoted to develop motion charts to support complex data interplays in order to be intuitively investigated.

2.3 Temporal data analysis and careflow mining: data driven methods to infer new evidences from longitudinal data

Information Retrieval and ***Data Mining*** methods, aimed at automatically extracting information from data, were successfully applied in a wide range of fields such as marketing, surveillance, finance, meteorology, fraud detection, web usage monitoring, healthcare and scientific discovery. Data Mining represents a single step in the process of knowledge discovery in databases and can be described by the following definition, originally given by (Fayyad et al. 1996): "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". This definition highlights that the target of the discovery is focused on valid information, not already known and plausible of a practical exploitation to support the knowledge discovery process.

Data collected in current health care systems (EHRs, Hospital Information Systems, Electronic Case Report Forms) allow incorporating the temporal dimensions to which data are related, and according to suitable procedures of historical management, data can be linked to a proper time of validity. The importance of dealing with longitudinal data embraces several settings: (i) data gathered over long periods of time at a slow pace, like data stored over years in chronic-care settings, (ii) data collected in a continuous way through monitoring and wearable devices, (iii) data rapidly accumulated in domains that require fast actions, like intensive care units. Literature increasingly reports examples of methods implemented to tackle the challenges related to analyzing temporal clinical data that include a large numbers of variables, different sampling frequencies, and heterogeneous types of events, which can be either instant or interval based (Bellazzi et al. 2009). This necessity to fully understand temporal data implies the need to explicitly take into account in the data mining process also the temporal dimension. For this reason, traditional data mining methods, capable of performing only "static" analyses, have been adapted and extended to explicitly handle temporal reasoning and to incorporate the recognition of temporal features, giving rise to a research field called temporal data mining (Brown 2008).

Clinical time series data have some peculiar features that distinguish them from more traditional information recorded in time (Benin et al. 2011). First of all, excluding regularly monitored physiological signals (e.g. ECG, respiratory rate, etc.), clinical data are often measured without a specific sampling scheme. For example, laboratory tests are executed when needed, and temperature and blood pressure can be taken at different times during the day (as they are manually collected by a health care operator). This results in the collection of time series that are most often characterized by an uneven sampling grid. In addition, the number of recorded samples is usually low. For these reasons, traditional signal processing techniques are often not applicable to medical time series data. To cope with these peculiarities, temporal data mining (Mitsa 2010; Post & Harrison 2008) has started to be extensively used for analyzing clinical time series. Temporal data mining techniques give the opportunity to perform deep analysis about the temporal behavior of complex processes, and may help to forecast the future evolution of a variable or to extract causal relationships between the variables involved in a multi-dimensional scenario. The capability of analyzing the temporal behavior of patients' conditions may allow a more effective adaptation of clinical interventions. The understanding of complex ***multivariate clinical histories*** produces multiple advantages, it can offer new insights at clinical and organizational level on clinical

processes, treatments or about associations between patient's clinical status and care delivery processes (Parsons et al. 2012; Mouttham et al. 2011; Kahn & Ranade 2010).

Temporal data mining techniques and tools enrich and integrate traditional data mining methodologies by explicitly taking into account temporal aspects (Post & Harrison 2008) to extract new and useful clinical knowledge (Brown 2008). A comprehensive introduction to this class of methods is given by the following definition (Benin et al. 2011): "temporal data mining is a single step in the process of Knowledge Discovery in Temporal Databases that enumerates structures (temporal patterns or models) over the temporal data".

The biomedical domain naturally has to deal with temporal issues due to the intrinsic longitudinal nature of clinical data. Within ***medical informatics***, time had been recognized as a fundamental aspect of clinical information systems (Combi & Shahar 1997) to represent (Adlassnig et al. 2006), maintain, query (Das et al. 1992), and reasoning (Shahar 1999; Augusto 2005) about clinical data. Further and innovative contributions are still being suggested to mining of electronic health records for revealing novel patient-stratification principles or unknown disease correlations (Jensen et al. 2012), to find functional dependencies in clinical database (Combi et al. 2014), to retrieve suitable time parameterization for association and clustering experiments (Hripcsak et al. 2015) and to classify multivariate time series (Moskovitch & Shahar 2015). Other recent temporal data analytics examples include applications in biomedical informatics (Singh et al. 2015; Gálvez et al. 2014; Hauskrecht et al. 2013; Warner et al. 2016), the development of novel data mining algorithms (Fei Wang 2013; Last Mark, Tosas Olga, Cassarino Tiziano Gallo, Kozlakidis Zisis 2016), time series analysis (Liu & Hauskrecht 2015; Mitsa 2010), pattern discovery (Wang et al. 2012), text mining application (Savova et al. 2009; Sun et al. 2013; Nikfarjam et al. 2013; Goldstein & Shahar 2016) and the effective implementation of frameworks able to represent the longitudinal nature of data through temporal abstraction (Batal et al. 2009a; Lucia Sacchi et al. 2015a).

Temporal abstraction**s** (TAs) are able to give common ***representation*** of heterogeneous data transforming raw clinical time series into meaningful representation of clinical events. Knowledge-based Temporal Abstraction were theorized in the late nineties (Shahar 1997), and have become more and more popular in the clinical data mining community through the years (Stacey & McGregor 2007; Post & Harrison 2007a; Verduijn et al. 2007; Combi & Chittaro 1999; Lucia Sacchi et al. 2015b), especially when the need for data integration has become a primary issue in the research. In general, TAs are a way to perform a shift from a quantitative time-stamped representation of raw data to a qualitative interval-based description of time series, with the main goal of abstracting higher-level concepts from time-stamped data. According to the models specified in (Post, Kurc, Cholleti, et al. 2013; Batal et al. 2009b), it is in general possible to identify two types of abstraction tasks, which lead to the definition of two types of TAs: a basic (or low level) and a complex abstraction task. ***Basic TAs*** take as input raw time series of time-stamped clinical data and output a series of intervals where specific behaviors hold. Basic TAs identify simple behaviors, such as states and trends, in the data and offer a way to represent quantitative variables through qualitative labels holding on time intervals. ***State TAs*** extract the intervals in which a variable hold within a defined set of meaningful values and correspond to expressions like "high blood glucose values between March 11th and September 22nd 2015" or "normal cholesterol in the last six months". ***Trend TAs*** represent increase, decrease, and stationary patterns in quantitative time series.

Clinical time series processed using basic TAs are thus already represented by sequences of (interval-based) events. ***Complex TAs***, also referred to as temporal patterns (Post & Harrison 2007b; Batal et al. 2009b) works on sets of intervals rather than on raw time series. They require two intervals sets as inputs, associated to two TAs, and provide an interval set as output. Such interval set is associated to a new TA, which is evaluated according to a pattern specified between the intervals of the two composing sets. The definition of a complex TA pattern is based on temporal relationships detected among the TAs on which it is built up and are usually exploited to detect temporal patterns are the temporal operators defined in Allen's algebra (Allen 1984). TAs representation offers several advantages, for example they allow medical experts to have a qualitative idea of the abstract patterns they would be interested to see in the data, as they are able to automatically translate such pattern into a formal representation. Moreover, TAs are particularly well suited for dealing with unevenly spaced time series, in fact, once the time series data are abstracted using TAs, their representation becomes uniform to the one characterizing data represented as sequences of events.

Among TDM techniques, those that resulted more suitable to analyze complex patients' temporal histories are the ones related to the efficient mining of ***frequent temporal patterns*** from data. Such patterns may identify the most common sequences of events in the data (e.g. sequences of hospitalizations, drug prescriptions, etc.). Methods related to ***sequential pattern mining*** are aimed at discovering frequent patterns that occur in series of time- stamped data (Agrawal & Srikant 1995). They can be used to predict future events and effectively exploited in decision support applications to extract meaningful information from the daily activity of health care organizations or patients monitoring. An initial group of algorithms was developed to deal with time stamped events without a duration (Zaki 2001; Ayres et al. 2002), whereas further implemented methods were able to deal with interval based data to mine sequential patterns (Kam & Fu 2000; Höppner & Klawonn 2001; Zhang et al. 2008; Patel et al. 2008) and to learn temporal rules, in which a set of heterogeneous events is temporally associated - it precedes - to another event of interest, deriving association rules from sequential data (Concaro et al. 2011; Sacchi, Larizza, Combi, et al. 2007; Moskovitch & Shahar 2009; Winarko & Roddick 2007).

Temporal Association Rules (TARs) are association rules of the kind A → C, where the antecedent (A) is related to the consequent (C) by some kind of temporal operator. TARs mining algorithms are aimed at extracting frequent associations, where frequency is evaluated on the basis of suitable indicators, the most utilized being support and confidence. The support gives an indication of the proportion of cases verifying a specific rule in the population; confidence instead represents the probability that a subject verifies the rule given that he verifies its antecedent. In the case of TARs such indicators are extended to take into account the temporal nature of the data.

Within the University of Pavia working group, where the presented research took place, several efforts have been done in the past to develop a framework for ***mining TARs based on TAs*** (Bellazzi et al. 2000; Sacchi, Larizza, Magni, et al. 2007; Concaro et al. 2011). In (Bellazzi et al. 2000), a method to derive temporal association rules using Basic TAs in the antecedent and in the consequent is presented. Basic TAs are though not enough to represent the complex qualitative behaviors that clinical experts want to retrieve when reasoning about temporal data representation. The methodology has then been extended to take into account complex patterns both in the antecedent and in the consequent of the rule (Sacchi, Larizza, Magni, et al. 2007). Patterns of

interest can be specified on the basis of domain knowledge into a set called Abstractions of Interest, and rules containing such pattern in the antecedent and in the consequent are extracted. After the development of a TARs mining framework mainly oriented to the analysis of clinical data, the framework had been extended to incorporate also administrative healthcare information into the data set. This required the methodological effort focused on several aspects, like the necessity to integrate clinical and administrative information (Bellazzi et al. 2009). This task was performed using the TAs framework set up in (Sacchi, Larizza, Combi, et al. 2007). In (Concaro et al. 2011) a methodology to mine disease-specific TARs is presented: rules are mined in a population of diabetic patients and in a group of controls, and only significant disease-specific patterns are extracted. In addition, the mining process includes additional filtering strategies that are coupled to the more traditional cut-offs based on confidence and support. More recently, a framework that includes a library of algorithms for time series preprocessing, abstraction and execution of temporal data processing workflow has been developed (Lucia Sacchi et al. 2015b).

Sequential pattern mining and TARs methods are are powerful tools and allow detecting interesting temporal relationships between diagnostic or therapeutic patterns from clinical histories. Although the mined patterns and rules can contain arbitrarily complex patterns in the antecedent and in the consequent, they are usually able to capture histories with limited length. As a matter of fact, ***frequent clinical histories*** made up of longer chains of events are not straightforward to be mined through these techniques, which do not have the capability to represents a whole care process. To overcome this limitation, the reconstruction of the so-called clinical pathways is becoming nowadays one of the most challenging fields in data mining in healthcare. The opportunity of developing these novel algorithms is primarily offered by the availability of hospital information systems, which allow collecting large amounts of data related to complete clinical histories.

Careflow mining techniques allow extracting frequent histories from collection of events (event logs) and have been used to describe process-related information in chronic diseases (Panzarasa et al. 2004; Chesani et al. 2008), to analyze set of mixture clinical and administrative data (Concaro et al. 2011), and to depict interactions of patients with healthcare providers (B. Huang et al. 2012). Careflow mining methods are intended to automatically learn the most typical patient careflows from routinely collected administrative and/or clinical data. Typically, careflow mining is performed using algorithms derived from the process mining domain (W. M. P. van der Aalst 2011; Jagadeesh Chandra Bose & Van Der Aalst 2012; Bouarfa & Dankelman 2012; Rebuge & Ferreira 2012; Caron et al. 2014; West et al. 2014; Z. Huang et al. 2012; Tsumoto et al. 2014). Some recent approaches tackle the problem of clinical pathway discovery through probabilistic topic models (Huang et al. 2014) and by using a dynamic-programming algorithm (Huang et al. 2013). Other works show how the adaptation of frequent sequence mining algorithm can successfully be applied to electronic medical records to derive care pathways (H. Liu et al. 2015; Perer et al. 2015; Li et al. 2015).

Once mined, the careflows can be evaluated through a comparison with clinical guidelines and hospital-specific protocols to check their adherence to best practices and hospital management directions.

In a time-oriented setting, the investigation of temporal analysis and careflow mining techniques not only increases the computational efficiency of analytics, but it has also the potential to improve

the quality of patient care through the detection of meaningful clinical knowledge, which can be delivered through ad hoc CDSS.

Since their first introduction (Cook & Wolf 1998; R.a et al. 1998), approaches for ***clinical pathways mining*** have been often borrowed from business process analysis. In analogy to this discipline, in clinical pathways mining the sequences of events occurring to each patient during his clinical history are commonly referred to as event logs. Differently from clinical process modelling, where workflows are manually constructed on the basis on medical evidence (e.g. clinical guidelines), and clinical pathways mining has the advantage of exploiting the huge amount of data collected through the hospitals information systems to reconstruct the most frequent histories that took place in a particular medical center. This gives the possibility of detecting anomalous pathways or site-specific behaviors that can be operated in a hospital for specific reasons possibly not stated in the current clinical guidelines. On the other end, though, given the high heterogeneity and variability of the processes of care, the interpretation of the results is not straightforward.

The techniques that have been most often exploited to analyze clinical histories (Rebuge & Ferreira 2012; Z. Huang et al. 2012; Yang & Hwang 2006; Lin et al. 2013) are the ones coming from Process Mining, a general method used in business process analysis (Van Der Aalst et al. 2004; R.a et al. 1998; Cook & Wolf 1998). To cope with the variability of healthcare processes, techniques to help interpreting and synthesizing PM results have been developed (Huang et al. 2013). Some recent approaches tackle the problem of clinical pathway discovery through probabilistic topic models (Huang et al. 2014) and by using a dynamic-programming algorithm (Huang et al. 2013). Other works show how the adaptation of frequent sequence mining algorithm can successfully be applied to electronic medical records to derive care pathways (H. Liu et al. 2015; Perer et al. 2015; Li et al. 2015).

Clinical pathways mining usually works on event logs made up of data coming from administrative data streams. Interestingly, only a few works in the literature deal with the exploitation of clinical data into the mining process (Fernández-Llatas et al. 2010).

Process mining has been defined as the method for retrieving a structured process description from a set of real executions (van der Aalst & Weijters 2004), the event log. An event log contains information about events, which refer to activities or tasks executed in a particular process and for a specific case. Typically, event logs also record the time when these tasks were executed (the timestamp of an event). ***Event logs*** also typically store information about the originator of a task, i.e. who performed which task or initiated an event. Based on these event logs, the goal of process mining is to extract process knowledge (e.g. process models) in order to discover, monitor, and improve real processes (W. van der Aalst 2011). It's necessary to characterize event logs with specific attributes. In healthcare, it is possible to consider patients' ids as cases, which are the process instances, and procedures as activities, defined as the steps of the process. As other data mining frameworks, also process mining can be thought as a three-phase process: pre-processing, processing and post- processing.

- In the pre-processing phase, the event log is cleaned and prepared to be read from process mining tools;
- In the processing phase, a mining algorithm is applied to the event log and the ordering relationships between tasks serve as the input;

- For the post-processing phase, both the event log and the generated process model serves as input. They can be used to find additional information about the process or to adjust the process model as well as represent it graphically.

In the process mining area, the ***ProM framework*** (www.processmining.org) has become the de facto standard for process reconstruction and analysis. ProM is an open source platform developed by the Process Mining Group of the Eindhoven University of Technology [http://www.processmining.org/]. It is an extensible framework that supports a wide variety of process mining techniques and algorithms (Verbeek et al. 2006). The ProM architecture offers several functionalities (Verbeek et al. 2006) like mining devoted plug-ins that implement a variety of algorithm (e.g. Alpha algorithm, Heuristics Miner, Genetic algorithm) and export/import functionalities to save mining results in form of different objects.

Among the process mining algorithms mentioned above, one of the most frequently exploited for clinical applications is the ***Heuristics Miner***. This algorithm, described in detail in (Weijters & Ribeiro 2011), is a practical applicable mining algorithm that can deal with noise and low frequency behaviors. The algorithm focuses on the control flow perspective and generates a process model in form of a Heuristics Net for the underlying event log. A recent process mining review paper (Rojas et al. 2016) shows that in the healthcare context the Heuristic Miner is the most popular technique for mining event log derived from Hospital Information System (HIS).

The Heuristic Miner algorithm, together with other process mining methods, like alpha-algorithm and fuzzy miner, have been explored in this research program and applied to studied population. However, they showed some intrinsic limits for their use in the context of CDS, mainly related to unclear and poorly readable results. Other drawbacks in their use to reach this research objectives were related to their inability to identify sub-cohorts in the studied population, namely performing electronic phenotyping. Limits of these approaches are described in detail in section 5.1.1.

2.4 Electronic Phenotyping, Use of Temporal Analytics on Big Data to Deliver Decision Support

One of the main scientific goals of this research is to show how careflow mining (Quaglini et al. 2001; Caron et al. 2012) can be a useful instrument to perform electronic temporal phenotyping (Pathak et al. 2013; Hripcsak & Albers 2012; Rasmussen et al. 2015; Yahi & Tatonetti 2015) from EHR, by identifying the evolution of clinical states over time across a specific patient population.

Electronic Phenotyping has been defined as "the collection of methods able to exploit the data that are captured in routine health care delivery to identify groups of patients who meet specific conditions that are relevant to clinical studies of interest" (Richesson, Hammond, et al. 2013; Tu et al. 2011; Solti et al. 2008). Careflows allow identifying different sub-groups (sub-cohorts) of individuals in a larger cohort of patients (i.e. all the patients who undergo to the same - or similar - temporal history). Such groups may show differences in terms of patient complexity, disease stage, or treatments. As a consequence, careflow mining can be easily seen as a type of electronic phenotyping.

A recent review (Xu et al. 2015) of available electronic phenotyping tools, state current challenges in the field and, through a review based on tools capabilities, identify twenty-four state of the art tools for electronic health record–driven phenotype. Even though the authors consider as one of the evaluated capability the possibility to deal with temporal operators, none of the presented tools is designed to define computable phenotypes through careflow mining. A less recent review (Shivade et al. 2014) better contextualizes this research program efforts: authors identify ninety-seven articles and classify them through the exploited approaches (natural language processing techniques, rule-based systems, statistical analyses, data mining, and machine learning techniques) recognizing the increased popularity of systems based on statistical analyses, machine learning and data mining. Although this growing trend, the reported methods are mainly univariate and cross-sectional. Authors also report the top ten phenotypes of interest, among which Diabetes is listed second. As conclusion, authors underline the necessity of a broader use of statistical and machine learning methods to develop more generalizable solutions. A very interesting approach to perform electronic phenotyping with machine learning is presented in (Peissig et al. 2014). Authors identify common challenges derived by the nature of EHRs, derived by interdependent heterogeneous observations (biological, anatomical physiological, and behavioral) and the presence of events that represent a patients' medical history. The tackled problem is the same of this work: to discover computable, relevant and reliable phenotypes while discovering hidden relationship among noisy information. In this case authors apply inductive logic programming, switching from a prepositional to an inductive approach, which identified phenotypic models of T2DM with an accuracy of 0.945 and Diabetic Retinopathy with accuracy of 0.988. Although the notable results, this method is not able to discover longitudinal phenotypes.

In (Pathak et al. 2013) and (Hripcsak & Albers 2013) authors discuss the importance of phenotyping through data mining methods and highlight the importance of taking into account the temporal dimension through proper temporal modeling and temporal abstraction to identify specific sub-cohorts of patients. Examples of such an approach are reported in (Albers et al. 2014; Hripcsak et al. 2015; Pivovarov et al. 2014) where temporal modeling is used to analyze the dynamics of pathophysiological variables in clinical records. Another important function of temporal phenotyping is to provide proper representations of the retrieved patterns. Some interesting examples are provided by the exploitation of temporal graph to depict disease evolving patterns and complication arising (C. Liu et al. 2015; Ng et al. 2014), of visual analytics methods for interactive pattern mining (Gotz et al. 2014) and of web-based platforms for information combination (Yu et al. 2014). However, this body of works does not address clinical situations like heterogeneous data collected both from electronic health records (EHR) and administrative sources data, explicit temporal information and the employment of clinical data to characterize the transitions between states of care. In order to address these shortcomings, careflow mining was applied in the context of electronic phenotyping, and a new algorithm was developed.

The importance of enhancing phenotype description through temporal information is highlighted also in (Frey et al. 2014), where authors discuss the current scientific opportunities that the use of Big Data technologies could convey in the definition of phenotypes in medical records. Beyond the first step of creating a common representation of data to harmonize the phenotype research, electronic phenotyping activities should be more largely supported using so called Big Data approaches, which enable scalable classification of EHR events and semantic and temporal similarity analysis.

Examples of frameworks that automatically extract phenotypes from EHR data are still limited and, to the best of our knowledge, none of them is able to analyze, or most importantly represent, longitudinal data. For example, while they combining different sources of EHR data, they tackle the complexity through rule-based approaches (Nadkarni et al. 2014), or they rely on semi-supervised methods to extract patient narratives, but without explicitly representing temporal correlations among events (Halpern et al. 2016). Recent applications of machine-learning model to large sets of EHR data (Warner et al. 2016) have shown the possibility of performing accurate predictions of hospital acquired complications using temporal clinical data as a function of healthcare system exposure.

Yet few works are applied to characterize T2DM. Some approaches are explicitly cross-sectional (Anderson et al. 2015), as they use logistic regression and random-forests model. However, they demonstrate improved performance in using EHR data, when compared to basic covariates, for T2DM cases. In (Wei et al. 2013) authors apply the eMERGE high-throughput clinical phenotyping algorithm to recognize T2DM case and controls, and demonstrate the impact of insufficient longitudinal data on the accuracy of an algorithm, thus suggesting to carefully consider temporal aspects in the design and execution of T2DM phenotyping algorithm. (Yahi & Tatonetti 2015) exploit a knowledge-based method to generate pathology signature for the terms of a given ontology, and apply the algorithm to the T2DM case. They define and validate these signatures on longitudinal medical record, but do not consider their co-occurrences across time.

These literature findings and limitations suggested that electronic phenotyping has the potential to leverage on temporal methods to integrate into CDSS an innovative and holistic approach for chronic disease management. The cited body of works does not address clinical realities like heterogeneous data collected both from electronic health records (EHR) and administrative data sources, explicit temporal information and the employment of clinical data to characterize the transitions between states of care. In order to address these shortcomings, careflow mining had been applied in the context of electronic phenotyping, and a new algorithm was developed that can process data related to health care events and enrich the mined patterns with clinical data.

Type 2 Diabetes is a multifaceted disease and the differences across its phenotype definitions might affect their use in healthcare setting and the consequent interpretation of data. To find appropriate definitions of T2DM phenotype for specific sectors, like healthcare, research or policy making, is very important. Several efforts have been done to defining the clinical characteristics of diabetes cohorts (Richesson, Rusincovitch, et al. 2013). The NIH Health Care Systems Research Collaboration released the support material to guide researchers in identifying the more suitable T2DM phenotype definitions by the purposes of a specific study [https://www.nihcollaboratory.org/], considering the data gathered from three main domains (ICD-9- CM–coded diagnoses, laboratory test results, and medication data) in different combinations and thresholds. They propose to use:

- the eMERGE phenotype definition, which is based on the diagnosis codes supplemented by relevant laboratory results and medication prescriptions, for researches that require high sensitivity for cases identification (Wei et al. 2012; Newton et al. 2013; Kho et al. 2011; Pacheco et al. 2011);
- the Center for Medicare and Medicaid Services, Chronic Condition Warehouse definition [http://www.ccwdata.org/index.htm], which is based exclusively on ICD-9-CM codes, for

studies gathering information as a baseline characteristic, risk factors and comorbidities, in particular to support health services monitoring and quality reporting needs (Gorina & Kramarow 2011);

- the Durham Diabetes Coalition definition, which includes ICD-9-CM diagnosis codes, laboratory values and medications, with the intent of providing a broad EHR-based definition of diabetes, for prevention or behavioral studies aimed at supporting regional public health intervention (Spratt et al. 2015; Rusincovitch et al. 2013).

The above phenotype definitions serve to identify T2DM cases, while they are useful to understand what characterize the diabetic disease they have to be modified in order to apply to this research requirement, which is focused on recognizing sub phenotypes in a cohort of diagnosed patients and stratify them on the basis of risk profiles. Another, more general, scientific issue arises from the need of customize phenotype definitions due to the heterogeneity of EHR systems, and the variety of practices in different clinical settings. The concept components of phenotypes definitions have to be translated into specific operational entities (Richesson, Rusincovitch, et al. 2013).

CHAPTER 3

3 Data Gathering and Integration (Aim 1)

The first phase of the research was dedicated to the data gathering activities necessary to collect and make available a large temporal multivariate data set derived from heterogeneous sources, including different hospitals EHRs and data collected by healthcare agencies for reimbursement proposes. The data stored in hospitals EHR or in administrative databases do not have the same consistency and accuracy of data collected for experiments (Weiskopf & Weng 2013; Bowman 2013). As consequences, while this database offers some interesting chance to get new insights into patients' disease evolution, it also sets several issues, largely related to data variability and heterogeneity.

These data originate by the careflows followed by patients, who are managed by different professionals during the evolution of the disease, including GPs, doctors in specialist centers, doctors in hospitals for acute events or comorbidities, pharmacists. Data are collected over long time spans by different health care giver, in different format and for different scopes. To get a complete view of patients' histories requires an accurate preliminary analysis to fully understand the relevant information to consider from each data stream, or when this information is overlapping. Another related problem concerns encounters frequency, related to the possibility to work with data collected continuously over time and missing data in time series.

The need for creating a common and sharable data model to best exploit the repository is the first essential step to the following research activities. As a matter of fact, each hospital database has a different structure and different variables might potentially be collected by different centers. On the other hand, though, in order to query data consistently, it is important to share a common data model with a homogeneous representation of the collected parameters.

The ***objective*** of this step of the research is to finalize the design of a common data model and the implementation of a data warehouse that efficiently integrate and handle temporal multivariate data from heterogeneous sources. Medical informatics state of the art tools were exploited.

To reach this objective, the ***activities of this research program*** were focused on the definition and implementation of a data model able to represents heterogeneous, and meaningful, information for the T2DM management. Risk scores, and a complexity index, were defined and computed. Extraction, Transformations and Loading procedure were executed to fill the Data Warehouse. In this phase, several efforts were dedicated to check data quality. Relevance of each information, and possible issues for their visual representation in the CDSS, were discussed with clinicians.

All the activities described in this chapter extend the concepts included in the ***Key Area 1*** and, in particular, the segment of the Learning Healthcare System cycle related to informatics solutions to facilitate data integration and to turn Clinical data (in this case enhanced with administrative and

environmental data) into information which are the base for knowledge discovery Research activities.

The content of the following sections has been published and disseminated in several works (Dagliati, Sacchi, Bucalo, et al. 2014; Segagni et al. 2015; Bellazzi et al. 2015; L Sacchi et al. 2015)

3.1 Collecting Data in a Common Framework

One of the main challenges in collecting a data set that aims at representing the whole medical history of a chronic patient is related to data sources' diversity in terms of data structures and acquisition purposes. Complex multivariate temporal data sets have been defined as data sets where data instances are traces of complex behaviors characterized by multiple time series (Batal et al. 2009a). Ideally, for each patient there are at least two complex time series, one linked to his clinical history, and another collecting the succession of his contacts with the healthcare services. As a matter of fact, healthcare organizations have become aware that a new level of data aggregation and reporting is needed to fulfill existing and future requirements.

As already illustrated in the literature review of networks to support big data research and data models frameworks, one of the state-of-the-art open source tool devoted to collect longitudinal heterogeneous data is the i2b2 framework, which gives the possibility to integrate and query multidimensional data, while preserving their temporal nature and complexity (Murphy et al. 2010). Given its open source nature, i2b2 can be highly customized for specific medical applications and the development of ad-hoc plugins for data extraction and visualization. In addition, i2b2 is characterized for being a tool developed specifically to deal with clinical data management and translational research applications. For these reasons it has become widely popular among health care organizations and it is usually preferred to other commercial tools, such as Business Objects (SAP) [http://www54.sap.com] or Cognos (IBM) [http://www-01.ibm.com/software/analytics/cognos]. One of i2b2 key features, which makes it highly suitable to handle complex multivariate temporal data, is that it interlocks medical record data and clinical data at a person-level so that diseases, medical events, and outcomes can be related to each other.

Within the i2b2 model, data are stored in a star-schema relational database. The star-schema architecture is based on a central table where each row represents a single fact. Since in i2b2 a fact is an observation about a patient, this table is referred to as the "fact table". The fact table contains all the quantitative or factual data coming from observations about each follow up (or contact with the other services, like drug purchases, in the case of administrative data) related to each patient, and it is the table where all the values of each observation are stored. Each row of the fact table identifies one observation about a patient (described in the Patient Dimension table) made during a visit (stored in the Visit Dimension table). All the observations derived from different events about a patient are recorded in a specific time range, defined by start and end dates, and are related to a specific concept. The concept can be any coded attribute, such as an ICD9 code for a certain disease or a medication or a specific test result. Figure 8 shows how the "Observation Fact Table" collects for each patient categorized data on a single timeline, divided into consecutive visits.

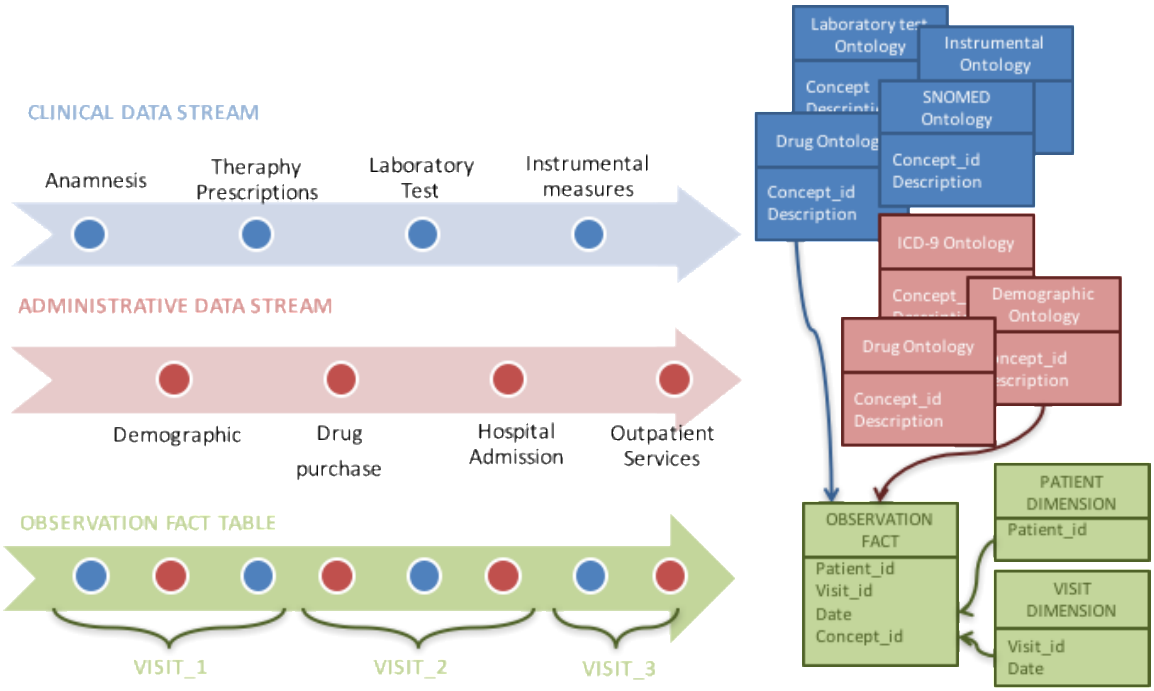


Figure 8 The i2b2 data warehouse to integrate clinical and administrative data streams

To facilitate the query process for the user, data are mapped to concepts organized in an ontology-like structure. I2b2 ontologies aim at organizing concepts related to each data stream in a hierarchical structure. This solution allows the separate management of different data sets and the informative formalization of their content, since each ontology contains the necessary fields to associate each patient's observation with a specific concept. Figure 8 shows the clinical and administrative data streams for an example patient, how they are merged together in the fact table and how they can be described through a set of ontology tables. In this example, drug prescriptions are represented through their ATC drug codes in the Drug Ontology and the subset of Laboratory test from Anatomy Pathology are linked to the SNOMED (Systematized Nomenclature of Medicine) Ontology. Thanks to this approach, data entered in the fact table can be saved with a shared structure that allows to store complex observations with different granularity and extract them in a common format in which each event (identified as an observation on the patient) is associated with specific visits, prescription or hospital services identified by a precise time frame and a coded description. I2b2 data layer is described in details in the following paragraph dedicated to the data modeling activities.

There are several reasons why i2b2 has been selected among the variety of available data warehouses development tools. (i) i2b2 is the only available software that is patient-specific and supports the use of ontologies for querying data. For this reason, there's no need to use dedicated languages to perform a query. (ii) The i2b2 model meets the requirements to reach the goals of an efficient data gathering strategy as it will be possible to: collect multidimensional data from heterogeneous sources, aggregate and export them in a format suitable for temporal dimension analysis. (iii) Another key feature of i2b2 is that gives the opportunity to represent together medical record data and clinical trial data at a person-level so that multiple diseases, exposure and outcomes can be related to each other. For this reason, it is perfectly suitable to the needs of a project such as MOSAIC, especially considering the variability of diabetic patients' events and the disease related

comorbidities. (iv) From a technological perspective, the i2b2 solution provides all the medical centers with a common substrate to store their data. I2b2 aggregate the data repositories under a common framework, in this way it is possible to perform integrated queries while maintaining the data inside each hospital facilities.

For all these reasons, i2b2 was suitable to the needs of this project, where the platform might be used on top of different models for fast data exploration and interrogation, as well as for retrieving data.

3.2 The i2b2 Data Layer

As anticipated in the previous paragraph about common data framework, there are several reasons why i2b2 has been selected among the variety of available DW development tools. The i2b2 DW meets the requirements to reach the goals of an efficient data gathering strategy. The i2b2 data model is based on a star-schema. The star schema has a central "fact" table where each row represents a single fact. A fact is an observation about a patient.

Observations about a patient are recorded by a specific observer in a specific time range (defined by start and end dates) and are related to a specific concept, such as a lab test or diagnosis, in the context of an encounter or visit. The concept can be any coded attribute about the patient, such as a code for a disease, a medication or a specific test result. This way of representing concepts is based on prior work known as the entity-attribute-value (EAV) model (Murphy et al. 2010). The central table of the i2b2 Star Schema is the observation fact table. It contains all the quantitative or factual data coming from observations about each visit related to each patient, and it is the table where all the values of each observation are stored. The observation fact table, as the central fact table of the schema, is the intersection of the dimension tables (visits, patients, concepts and providers). In the observation fact table, facts are defined using concept codes.

Concepts are organized in a hierarchical structure: the i2b2 ontology (also called metadata). Each concept in the ontology is represented by a metadata table [https://community.i2b2.org/wiki/display/releases/1.6.00+Release+Notes]. All metadata tables have the same basic structure, which is the underlying structure for querying the data. The hierarchical path that leads to the term represented in each table is stored in a field called fullname, formed by the names that identify the complete path of the terms: from the root to the leaf value. Below there is an example of fullname for the term "Diabetes with ophthalmic manifestations" contained in the ICD9-CM ontology. It is shown on several lines but is actually one concatenated line in the fullname field and each '\' represents a hierarchical level.

\i2b2\Diagnoses

\ICD9

\(001-999.99) Diseases and injuries

\(240-279.99) Endocrine, nutritional and metabolic diseases, and immunity disorders

\(249-259.99) Diseases of other endocrine glands

\(250) Diabetes mellitus

\(250.5) Diabetes with ophthalmic manifestations

The i2b2 database design allows new blocks of data to be added without disturbing the integrity of the old data. In addition, it allows blocks of data to be copied out of a larger original database to a smaller one. The data about a set of patients can be copied from the i2b2 enterprise database and placed into an i2b2 project database with the same data format and with the same data descriptors while preserving powerful methods for querying the data.

3.3 The Collected Datasets

As stated in the background section, the MOSAIC project involved several research and medical centers and hospitals. Detailed descriptions of hospitals clinical databases and the necessary medical and scientific knowledge to set up the processes of data mapping and parameters selection were retrieved. The analysis of the data available in each of the involved hospitals was aimed at defining the most efficient data gathering strategy to apply.

The data gathering and parameters mapping strategy is detailed for the Pavia data set, as is the most complete data available at the time and the main part of the modeling activities results are based on these data.

The datasets collected in the Pavia area come from two different sources: one is the Fondazione Salvatore Maugeri (FSM) hospital EHR, which collects the clinical data related to its routine medical activities, while the other is an administrative entity (the local healthcare agency - ASL), which mainly collects process data for organizational purposes. An integrated version of these two data sources provides a complete view on the clinical histories of diabetes patients, ranging from their clinical data to the different accesses they performed to national healthcare services.

While the data collected at the hospital level provide detailed information about clinical parameters, they have the intrinsic limit of not being able to supply general and complete information about the history of the patient in terms of time and space. The reason is that the patient, after the diagnosis, is usually treated for a certain period of time by his/her General Practitioner (GP) and assigned to a specific center for the diabetes pathology when the disease status deteriorates. Moreover, even if followed in a specific center, the patients may undergo visits or laboratory tests elsewhere through the national health care service. The administrative data allow a broader view of

the patient's history, supplying information about his treatment pathway, contacts with all the regional health services, drug prescriptions made by his general practitioner, etc.

The integration of these different data sources is particularly suitable to manage chronic patients' histories. In fact, this gathering information strategy allows the access to all the data required to trace full paths clinical but also socio-environmental evidences. This wide view of the diabetic population represents an added value to the research, and the temporal analysis techniques developed greatly benefits from it, as shown in the following chapters.

In the following there is a brief overview of the data coming from the two structures, focusing in particular on the description of the ASL dataset, whose features are less widely used in common data mining and statistical analyses.

Italian Local Health Care Agency (ASL) organization and Administrative Data. In Italy, healthcare is provided to all citizens by a mixed public-private system. The public part is the national health care service (SSN: Servizio Sanitario Nazionale), which is administered on a regional basis and by local health authorities (ASL: Aziende Sanitarie Locali). ASLs purchase the health services required for the care of their population from public hospitals or from licensed private hospitals. The ASL of Pavia data warehouse (DW) was created in 2002, when the Local Healthcare Agency started to develop and maintain a central data repository to trace all the main healthcare accesses to SSN services of the population of the area. This repository includes healthcare administrative data, which have the main purpose of providing the information required for reimbursing service providers. The main information stored in the ASL DW are related to hospital admissions, drug prescriptions, ambulatory visits, lab examinations, disease exemptions, thermal treatments, retirement homes and hospice care. Currently, The ASL data warehouse collects around 160,000 hospital admissions, 5 million drug prescriptions and 9 million outpatient visits related to about 530,000 people every year.

Figure 9 shows a schematic overview of the available administrative data in the ASL DW relevant to the research purposes. In the following these three data streams (Hospitals Admissions, Drug Prescription and Ambulatory visits and laboratory tests) are briefly described to illustrate how data are collected within the Pavia local administration, also considering the Regional [www.sanita.regione.lombardia.it/] and Italian Healthcare System Government Organization [www.salute.gov.it/].

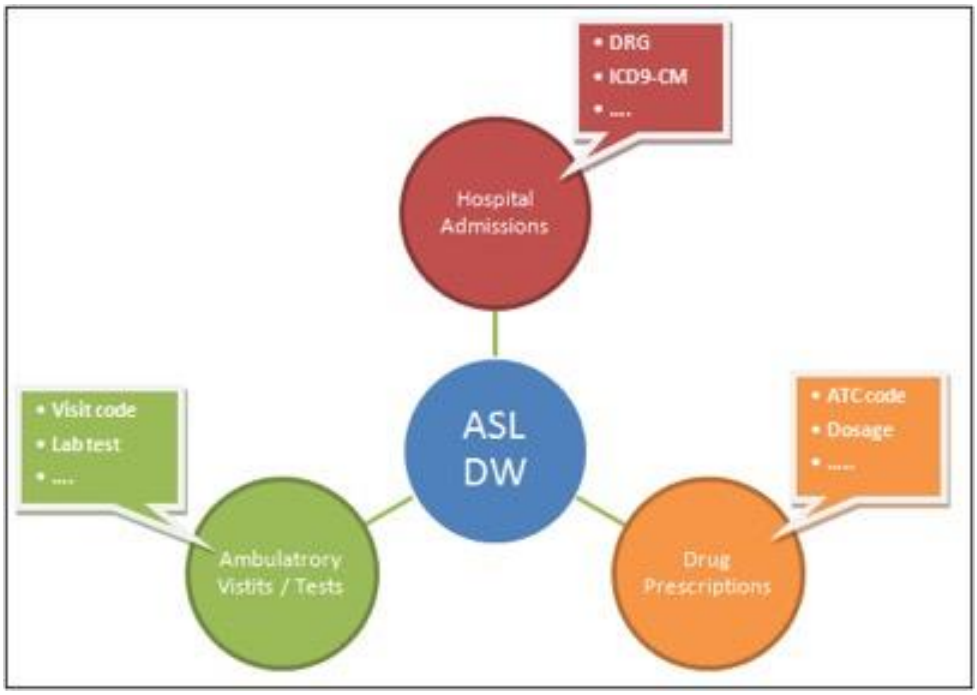


Figure 9 Schematic View of the Administrative Data Included In The ASL DW

Hospital admissions. One of the main administrative information flows collected in the Data Warehouse is the one related to the hospital discharge record. This includes all the hospitalizations occurred anywhere to a patient living in the Pavia area (local health demand) and all the hospitalizations provided to anyone by the healthcare institutions of the Pavia area (local health supply). Besides a patient identification number, the temporal details of the hospitalization are provided through the "Admission date" and the "Discharge date" fields. The clinical characterization of the hospitalization event is provided by the fields including the principal and secondary diagnoses and procedures, both encoded through the ICD9-CM classification system [http://www.cdc.gov/nchs/icd/icd9cm.htm]. The Diagnostic Related Group ("DRG"), used to synthesizes the hospitalization event through an iso-resource grouping (Fetter & Freeman 1986), is also reported. Moreover, for each hospitalization, it is possible to calculate a value for the cost, which reflects the standard expectation of resources usage and represents the charge of the ASL for the reimbursement to the service provider ("Refund").

Drug prescriptions. A further example of the main administrative information flows is represented by the database tracing all the pharmacological prescriptions required for drug purchase at the pharmacy. Over-the-counter drugs are not included, because they do not require a medical prescription to be purchased. When the drug is actually purchased at the pharmacy, each drug is identified by two different coding systems. The first is the "ATC code" (Anatomical Therapeutic Classification system), a standard coding system proposed by the World Health Organization to identify the active ingredient of each drug molecule [http://www.whocc.no/atc_ddd_index/]. This system provides also a hierarchical classification in the medical domain through five different levels of specificity. The second is the so called "AIC code" ("Authorization to the Market Introduction"), a code assigned to each product by the Italian Ministry of Health that identifies the drug from a marketing point of view, providing information about pharmaceutical company and drug package features (Abraham & Reed 2002). An additional information, strictly related to this coding system, is the "Defined Daily Dose (DDD)", defined

from the WHO as "the assumed average maintenance dose per day for a drug used in its main indication in adults" [http://www.whocc.no/ddd/definition_and_general_considera/]. This measure offers an indicator of the supposed average daily dosage of the prescribed drug. This measure allows deriving an estimation of the duration of the specific pharmacological therapy. In the data set extracted specifically for the MOSAIC project, DDD values have been already converted in days. For each prescription the DDD value indicates the number of days for which that drug should be assumed by the patient. For example, "A02BC01 prescription – DDD = 7" means that the patient should be treated with drug A02BC01 (ATC code for omeprazole) for one week. The economic dimension is provided by the drug reference price ("Cost"), extracted from the Italian national pharmaceutical price list (AIFA – Italian National Pharmaceutical Agency). This amount is the one that will be directly charged to the ASL, as the pharmaceutical therapy under medical prescription is a service under SSN coverage.

Ambulatory visits and Laboratory tests. Another of the main administrative information sources is represented by the database dedicated to the tracing of the health services provided outside the context of hospitalization, thus including ambulatory visits, specialist services and laboratory tests. The "Contact date" identifies the day when the service was provided to the patient (e.g. in case of a blood sample, it is the day when the test is performed, it doesn't matter when the sample will be analyzed by the clinical laboratory). Each health service is identified by a specific Italian National code ("Service code" - DGR n. VIII/5743 - 31/10/2007), which is based on a hierarchical classification system showing several analogies with the ICD9-CM system. Each code is associated with a reference fee ("Refund"), which determines the refund that will be charged to the ASL budget, since all these services under medical prescription are guaranteed by the Italian SSN coverage. Considering that in the clinical practice several services are usually prescribed simultaneously, in order to build a complete and enriched clinical scenario, a maximum of eight different services can be collected in the same record ("Service code1, …, Service code8"). Moreover, since some services can intrinsically involve a cyclic supply planned on the basis of multiple periodical sessions, also the number of sessions is reported for each service provided ("Quantity1, …, Quantity8").

As already mentioned, since the clinical dimension is not represented in the ASL data set, it can be highlighted that a patient underwent a blood glucose test on a specific day, but no information about the outcome of that test is available in the administrative records.

IRCCS Fondazione Salvatore Maugeri EHR. The FSM outpatient service manages more than 1,500 T2DM patients and has collected data on more than 5,500 patients over the last 12 years. FSM has outpatient and inpatient facilities, with computerized outpatient and inpatient EHRs. It has state-of-the-art laboratory and instrumental equipment and expertise for the detection and care of chronic diabetes complications, laboratories with molecular medicine equipment and performs continuous glucose monitoring (CGM) to assess glycemic control of T2DM patients. FSM clinical data are collected during day by day medical activity and stored in the hospital EHR. The data shared within the project are related to 1,000 T2DM patients and store information related to lifestyle, medical history, pharmacological treatment, results of laboratory and instrumental exams, and patients' monitoring.

Clinical and Administrative data integration. The step of integrating administrative (process) data with the available clinical information on diabetic patients represents an important step within the project, particularly with regard to the data analysis methodologies.

The advantages of this step of data integration are twofold: on the one hand, it allows monitoring some essential disease-related physiological variables, providing a feedback about the efficacy of the care delivery process for primary care. On the other hand, administrative data give the opportunity to build and monitor extended medical histories of the patients through the integration of the main healthcare data. This is an important advantage also from a research point of view, as the analysis of such integrated healthcare datasets could greatly help to gain a deeper insight into the health condition of the population and to extract useful information (e.g. the disease patterns) to support decision makers in the assessment of health care delivery process (e.g. improving the overall standards and quality of care).

The strategy adopted for the centers of the Pavia area is aimed at making available two sets of data, both integrating clinical and administrative data, made up of 1,000 patients (Diabetes Service at FSM and ASL). These 1,000 diabetic patients have been selected by FSM doctors on the basis of their clinical characteristics and of length and completeness of the available follow-up.

Two operations have been then performed in parallel: first, the clinical data of these patients are extracted from the hospital EMR and, second, their national identification numbers were sent to the ASL. On the basis of these identification numbers, the ASL will performed the extraction of the corresponding records from its own DW. The two extractions performed at ASL and FSM, are finally loaded in the MOSAIC i2b2 DW using suitable transformation procedures.

Geo-localized information. The environment's involvement in health issues has been conceptualized as the exposome (Sanchez et al. 2014). It has been recognized that, although expanding biomedical research capabilities, the combination of these wide-ranging data sources requires innovative approaches not only in data analysis but also in the visualization of their results. One of the main advantages of the MOSAIC data sets is that, thanks to the information derived from ASL DW, patients' addresses are associated to a precise municipality code. This fact allows to easily connecting environmental data to those contained in the MOSAIC DW. And, consequently, it is possible to precisely locate each patient and analyze his interactions with the territory. As matter of fact it is possible to deal with geo-referenced clinical data. Clinical and administrative data were jointly analyzed with satellite data to understand the effects of being exposed with environmental factor, such as air pollutants, the results are shown in 6.4.

3.4 MOSAIC i2b2 – Ontology definition

The technologies for implementation of the MOSAIC DW were designed to serve as a common substrate to store the data related to all the medical centers included in the project. Even though, as already stated, by the end of the research it was possible to apply the analysis methods only on the Pavia Data set, it is important to describe the adopted approach to collect data from different clinical centers, located in different countries and where different procedures and care processes are followed, and made them available for the research. Given the temporal nature of the data collected by clinical centers, they are particularly suitable to be analyzed by the developed temporal

and careflow mining techniques. One of the efforts that was carried out in parallel of the data gathering was be the implementation of suitable plugins able to transfer i2b2 queries results directly into the mining tools.

The strategy underlying the MOSAIC i2b2 concept consisted in implementing different i2b2 instances, one for each of the clinical centers involved. Each center was then supposed to establish its own necessary procedures in order to fill in its own i2b2 local data warehouse with clinical and administrative data. The ontologies of each i2b2 instance have a common core, although, since each local i2b2 framework is independent of other instances, each center have the possibility to create and insert specific concepts thus extending the original core ontology, especially in the case those concepts were peculiar of just one center.

The first step of this process was to map data available from each of the medical centers. The ***mapping procedure*** had several objectives:

- Identify the parameters that are in common between the different centers;
- For such identified parameters, share a common representation of the variables (same units of measurement, same coding system, same type of representation);
- Define a common data structure in order to:
    - build a DW for each hospital;
    - facilitate data sharing for data analysis purposes;
    - allow future DW aggregation on the basis of the adopted technologies.

The data mapping was performed through a multi-step process: first, each hospital shared a list of the available variables and parameters. An initial matching was consequently performed and comments by the hospitals were collected. A further refinement phase, based on these comments, was performed and a final agreed version of this mapping was delivered. Once the aggregation and data storing procedures were defined, the task following the data mapping was dedicated to the creation of the integrated ***core ontology*** representing common concepts. As a general guideline, those concepts that are represented in at least two out of the three participating medical centers were included in the core ontology.

According to the i2b2 structure, the main idea was to represent each clinical event happening to the patient (follow-up visit, hospitalization, lab test, drug prescription, etc) as a specific encounter in the DW, thus providing it with a start and end time and connecting it to the specific concepts related to that particular event. This is reflected in the resulting ontology structure.

Relying on these criteria, the ontology that we have defined contains seven high level concepts, which are:

- Patient data: collects all the concepts related to a patient and that are not depending on the follow-up ("static" data). This information is collected once, usually during the first encounter. It includes information related to the diabetes diagnosis, family history, vital status, ethnicity, level of education, marital status, etc.
- Contact details: collects all concepts that can be measured during an encounter. These include in particular the habits and lifestyle of the patient (eating habits, physical activity and smoking habits) and those related to the physical examination (weight, blood pressure, etc. In case an

encounter is describing a Hospitalization, also some related information such as the course and the discharge mode are included.

- Laboratory test results: collects all the concepts related to the main lab exams related to the disease and used by the MOSAIC models: Cholesterol, fasting glucose, OGTT, Fasting insulin, HbA1c, Triglycerides, Uric Acid.
- Complications and comorbidities: these are specific concepts related to the diabetic disease that the medical centers always collect. They both can arise during the disease course, being directly related to the disease itself (complications) or have an external cause (comorbidities)
- Drugs data: collects all the data related to the pharmacological therapy prescribed to the patient within health centers (Therapy prescriptions) and retrieved from administrative flows (Pharmaceutical Data). The concepts included in the Drugs section are related to the most common therapeutic solutions exploited for diabetic patients. Drugs are represented through their active principle descriptions and their ATC code (Anatomical Therapeutic Chemical Classification System).
- ICD9-CM: all the concepts related to the ICD9-CM coding system. ICD9 codes can be related to several types of events, such as hospitalizations, outpatient services, follow-up visits, comorbidities, complications, etc. Creating a specific class for those concept makes it very convenient to query the DW for specific diseases.
- Visits and Follow-up: describes the type of encounter that the patient undergoes
- Living area: geo-referenced information about patient's living area (from administrative flows)
- Temporal Abstraction: collects all the concept that derive from a temporal abstraction procedure performed on raw patient's data. Temporal abstractions are generated by a dedicated module integrated in the MOSAIC tool, which automatically retrieves time point data from the DW and computes interval-based qualitative abstractions on some variables such as HbA1c, weight and diet.

In Appendix A each class of the ontology is described in detail.

3.5 Risk score and complexity index

The high-level concept referred as Patient data include two concepts which are derived from literature and previous projects and integrated into the DW for analysis and decision support purposes. They are a cardiovascular risk score, calculated through the "Progetto Cuore" algorithm (Palmieri et al. 2004) and defined thresholds, and the patient status evolution levels, illustrated by means of complications onset and related hospitalizations. Each time new data is uploaded into the i2b2 DW, both these concepts are synchronized. They are computed via R scripts on the basis of the data gathered and simultaneously uploaded into the DW.

Progetto Cuore, Cardiovascular Risk. The "Progetto Cuore" indicators are defined on the basis of Italian population characteristics. The "Progetto cuore" algorithm calculates the cardiovascular risk (CVR) score for each time the appropriate clinical measures are obtained during the registered encounters. The "Progetto Cuore" is a project funded by the Italian Ministry of Health devoted to estimate the impact of cardiovascular diseases in the general population through a board of indicators like prevalence, incidence and mortality rates. The score has been selected after an evaluation of other cardiovascular risk indicators, like the Framingham (Kannel & McGee 1979)

and the QRISK (Hippisley-Cox et al. 2010), as complete information for its calculation is currently collected within the hospitals and thus the variables needed to apply it are present in the DW. The "Progetto Cuore" algorithm presented in Figure 10 gives as result CVR scores for each encounter, which allow to stratify the population in classes of risk. The application of the algorithm results in a set of continuous values representing the risk of macrovascular event at 10 years from when the set of clinical measures are taken. The values are discretized on the basis of the thresholds indicated by the project [http://www.cuore.iss.it/valutazione/carte.asp], as shown in Figure 11.

1 - [S(t)] * {EXP [**b1** * AGE (years)
+ 0.0 (if SBP (mmHg) <= 129) +
+ **b2** (if 130 <= SBP (mmHg) <= 149) +
+ **b3** (if 150 <= SBP (mmHg) <= 169) +
+ **b4** (if SBP (mmHg) >= 170) +
+ 0.0 (if Total Cholesterol (mg/dl) <= 173) +
+ **b5** (if 174 <= Total Cholesterol (mg/dl)<= 212) +
+ **b6** (if 213 <= Total Cholesterol (mg/dl)<= 251) +
+ **b7** (if 252 <= Total Cholesterol (mg/dl)<= 290) +
+ **b8** (if Total Cholesterol (mg/dl)>= 291) +
+ **b9** (if diabetic disease = yes) +
+ **b10** (if smoking habit = yes) - G(u)]}

Parameters for males	Parameters for females
S(t)=0.942	S(t)=0.986
b1=0.083	b1=0.088
b2=0.313	b2=0.212
b3=0.650	b3=0.614
b4=0.952	b4=1.073
b5=0.062	b5=0.025
b6=0.233	b6=0.147
b7=0.411	b7=0.163
b8=0.869	b8=0.437
b9=0.566	b9=0.499
b10=0.489	b10=0.715
G(u)=5.024	G(u)=4.978

Figure 10 The "Progetto cuore" algorithm –the figure shows the algorithm for cardiovascular risk computation (on the left) and the values of the parameters for male and female subjects (on the right)

Figure 11 CVR thresholds – from Progetto Cuore project site

Level of Complexity. The complexity of T2DM pathophysiology, together with the high number of heterogeneous factors (clinical, behavioral) influencing patients' profiles, pose the need of an exhaustive stratification to achieve an adequate description of disease status. Patients can be though classified according to the developmental stage of their disease in time. Thanks to the suggestions of MOSAIC project members derived from a previous experience within the project METABO (Georga et al. 2009; Guillén et al. 2011) four levels of complexity (LOC) patients might go through on the basis of complications and related hospitalizations, during his/her whole history from the

diagnosis, were identified. The first stage identifies Stable patients, who do not suffer any complications yet. The disease progression and complexity increasing are described by First Level, to which belong patients affected by only one complication, Second Level, representing the arising of more than one complication in multi-pathological patients and Third Level, which include those patients likely to suffer hospitalizations due to the status of their diabetes-related complications. For each patient, a LOC observations include the LOC value and its duration in terms of start and end date.

As for the Progetto Cuore score, the input data necessary to compute LOC values are gathered and stored in the i2b2 DW. The variables and the reference to the corresponding ontology concepts are listed in Table 1.

Variable	Information needed	Concept INPUT
Diagnosis	Date	..\Patient_Data\Anamnesis\Diagnosis of Diabetes\
Complications	Description + Onset Date	..\Complications\
Hospitalizations	ICD9-CM code + Date	..\icd9\Diagnoses\
Hospitalizations	Course of Hospitalization	..\Hospitalization\Course\

Table 1 Input data needed for the computation of the LOC

The pseudo code used for LOC computation is shown in Figure 12. In order to assess if a patient switched or not in Stage 3 (hospitalization related to a T2DM complication), the ICD9-CM data contained in the DW were preprocessed through the Clinical Classifications Software (CCS) for ICD-9-CM (Elixhauser et al. 2014).

Table 2 (referred to as Matching table in the pseudo code) indicates if a certain group of CSS codes (columns) can be related to a complication (rows) that arises before the hospitalization event. Value 1 indicates there is a match (e.g. the onset of neuropathy can lead to a hospitalization related to "Nervous System Sense" organs diseases).

```
PSEUDO CODE
PRE-PROCESS
DATA GATHERING from i2b2: Diagnosis, Complications, Hospitalization
HOSPITALIZATION → from ICD9-CM to CSS LABELS
DEFINE CORRESPONDENCES BETWEEN COMPLICATIONS and HOSPITALIZATIONS in the MATCHING.TABLE*
DEFINE EVENTS
Each events is defined by DATE, DESCRIPTION
EVENTS =        {DIAGNOSIS,
                 COMPLICATION.NEUROPATHY, COMPLICATION.STROKE [....],
                 HOSPITALIZATION.LIVER.DISEASE, HOSPITALIZATION.ENDOCRIN.DISEASE[....]}
EVENT.LOG = { ID PATIENT, EVENT, DATE }
SORT EVENTS BY DATE and FOR EACH PATIENTS ASSIGN SEQUENCE TO HIS/HER EVENTS FROM 1 to count(EVENTS)
EVENT.LOG = { ID PATIENT, EVENT, DATE, SEQUENCE }

LOC COMPUTATION find_loc_time_intervals(EVENT.LOG, MATCHING.TABLE)
For each patient create EVENT.LOG.ID          % subset of the EVENT.LOG
        For each EVENT ∈ EVENT.LOG.ID
                if EVENT == "DIAGNOSIS"
                        LOC ← "STABLE",
                        LOC.START.DATE ← DIAGNOSIS.DATE
                else if EVENT == "COMPLICATION%" and SEQUENCE == 1
                        LOC ← "STAGE1",
                        LOC.START.DATE ← COMPLICATION.DATE
                        LOC.END.DATE (previous step) ← COMPLICATION.DATE
                        STORE COMPLICATION TYPE
                else if EVENT == "COMPLICATION%" and SEQUENCE > 1
                        LOC ← "STAGE2",
                        LOC.START.DATE ← COMPLICATION.DATE
                        LOC.END.DATE (previous step) ← COMPLICATION.DATE
                        STORE COMPLICATION TYPE
                else if EVENT == "HOSPITALIZATION" and STORED COMPLICATION at STAGE1 or STAGE2 match
                                                        HOSPITAZIATION in the MATCHING.TABLE
                        LOC ← "STAGE3",
                        LOC.START.DATE ← HOSPITALIZATION.DATE
                        LOC.END.DATE (previous step) ← HOSPITALIZATION.DATE
                        LOC.END.DATE ← today
        End
End
```

Figure 12 Pseudo code for computing the level of complexity

	Circulatory System	Nervous System Sense.organs	Skin Subcutaneous Tissue	Endocrine Nutritional Metabolic	Genitourinary System	Digestive System	Infectious Parasitic
Occlusion and stenosis of carotid artery	1	1	0	1	0	0	0
Chronic ischemic heart disease	1	0	0	1	0	0	0
Acute myocardial infarction	0	0	0	1	0	0	0

Neuropathy	0	1	1	1	0	0	1
Peripheral vascular disease	1	0	0	1	0	0	0
Retinopathy	0	1	0	1	0	0	0
Fat Liver Disease	0	0	0	1	0	1	0
Nephropathy	1	0	0	1	1	0	1
Angina	1	0	0	1	0	0	0
Stroke	1	0	0	1	0	0	0
Diabetic Foot	0	0	0	1	0	0	0

Table 2 Matching table, ROWS represent complications and columns hospitalizations, value (0-1) indicate if there is a correspondence between these two kind of events

3.6 The Pavia Data Set

In the following chapters the analysis methods are applied and evaluated on the data collected in a population of 1030 T2DM patients both by the IRCSS Fondazione Salvatore Maugeri (FSM) and by the Local Healthcare Agency of Pavia (ASL). This data set is referred as the Pavia Data set. This paragraph illustrates some descriptive analysis on this population.

Figure 13 shows the age distribution in the Pavia Cohort for Female (426 subjects, 41% of the population) and Male (604 subjects, 59% of the population). By the end of the study, 7 patients deceased.

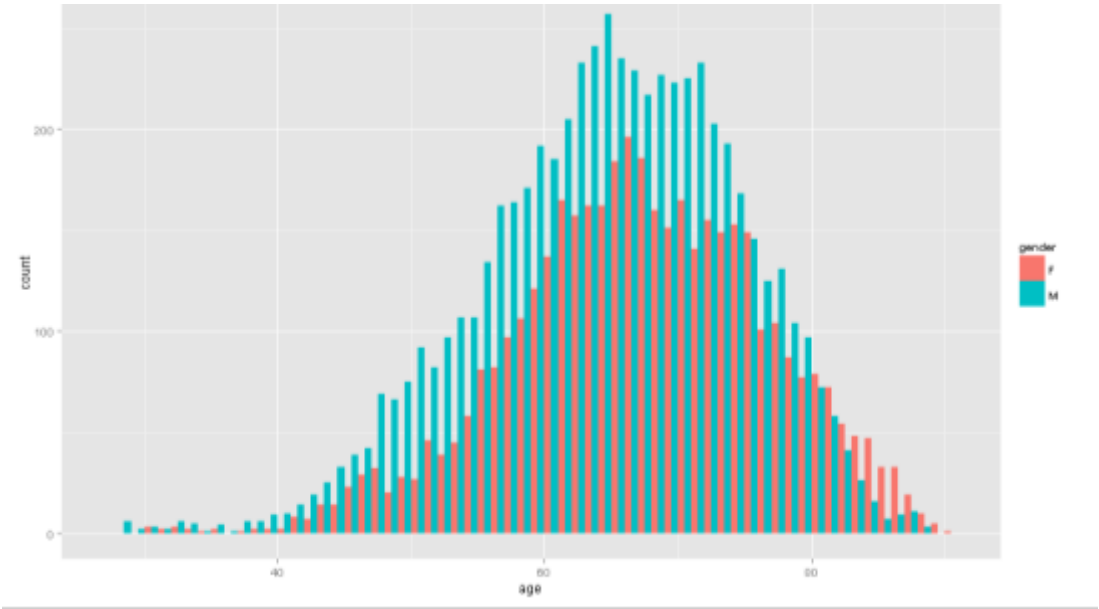


Figure 13 - i2b2 Web client, Analysis tool.

Table 3 shows means and standard deviations in male and female of some of the variables considered in the analyses, and includes demographic data (age and time from diagnosis when patients underwent to the first visit near the FSM hospital) and clinical data from the FSM hospital EHR: Body Mass Index, Glycated hemoglobin (Hba1c), lipid profile (Triglycerides and Total Cholesterol), Systolic Blood Pressure and Smoking habit.

	Female	*Male*
Number of Patients	426	604
Age at the First Visit	mean(63.97) sd 10.02	mean(61.13) sd 9.52
Diabetes Duration at the First Visit (Years)	mean(6.93) sd 9.34	mean(5.25) sd 6.55
Smoking (Yes)	110(25.97%)	172(65.44%)
BMI	mean(30.38) sd 5.62	mean(28.51) sd 4.29
Hba1c (mmol/mol)	mean(55.22) sd 12.9	mean(54.25) sd 13.36
Triglycerides (mg/dl)	mean(133.09) sd 66.38	mean(131.44) sd 74.19
Total Cholesterol (mg/dl)	mean(194.65) sd 31.96	mean(181.33) sd 31.81
Systolic Blood Pressure (mmHg)	mean(134.54) sd 14.73	mean(132.13) sd 13.98

Table 3 – Clinical variables in male and female

Table 4 shows the number of patients and the percentage over the total number of subjects in the cohort, affected by T2DM complications. The table reports all the complications without separating those detected before and after the first visit near the FSM hospital. Each patient can be counted more than one time if affected by more than one complication.

Type	Complications	Patients with complication	% on the total population
Macrovascular	Acute myocardial infarction	111	10.78%
	Angina	10	0.97%
	Chronic ischemic heart disease	183	17.77%
	Occlusion and stenosis of carotid artery	267	25.92%
	Peripheral vascular disease	95	9.22%
	Stroke	41	3.98%
	TOTAL	**707**	**68.64%**
Microvascular	Nephropathy	128	12.43%
	Neuropathy	145	14.08%
	Retinopathy	130	12.62%

	TOTAL	**403**	**39.13%**
Not Vascular	Diabetic Foot	33	3.20%
	Fat Liver Disease	238	23.11%
	TOTAL	**271**	**26.31%**

Table 4 - Complications

The distribution of the cardiovascular risk value in the population, calculated by mean of the Progetto Cuore score, for each encounter near the FSM hospital, is shown in Figure 14.

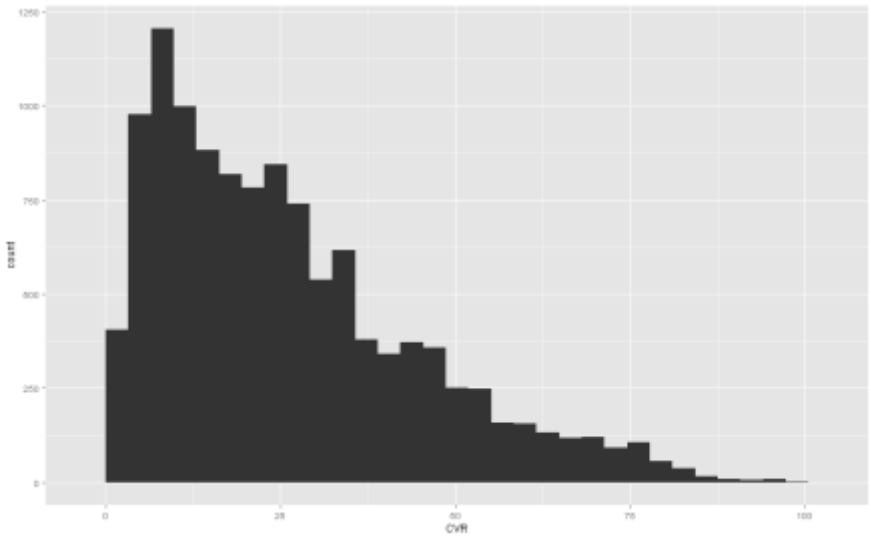


Figure 14 - CVR distribution.

Table 5 illustrates the number of patients (and the percentage on the total population) that purchased at least once one the listed drugs, as classified via ATC codes in the data warehouse. As extensively discussed in 5.2 this information is retrieved from the ASL data stream that records drug purchases near the Pavia area pharmacies and it is used as proxy for patients' treatments.

	Patients treated	% on the total population
Antihypertensive	792	76.89%
Diuretics	381	36.99%
Antithrombotic	745	72.33%
Lipid_Lowering	693	67.28%
Diabetes Treatment		
Biguanides_Sulfonamides	13	1.26%
DPP-IV_Inhibitors	66	6.41%
Glp-1_Analogues	225	21.84%
Insulin	244	23.69%
Metformin	756	73.40%
Metformin_DPP-IV_Inhibitors	67	6.50%
Metformin_SGLT2_Inhibitors	1	0.10%
Metformin_Sulfonamides	154	14.95%
Metformin_Thiazolidinediones	35	3.40%
Repaglinide	238	23.11%
SGLT2_Inhibitors	6	0.58%

Sulfonamides	327	31.75%
Sulfonamides_Thiazolidinediones	2	0.19%
Thiazolidinediones	61	5.92%
Alpha_glucosidase_inhibitors	215	20.87%

Table 5 – Drug treatments in the population

CHAPTER 4

4 Data Analytics (Aim 2)

The collection of longitudinal data is the essential requirement for the detection of temporal patterns, which can be used to dynamically reassess risk categories during the follow up and to estimate the probability of complications that might arise during the process of care.

The applied data gathering strategy allowed mapping clinical databases to reach common parameters representation. From a technological perspective, the adopted solutions provide medical centers with a common substrate to store their data. Aggregating the data repositories under a common framework enables to perform integrated queries while maintaining the data inside each hospital facilities.

The data collected within the MOSAIC project supports the study of the events characterizing the evolution of the disease after diagnosis through T2DM patients' data coming from heterogeneous databases. The implemented methods need to explicitly consider the longitudinal nature of data, in relation to T2DM evolution, especially in terms of complications and assessment of metabolic control, both at patient and population level. Moreover, to be proficiently integrated in a CDSS, these analyses need to be defined in collaboration with the clinical partners, considering their specific interests and the characteristics of the available data.

The ***objective*** of this phase of the research is to finalize the implementation of an analysis framework able to (i) handle temporal multivariate data, (ii) implement longitudinal models to extract meaningful information from these data and dynamically reassess risk categories during the follow up and (iii) deliver prediction models for decision support that take into account these characteristics.

To this end, the applied state-of-the–art temporal data mining methods were focused on the capability of recognizing changes in time (patterns, careflows) and suggest that a patient condition is worsening, in order to identify sub cohorts of patients who meet specific conditions that are relevant to clinical actions. This Aim, and all the tasks described in the following chapters, represents the ***core scientific efforts*** of this research program.

Aim 2 fits the concepts defined in ***Key Area 2***, which claims the need of defining computationally manageable phenotypes, which simulate disease evolution. It is the core of all the research efforts made to transform data into knowledge, especially thanks to the development of a novel algorithm for careflow mining that defines an innovative way of performing electronic phenotyping with a temporal dimension. The activities related to this research aim are related to different subtasks, which originate by several motivations and are exploited on the basis of different type of the gathered data thanks to the activities performed within Aim 1.

It is important to remark how the data collected within the i2b2 framework allowed the implantation of a scalable informatics platform, which enables the use of existing clinical data for discovery through the application of the temporal analytics methods. In the context of the Learning

Healthcare Cycle, this is the step that triggers the transformation of data into knowledge (Care informs Research). The implemented data layer formalizes events that patients undergo and facilitates their representation in various clinical and research scenarios. The availability of a common data model is the pillar for the application of various methods to convert datasets into clinically relevant knowledge and it is the first step in improving diagnostic accuracy and achieving precision medicine.

Figure 15 shows the division of Aim 2 in sub aims and tasks.

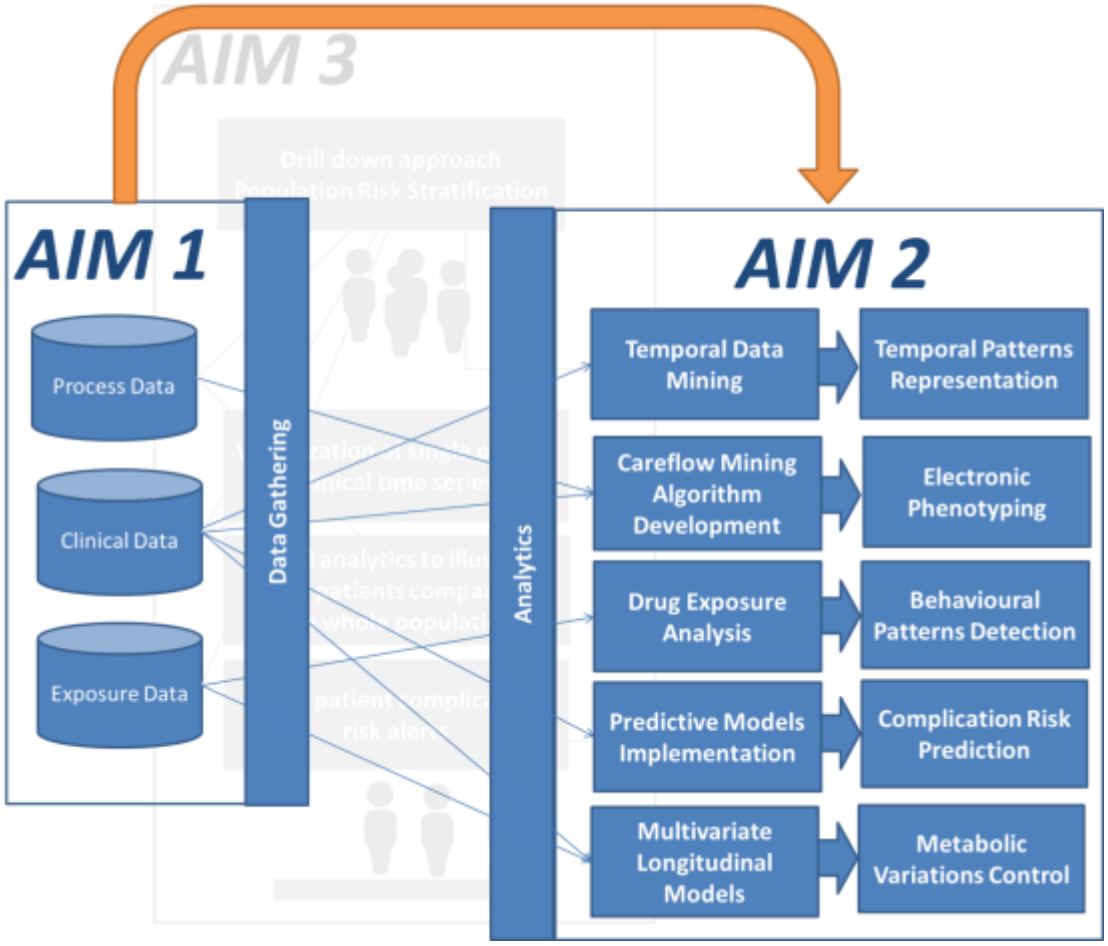


Figure 15 Aim 2 sub aims and Aim 1 relation

For the analyses purposes data had been conceptually classified as: ***Process data,*** which include the events a patient undergoes during his or her medical history, specifically recorded for business purposes. Examples of events are inpatient stays, outpatient visits or discharge diagnoses and procedures derived from ICD9 – CM codes. Process data are derived from billing information, and can be also referred to as administrative data in a clinical setting. Administrative data can be used as a proxy for pathological conditions. For example, in the diagnosis process the overlap of billing and clinical information is clear: when clinicians apply a diagnosis code to a patient they are applying a procedure, but this action also constitutes an event as it changes the patient state and it is a meaningful event in the disease evolution from the clinical point of view. ***Clinical data*** include clinically relevant variables, recorded during hospital-related events, such as laboratory test results and surgical pathology results. While this first distinction of process and clinical data well depict the whole corpus of data collected in the data gathering phase, in the analysis context it is important

to provide a further category: exposure data, which can be derived both from clinical or billing data stream, or from other sources of information like environmental factors or social analysis. In epidemiology ***Exposure data*** refer to the potential causal characteristics as exposure (Rothman et al. 2014), and are identified to distinguish them from outcome data. Exposure can refer to behavior, treatments, trait, exposure to environmental factor (Sanchez et al. 2014).

The next two chapters describe the methodological innovations, how the developed methods are applied to the type of data collected in the i2b2 DW and their technical implementation to achieve Aim 2.

The chapter ***Longitudinal Data Analytics (Aim 2a)*** illustrates the implementation of several techniques to analyze longitudinal data. It is introduced the use of ***temporal data mining*** techniques to detect the temporal events that characterize the clinical history of patients after diagnosis. Temporal abstractions have been performed on the basis of clinical time series extracted from laboratory test data and physiological parameters recorded in the EHR. As this task implements established methods, they are not in depth discussed. Although a detailed description of their implantation is given in the CDSS description chapter.

The main focus of the chapter is on clinical pathways discovery indeed. This task is dedicated to the development of a new ***careflow mining algorithm*** devoted to detection of complex patterns. On the basis of the detected patterns it is possible to reconstruct the clinical pathways patients undergo during the whole process of care, and to identify clusters of patients with similar care histories so to re-assess their risk profiles accordingly. This last feature is related to patient stratification and electronic phenotyping activities. The discovery of the clinical pathways is based both on the jointly use of process and clinical data. Careflow are mined on process data and used to stratify the population, and then clinical data are exploited to prove the clinical relevance of the data-driven phenotype.

The last part of the chapter describes the performed ***drug exposure*** analysis by ***mining patterns*** from purchasing data. The discovery of well-defined patterns behaviors derived from administrative data streams is one of the central objectives of Aim 2. The detection of drug consumption patterns allows stratifying the population on the basis of subjects purchasing attitude, showing how secondary data can be exploited as a population marker. Process data, which in this case can be defined as exposure, allow detecting the patterns of drug purchases. These findings are used as a proxy of the complexity of clinical conditions, in order to detect specific behaviors over time, understand how purchasing behaviors affect the disease evolution and enhance patient monitoring.

The chapter ***Risk Models for T2DM Complications and Metabolic Control Variations (Aim 2b)*** illustrates the implementation of models to assess the risk of devolving worsening condition during the disease evolution. Firstly, the chapter illustrates the implementation of a set of models for the prediction of T2DM ***microvascular complications***. To make available indicators of the risk of developing microvascular complications, exposure data collected from the hospitals' EHR have been exploited to develop multivariate risk models of the onset of retinopathy, neuropathy and nephropathy. The developed models incorporate information related to physical, metabolic, phenotypic and lifestyle factors and have been integrated in the CDSS. This chapter also describes

other ***multivariate longitudinal models*** implemented to study the evolution of the ***metabolic control*** during the disease progression. To better explore the human metabolic response in diabetes, further efforts had been focused on exploring advanced mining methods able to explicitly take into account the longitudinal nature of the available data. The disease evolution has been in depth analyzed through the exploitation of information derived from environmental context and the behaviors of the patients. Cox regression analysis, Continuous Time Bayesian networks, Hierarchical Bayesian models and Remote Sensing analysis methods have been exploited to respond to specific clinical questions, especially for the valuation of the metabolic control. However, as the clinical research questions that guided these further analyses arose during the development of the CDSS, taking into account that the analyses and the system implementation have to concurrently move forward, these methods and their findings has not yet been implemented in the tool.

CHAPTER 5

5 Longitudinal Data Analytics (Aim 2a)

In the complex scenario of a patient followed inside and outside the hospital, what is recorded is isn't only strictly clinical information. Both hospitals' EHRs and local healthcare services collect some administrative information. If joined to clinical data, administrative data represent an added value to the process of knowledge discovery, as they contribute to build up a whole picture of patients' histories. Administrative data contain the collection of all the accesses a patient performs to the national healthcare system: hospital admissions, drug purchases, outpatient visits, etc. Given the purposes they are originally collected for; this data does not contain medical information. Administrative record may report that a patient has been admitted to the hospital with a specific diagnosis and that some procedures and lab tests have been performed, but the specific results of such tests are not reported. The integration of administrative and clinical data is crucial to get the best knowledge out of both.

The implementation of the MOSAIC i2b2 framework supports research through a robust system, which integrates clinical, administrative and environmental data, allowed supporting the application of advanced temporal data mining techniques. Thanks to the possibility to query a system where administrative data had been coded and related to clinical procedures, it is possible to identify patients with a specific history of the disease and to retrieve from the fact table structured data logs suitable to perform temporal and process data mining analyses to highlight meaningful clinical careflow patterns. Although this process in not straightforward, since the structure of the raw data is naturally different between the two. While clinical time series are stored as time-stamped data associated to a quantitative measurement (temperature, blood pressure, creatinine value, etc.), administrative data are stored as temporal sequences of events.

In ***Temporal Data mining***, an event is in general defined as a temporal variable that is associated to a timestamp or interval of occurrence. A sequence of events is defined as a list of events, where each event is associated to the same transaction or individual (Mannila et al. 1995; Kam & Fu 2000). When dealing with clinical data, the individual is usually the patient. Differently from time series that contain only time-stamped data (i.e. a measurement is collected at a specific time point) temporal sequences of events can contain both time-stamped events and events with duration. The duration of an event can be defined on the basis of the temporal granularity a specific set of data is collected with. The granularity represents the maximum temporal resolution used to represent an event (Combi 2004; Bettini et al. 1996). Accordingly to this definition, it is possible state that, on the basis of a specific granularity, sequences of events can contain both zero-length events and events with non-zero duration. In addition to the described differences in the purpose underlying the collection of clinical and administrative data, there are thus also some differences in the way they are represented. As already underlined, the joint analysis of clinical and administrative data might provide a deeper insight into the temporal mechanisms underlying patients' histories.

Given the nature of the different types of data a first step is to integrate them accordingly to reach a common representation format enabling to treat them in a uniform way. ***Temporal Abstraction*** (TAs) implementation within the CDSS tackle the issue of including a functionality that allows clinicians to look at the clinical temporal data of the patients both as raw time series and using an abstract representation able to highlight the occurrence of specific patterns in data (described in 7.2.2)

5.1 Careflow mining for electronic phenotyping

EHRs are useful tools in clinical practice, though they store and visualize data in a way that is not useful to make clinical decision during care delivery neither, at a higher level, to manage cohorts of patients. Building a system able to detect tailored sub cohort of patterns and allow representing characteristics that make these cohorts similar can be fundamental in CDS. Clinicians can explore the disease evolution in specific group of patients and derive specific insight for better treatments of them.

Patients can be stratified into specific sub cohorts according not only to their actual disease conditions but in term of the evolution of the disease. This is crucial to reach a more advanced management of the disease, which takes into account the effects of changes in treatments, patients' contacts with the health care systems or the way clinical procedures are performed.

Temporal analytics serve to discovery novel insights in care patterns; the further step is to exploit novel techniques to get insights in the whole diseases evolution through the mining of clinical pathways.

A novel ***Careflow Mining algorithm*** had been integrated into the CDSS as the central component of the so called ***Drill Down*** approach, which allows stratifying the population on the basis of its dynamic features and shown complications distributions in each identified sub cohort.

5.1.1 First Results and Limitations

In a first phase of the research, which results has been published in (Dagliati, Sacchi, Cerra, et al. 2014), the Heuristics Miner algorithm, as one of the most robust algorithms implemented in the ProM framework, was selected to preliminary investigate clinical processes. The main weakness of this algorithm is that it generates complex and not readable, so called spaghetti-like (W. van der Aalst 2011), models for barely structured processes, which are the most frequent in the healthcare domain. To overcome this first drawback, as first attempt the analysis relied on event logs composed by activities processed through TAs, which allow reducing the number of states while considering abstract variables.

Patients were stratified on the basis of their ***cardiovascular risk*** by means of the "Progetto Cuore" algorithm in three classes of risk: High, Moderate and Low. ***Basic TAs*** were then exploited to shift from a time point quantitative representation to qualitative interval-based description of ***glycemic control*** over time. State TAs intervals were determined on the basis of glycaemia (mg/dl) physiological thresholds fixed by the European guidelines on diabetes care [http://www.easd.org]. Trend TAs were applied to represent increase, decrease, and stationary patterns in the clinical time series through specifically fixed parameters. A further step was to create more expressive TAs,

which included both state and trend abstraction. Trend events over time intervals ranging from the one visit to the next one were detected, where the state abstraction was already mined for the second visit. In this way, trend information was associated to a consequent state in order to detect glycemic variations and it was possible to obtain mixed events that illustrate both current state of the glycaemia value and its following variations, as shown in Figure 16. Event logs were built on the basis of data preprocessed and represented through mixed State and Trend TAs and analyzed through ProM with the Heuristics Miner algorithm.

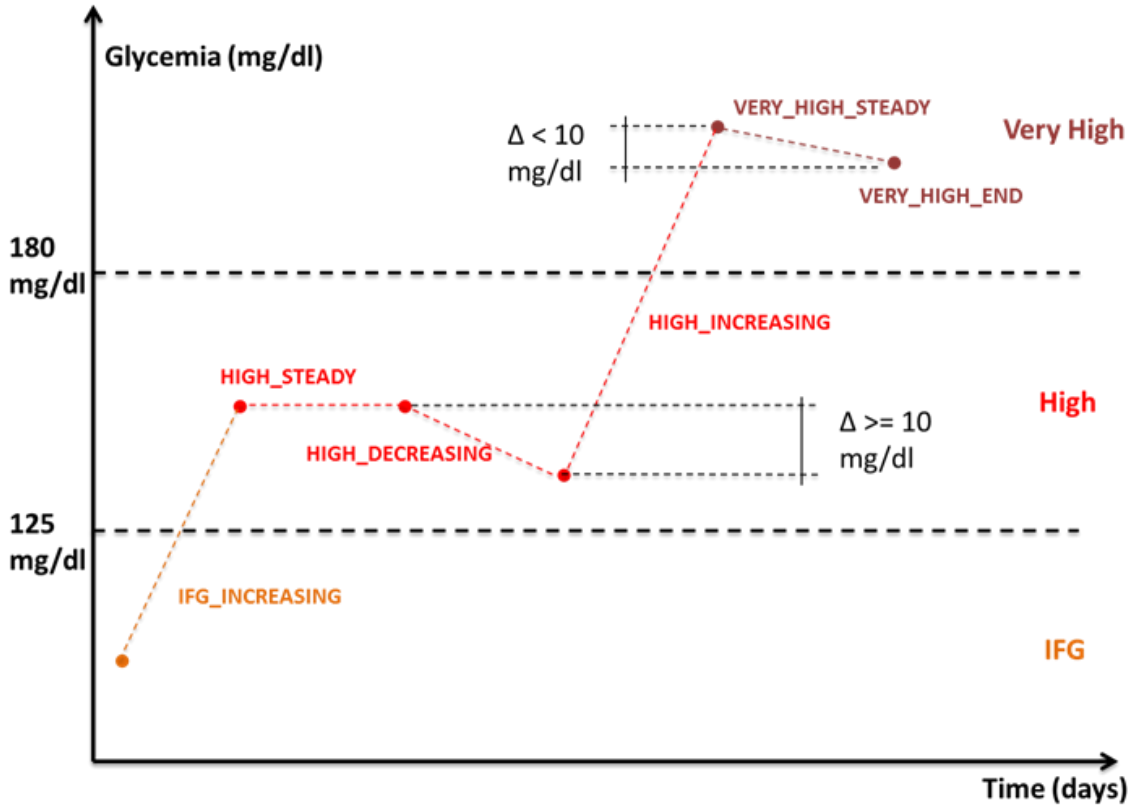


Figure 16 - Creation of complex TAs representing the coexistence of states and trends in a time series

The goal of first part of the research was to preliminarily investigate the potentials of process mining techniques to tackle the analysis of multivariate temporal data and detect interesting healthcare pathways. Some interesting insights were founded. Physicians for example found interesting that the patient with low cardiovascular risk show the poorest glycemic control. Transitions start with High blood glucose values and end in the same states or worse. A possible explanation is that these patients are less strictly controlled.

These results show how heterogeneous data can be processed and analyzed to derive consistent representations and to reconstruct clinical pathways. This approach though has several limitations; some of them were derived by the initial setup of the analysis and by the nature of the available data. For example, the cohort stratification was based only on clinical variables (glycaemia values) and not on process events, and the re-assessment of risk profiles was performed separately through a coarse and static stratification on the basis of a predefined index (the "Progetto Cuore" cardiovascular risk calculator), which is widely used in clinics but does not take into account the evolving characteristics of the specific cohort.

Further limitations of this approach are intrinsic of classic process mining techniques in providing explanatory and complete portraits of patients care and in performing electronic phenotyping. Even if spaghetti like issues was tackled through a more informative representations of events, like has been done through TAs, current process mining methods are not able to give to the specific timing of events occurrence in patient histories. Available algorithms consider frequent occurrences of sets of events within sequences or event logs, without taking into account the moment of occurrence of such events within the real temporal history of the subjects. This does not fit well the healthcare context, where the same type of event has a different clinical meaning if it occurs at the beginning or the end of the treatment. Clinicians explicitly request to considered as different the events of the same type (i.e. the same label) occurring at different time points of a clinical history, while available process mining algorithms don't allow, for example, to start from the first events of the sequences of all the patients.

Moreover, process mining algorithms typically consider all the events as process data, without the possibility to enrich them with clinical time series. In this example, the pathway mined with the Heuristic Miner algorithm were derived by clinical time series preprocessed via TAs, though ProM framework did not have the capability of showing the time of transitions between events neither the possible dependencies between events.

Beside some technical and visualization issue, like the difficulties of easily recognizing and show a specific pathway in cyclic graphs, like the results of the Heuristics Miner, one of the current requirements in electronic phenotyping is to leverage both on structured and unstructured data, to provide personalized recommendations and CDS. At this point of the research the necessity to implement a novel approach was clear. An approach that allow to facilitated the identification of the most frequent pathways in patients' care the enrichment of the results using temporal and clinical information

5.1.2 A novel algorithm for careflow mining

Taking into account the limitations mentioned so far, in order to apply careflow mining in the context of electronic phenotyping, a new algorithm that can process data related to health care events and enrich the mined patterns with clinical data was developed. This approach couples sequential pattern mining (used to extract careflows) with temporal analysis to characterize the transitions between events, thus combining the approaches taken in (Conway et al. 2011; Post, Kurc, Willard, et al. 2013) with the strategy that is reported in (Albers et al. 2014; Hripcsak et al. 2015).

The developed analysis pipeline avoids some of the drawbacks of applying temporal data mining and process mining algorithms to clinical data, like the already mentioned spaghetti-like networks of events, which can be increased by high variability of the population (W. van der Aalst 2011). The approach is able to exploit careflow mining on heterogeneous data while reducing sources of variability and improving the clarity and readability of results.

The developed approach is aimed at ***automatically deriving careflows*** and ***comparing*** them using clinical information of interest. Once careflows are extracted, it is possible to rely on them to stratify the population by ***undirected dynamical phenotyping***, described in (Albers et al. 2014). Distinguishing novelties of this approach in this context are related to the fact that the developed method takes into account the temporal nature of the data, explicitly including both

process and clinical information. This is a crucial feature, as most electronic phenotyping approaches, though considering the temporal nature of data, have so far been focused on clinical data only (Hripcsak & Albers 2012; Hripcsak et al. 2015; Wang et al. 2015). While such data are quantitative in nature, and are thus more suitable to be analyzed using time series analysis methods, process data provide insight into the sequence of events that patients undergo during their care. Furthermore, the inclusion of temporal and clinical information in an analysis oriented to process data is also new (Mans et al. 2015). Different durations of entire careflows or even of single events could be used to characterize the population undergoing a specific path.

Through following paragraphs, the following ***notations*** are used:

- an ***event E*** is defined by a pair ($[t_{start},t_{end}]$,L), where $[t_{start},t_{end}]$ is the time interval of occurrence of event E and L is the label that identifies the event (e.g., hospitalization). The definition $t_{start} \leq t_{end}$ makes it possible to include events of zero-duration according to the temporal granularity used to (Lucia Sacchi et al. 2015c), to consider a general case of an event with represent the data (Bettini et al. 1996). This definition has been extended from duration = 0, as well as events with duration > 0. In the case where t_{end} is not known, there are three choices based on the knowledge of the event type. In the first case, where an event is known to have zero-duration t_{end} is set to t_{start}. In the second case, where an event is known to have a duration greater than zero and is not followed by another event, then t_{end} is set to the end of the observation period, effectively right-censoring at this point. In the third case, where an event is known to have a duration greater than zero but is followed by another event, then t_{end} is set to t_{start} of the subsequent event.
    - a ***sequence*** of events is the collection of all the events happening to a patient, ordered by t_{start}. The length of a sequence is defined as the number of events in that sequence. The sequence of events for one patient constitutes his/her history. A generic history, including k-ordered events $E_1,E_2 \ldots E_k$, is represented as $H=<E_1,E_2 \ldots E_k>$. For example, it can be the sequence of hospitalizations a patient underwent during the evolution of his/her disease or the procedures performed on the patient during a single hospitalization.
    - a ***careflow*** is the result of the mining performed by the algorithm. It corresponds to the generalization of the history to a population of patients.

5.1.2.1 Mining careflows: An algorithmic pipeline

Figure 17 shows the analytic pipeline of our study, which is explained in details in the next sections. The careflow mining framework works in two phases: *discovery* and *enrichment*. These phases integrate a data-driven approach with pre-existing clinical data. In the *discovery phase*, the focus is on mining and reconstructing patients' careflows from process information. This phase results in a set of sequences of events that depict the services delivered to patients, their inpatient stays, and outpatient visits. The *enrichment* of the resulting careflows takes place using both temporal and clinical information. Using this information, it is possible to better characterize the groups of patients who experience different careflows, thus enabling the assessment of the clinical relevance of the extracted patterns.

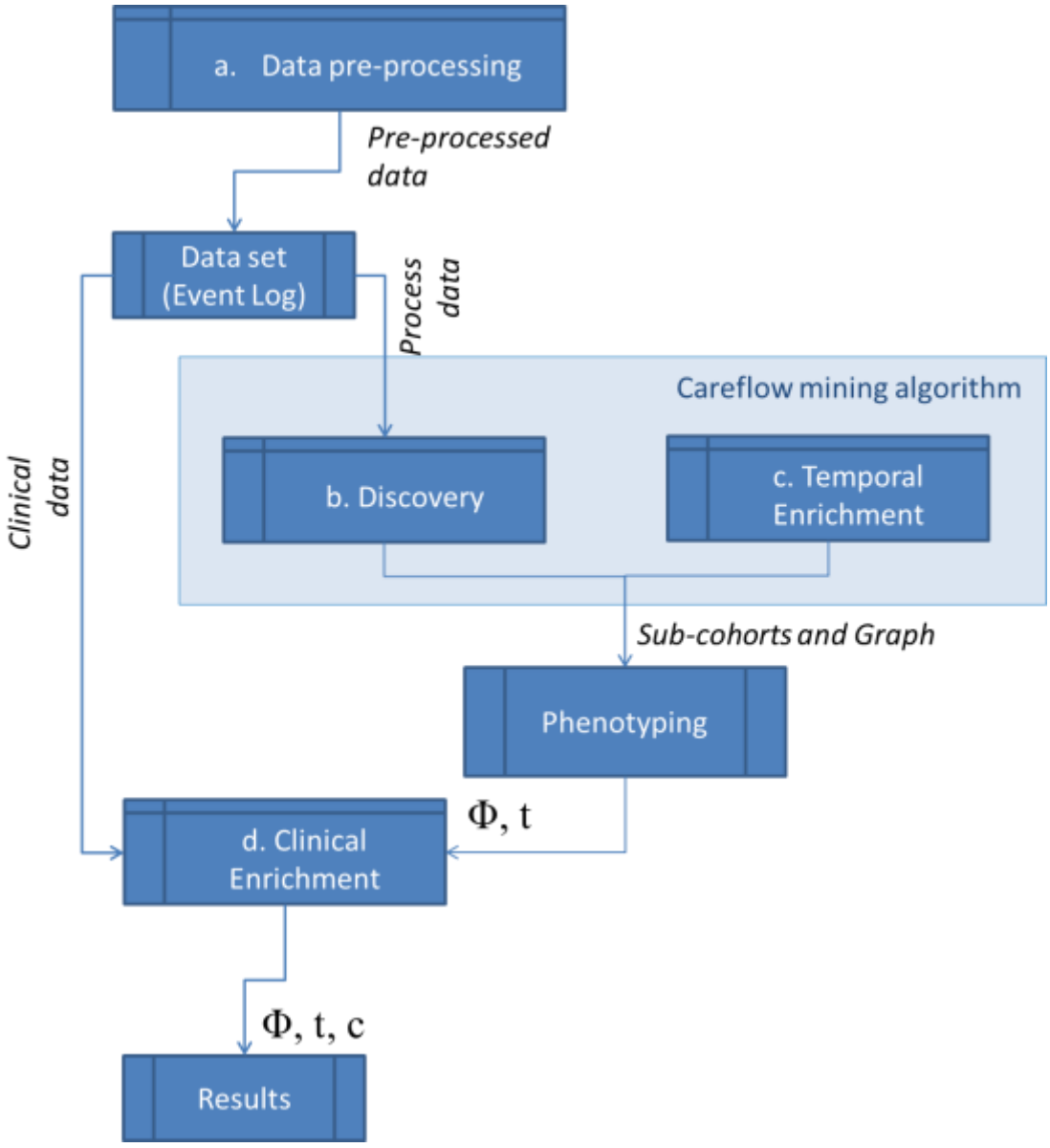


Figure 17 The analytic pipeline. Φ indicates the set of extracted phenotypes, Φ_t the phenotypes enriched with temporal information and Φ_{tc} the phenotypes enriched with temporal and clinical information

The careflow mining algorithm identifies different groups (sub-cohorts) of patients based on the careflows they experienced. Thus, a sub-cohort of patients is composed of those patients who experienced a specific careflow. The extracted careflows are represented using a directed acyclic graph (DAG). In this DAG, rectangles represent events, and arcs represent the temporal relation between them. Figure 18 illustrates an example of results obtained from applying the careflow mining algorithm. The name and duration of the event is specified in a rectangle, together with the number of patients experiencing that event. The algorithm implements the possibility to define and represent a higher level of specification of events as event type. The color of the rectangles graphically displays the information related to this level of events' representation. Numbers on the arcs represent the number of patients who go through that flow and the median duration of the flow, calculated over those patients. Patients who exit the flow after a specific event are represented as leaf nodes, labeled with "End". Using this representation, we can identify individual careflows that represent individual sub-cohorts. For example, in Figure 18, five sub-cohorts are identified,

respectively representing the event sequences <A, End>, <A, B, End>, <A, B, A, End>, <A, C, A, End>, and <A, C, B, End>.

The most important features of the algorithm results are: the explicit characterization of the events duration and of the time elapsed between consecutive events, and the identification of patient sub-cohorts.

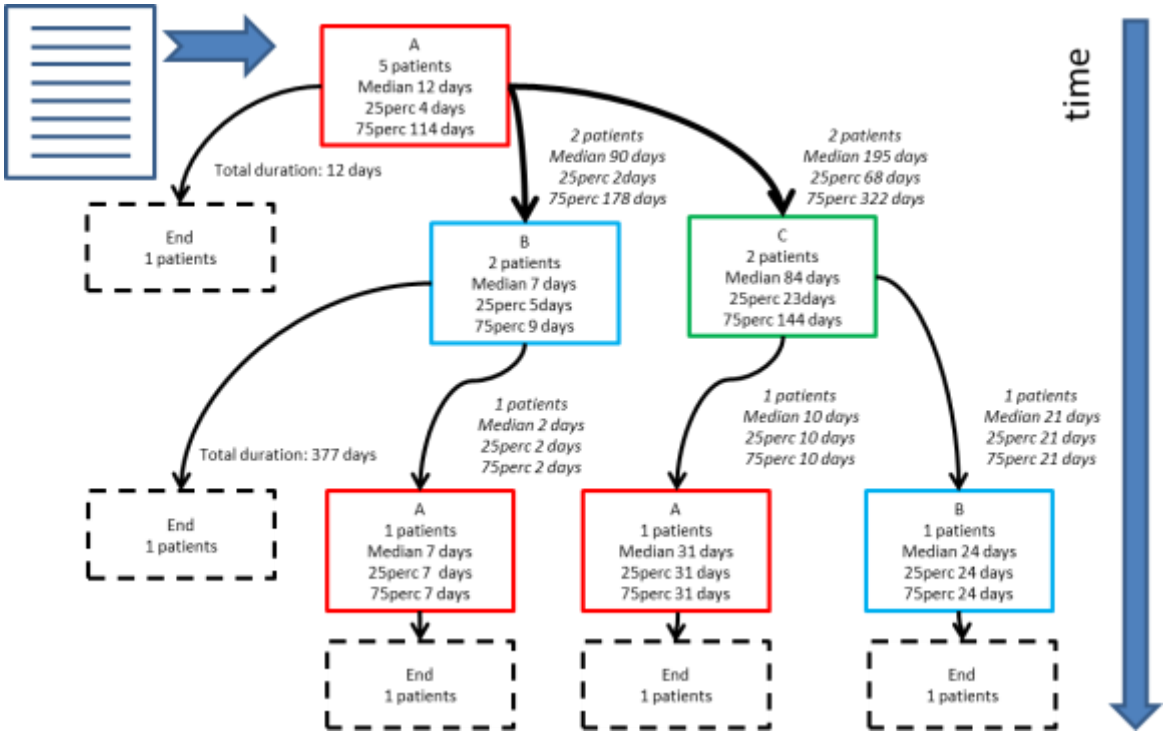


Figure 18 Careflow mining algorithm results, represented as a temporal directed acyclic graph

As illustrated in Figure 17, data analysis is accomplished in four phases: data preprocessing, discovery, temporal enrichment, and clinical enrichment.

In the following sections these phases are described in detail.

5.1.2.2 Preprocessing of events.

The input of the algorithm is an event log file (W. M. P. van der Aalst 2011), where each row is referred to a single event and each column includes patient's ID, t_{start}, t_{end} and the event label. The activities performed in the previous steps (Aim 1) to gather, integrate and store data in a common model, allow to easily extracting, from the i2b2 observation fact table data already structured in a format suitable for the analysis. Although, few preprocessing steps might be necessary, especially considering the dependencies among events for the specific analysis.

As already done in other mining approaches (van der Aalst et al. 2015), it has been assumed that events cannot influence each other directly, considering them as independent. The careflow mining algorithm considers all the events of a history as they were different. For this reason, when creating the event log, it is crucial to define how to deal with repeated consecutive events (Leemans & van der Aalst 2015). The choice of keeping certain events separate or merge them primarily depends on the analysis context. In the pre-processing phase, the algorithm gives the possibility to merge events, and to choose which events to merge, as an option.

For example, where there exists a history H = <A, A, B>, the user may consider to pre-process the two consecutive occurrences of the event A to create the history H' = <A*, B>. Repeated events <A, A> are aggregated into a single event A* that has the same t_{start} as the first occurrence of the event A in H, and the same t_{end} as the second occurrence of A in H. When no pre-processing is performed, the algorithm processes the first two events as if they were different events.

The result of the algorithm depends on the level of detail used to describe real clinical activities. Therefore, the aim of the analysis has to be clearly defined in advance, because it affects the decisions about data preprocessing and event logs creation. Thus, decisions about events selection and their aggregation should follow a careful assessment of the opinions and needs of clinicians and of the scope of the analysis itself.

5.1.2.3 Discovery, the Mining Algorithm.

The algorithm we developed to extract frequent careflows from process data is inspired by sequential pattern mining techniques (Agrawal & Srikant 1995; Zaki 2001). To assess the frequency of a particular sequence of events, our algorithm relies on the notion of *support*. Following the definition given by (Agrawal & Srikant 1995; Garofalakis & Rastogi 1999), the support of a sequence S is a proportion defined as the number of patients (N_S) who experience a specific sequence of events divided by the total number of patients in the analyzed population (N):

$$\boldsymbol{Support\ (S) = \frac{N_s}{N}}$$

Frequent sequences are those that have a support greater or equal than a user-defined threshold, θ. Thresholds are used to guide the search process such that only the most frequent patterns are extracted. As it happens for many support-driven search strategies in sequential pattern mining (Agrawal & Srikant 1995), the selection of the threshold on support has an impact on the overall results in terms of the number of patterns that are generated.

The algorithm starts to mining careflows from the most frequent observed event at the beginning of the considered sequence, where time is equal to zero. For this reason, the first event of the sequence is the one shown at the top of the produced graph start to illustrate the pathway originating from the first observed event (e.g. event A in Figure 19). This choice is based on the need to observe the complete evolution of a clinical phenomenon in time, such as the disease evolution from diagnosis.

It is possible to stop the search process through the *maximum length parameter*, which is a constraint on the maximum number of events included in the careflow. Figure 19 shows how, depending on the selected maximum length, it is possible to identify careflows with different lengths.

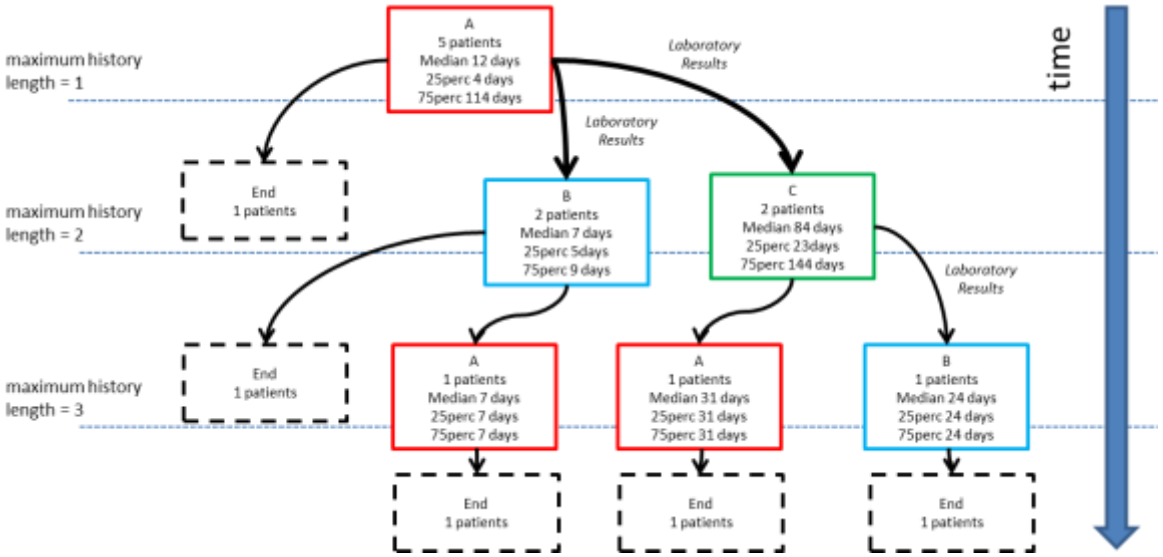


Figure 19 The exploitation of the maximum length parameter to define different careflows and identify different number of patients' sub-cohorts. If the maximum length parameter is set to 2, it is possible to identify 3 sub-cohorts undergoing the sequences <A, End>, <A, B > and <A, C>.

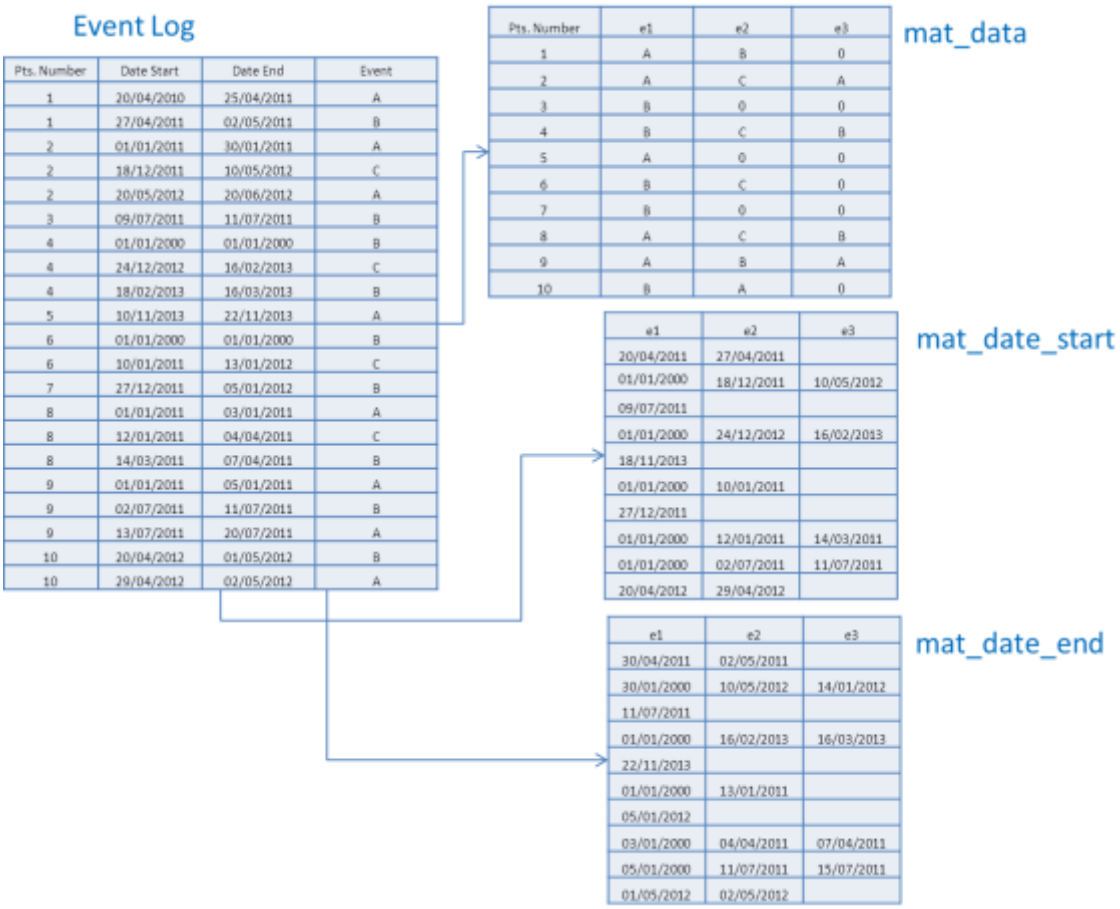


Figure 20 Event log and data matrix. The algorithm performs a transformation from long to wide format in the pre-process phase, transposing each observation of patients. mat_data contains the order events for each patient, mat_data_start and mat_data_end the start and end dates of events, in the same order of mat_data.

As first step, the algorithm transforms the event log in a data matrix, as shown in Figure 20. The algorithm starts by considering a set made up of all the starting events of the available clinical histories (column e1 in mat_data), and computing the support of each of these events; only the events that exceed the support threshold are selected. The algorithm adds steps to the careflows by iterating the support computation on the events following the initial set of selected events, until no more frequent histories can be extracted. The selection of the observation window, and thus of the starting point for each analyzed sequence of events, depends on the analysis context. Examples could be: the diagnosis of a disease to study disease evolution, first visit at a specific health center to analyze the local workflows.

The algorithm pseudo code is illustrated in Figure 21. Additional material can be find in Appendix B, where all the steps of the algorithm are graphically represented using the same example shown in this section. As output, the algorithm produces a file containing a direct graph representation that can be visualized through an editor such as GraphviZ [http://www.graphviz.org/content/dot-language], or using the process mining suite ProM [http://www.promtools.org/doku.php].

5.1.2.4 Temporal Enrichment

In the medical context, analyzing the time between consecutive events (e.g. follow-ups or specific procedures like dialysis) can provide useful information for characterizing sub-cohorts of patients. For this reason, while mining frequent histories, the algorithm also enriches the results by calculating temporal information related to nodes (event duration) and arcs (time between consecutive events). Given that several patients take part in the same careflow, the temporal information is calculated by computing the median, the 25th percentile and the 75th percentile of the following distributions:

- Duration of each event: for each patient, the duration of a generic event A is calculated as $t_{end}(A) - t_{start}(A)$
- Time between consecutive events: for each patient, the time between two generic events A and B occurring in sequence in a temporal history is calculated as $t_{start}(B)-t_{end}(A)$
- Total careflow duration: for each patient the algorithm calculates $t_{end}(E_{Last})-t_{start}(E_{First})$, where E_{First} is the first event in the history and E_{Last} is the last one.

The results of the careflow extracted from the event log shown in Figure 18, and enriched with temporal information are the ones shown in Figure 17 and already used to present the output of the algorithm.

```
PSEUDO CODE (See AdditionalMaterial_A for an example)

DEFINE
E = {e1,e2…ek} All the possible events
n: the number of sequences (e.g. the number of patients)
l: the length of a single sequence
m: max (l1,l2,…,li,…,ln)
th: threshold on support
mat_data: is a n*m matrix. The i-th row of mat_data contains the ordered events of patient i.
MAT(:,1), where MAT is a generic matrix, indicates the first column of MAT
MAT(1,:), where MAT is a generic matrix, indicates the first row of MAT


PRE-PROCESS
From the Event log build
Data Matrix (mat_data)                    % These matrixes have dimension n*m.
Matrix of start dates (mat_date_start)    % If the length of a single patient's history is lower than m,
Matrix of end dates (mat_date_end)        %the rest of the matrix row is filled with 0.


CALL the FUNCTION find_history(th, mat_data, mat_date_start, mat_date_end)
FUNCTION find_history(th,mat_data, mat_date_start,mat_date_end)
Find FE ⊂ E : set of the distinct first events in the histories
FE = distinct(mat_data(:,1))
size = nrows (mat_data)

for each e∈ FE
        sup(e) = number of occurrences e in mat_data(:,1) / size % compute the support of each first event
end

Find the FFE ⊂ FE :set of frequent first events. An event e ∈ FFE iif e ∈ FE AND sup(e) > th
If FFE = {∅} which means the FFE is empty
        Return to the invoking function;
Else
For each e ∈ FFE
        row_ind =  index of the rows of mat_data starting with e
        mat_data_new == mat_data (row_ind, :)      % subset mat_data and create mat_data_new
        mat_date_start_new == mat_date_start (row_ind, :)
        mat_date_end_new == mat_date_end (row_ind, :)

        Compute statistics on event e duration considering mat_date_start_new (:,1)
        and mat_date_end_new (:,1) - median, 25th and 75th percentile;
        Delete the first column of mat_data_new, mat_date_start_new, mat_date_end_new;
        CALL the FUNCTION find_history(th, mat_data_new, mat_date_start_new, mat_date_end_new)
End
```

Figure 21 The careflow mining algorithm pseudo-code

5.1.2.5 Clinical Enrichment

The careflows mined from process information in the discovery phase do not explain if different patient sub-cohorts are characterized by a change in clinical values, complexity, or treatments. To this end, it is important to enrich the results of the algorithm also with quantitative time series. This step is performed after the discovery phase.

Clinical enrichment is fundamental in evaluating the relevance of the mined phenotypes. This procedure allows comparing sub-cohorts, extracted from process data, in terms of clinical values. Through the clinical enrichment is possible to test the hypothesis that the sub-cohorts identified

by the algorithm present differences in terms of patients' health status. Moreover, thanks to this step of the analysis, clinicians have a broader vision of careflows, which allows them to assess if sub-cohorts are well characterized.

Clinical values can be compared with two different approaches, illustrated in Figure 22:

- Horizontal enrichment: specific clinical values are evaluated at fixed points in the careflow (e.g. the first measure after the first visit near the hospital or the last measure before the last follow up). In the example of Figure 22.A, the enrichment is performed by considering the last measurement before the second event of the careflow.
- Vertical enrichment: In this approach, the entire clinical variable time series as measured between two events is used. Such time series can be analyzed either as raw data, or processed using temporal data mining methods, such as temporal abstractions (Shahar 1997; Stacey & McGregor 2007; Post & Harrison 2007a). In Figure 22.B, for example, is possible to analyze all the clinical measurements from the end of the first and the beginning of the second event, as shown by the dots on the arrows. Time series can be synthesized using qualitative labels, such as "increasing", "stable", "decreasing", extracted using temporal abstractions.

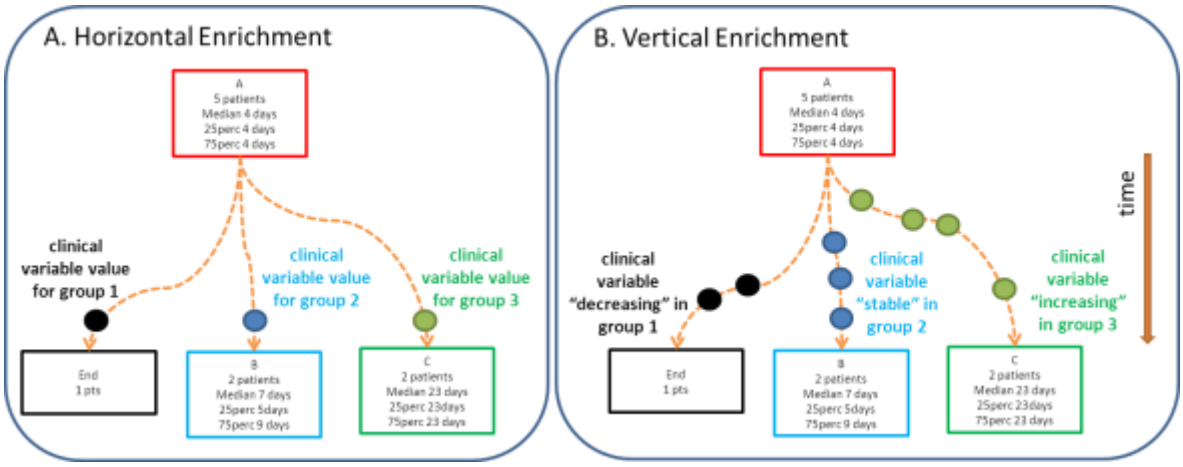


Figure 22 A schematic example of clinical enrichment of careflows

5.1.2.6 Recognition of AND events

The CFM algorithm results are acyclic graphs where events dependencies are expressed through arcs pointing from an event to the following one, indicating their temporal connection. In careflow mining, arcs that outgo from an event to more than one events indicate parallel executions of activities of all the successor once the predecessor is completed and are called split, incoming arcs from more than one predecessor to a single successor are called join (Panzarasa et al. 2002). From the control-flow perspective different modeling languages (like Petri Nets or Workflows Net including YAWL, BPMN, EPCs or Causal Nets) have well defined split and join semantics (W. M. P. van der Aalst 2011). Parallel activities, indicating two or more events that occur in the same time span in any order, are represented with AND split and join.

In business ***process modeling*** there are several advantages in exploiting parallel routing, as it can overcome classic physical limitations of sequential activities executed manually, like associated document that can only be in one place at a time. Within ***process discovery*** one of the first algorithm to be developed is the alpha algorithm (the heuristic miner is derived from it), which

scan event logs for specific patterns and represent them as petri net (Van Der Aalst et al. 2004). The algorithm captures ordering relations between events through the so called footprint of the event logs, which is a transition matrix representing relations for each pair of events. The alpha algorithm detects parallel events if the sequence of. The footprint matrix allows to represents parallel relations in events if sequences like <B, C> and <C, B> are detected. This is possible because in the footprint matrix each event is considered only once and, as typical in process mining, it does not consider the moment of occurrence of events within real histories.

One of the main differences of the developed ***CFM algorithm*** is that it produces a DAG, where events are shown as they occur in sequences during time. To represent parallel events, the same approach exploited with alpha algorithm is not applicable. Nevertheless, it was possible to apply a strategy for their detection inspired to the same relations among events. The applied strategy searches the mined pathways for sequences of events in the form <B, C> and <C, B>, but applies a further restriction on the moment when these sequences occur. Parallel events are detected in different paths of the history while satisfying the parallelism condition and occurring exactly at the same point of the history, in this case after the first event of the history (event A) and the fourth one (event D). When the condition is detected the two sequences <B, C> and <C, B> are merged and represented together as B AND C. The number of patients undergoing the events merged into the AND, and in the following event, are summed.

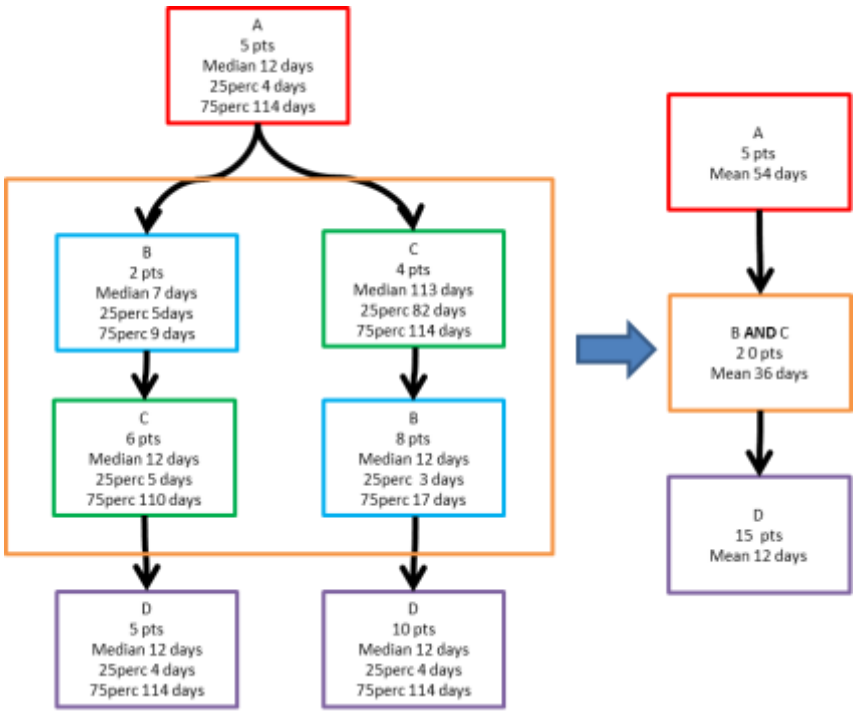


Figure 23 AND events recognition and representation in the CFM algorithm

Some syntactical differences from classical process mining approaches in depicting careflow raised further issues, in particular while dealing with temporal enrichment of events and arcs. The CFM algorithm enriches nodes and arcs with temporal information as median. While merging different events it was chosen not to use the median as measure of heterogeneous events duration: finding the appropriate median from the merging two events or arcs should require the computation of the measure from the original data set. For computational simplicity, it was used the mean instead, computing the mean of the mean of the original events. Moreover, it is clear that once the AND

detection is performed a possible loss of information can occur, especially in terms of temporal and clinical characteristics of the cohorts identified by the careflows. Several arcs (which might contain some information useful to physicians) are not visible anymore and, form the electronic phenotyping point of view, the number of sub cohorts is reduced. In the example of Figure 23 two phenotypes (A → B → C → D, A → B → C → D) are merged into a single one (A → B AND C → D). For these reasons, and also considering the scope of the CFM exploitation for electronic phenotyping in the CDS context, the detection of parallel events and their merging in AND was implemented as an optional function of the algorithm.

5.1.3 Exploitation for Electronic Phenotyping

Electronic phenotyping allows a shifting from identifying cohort of patients through static data base query to the examination of heterogeneous and unstructured data for searching the right features that might describe a clinical phenotype (Frey et al. 2014; Perer et al. 2015).

The CFM algorithm has the capability of searching for the most frequent histories in a cohort and representing them. Mining patients' histories means to illustrate the evolution of the disease through specific events and also to relate patients that follow similar trajectories. The possibility to use events temporal alignment to recognize similar profiles of patients add a temporal dimension to electronic phenotype. The following step was to enhance the phenotypes with relevant clinical descriptions, in order to prove that each mined history identified a well characterized cluster of patients. The clinical and temporal enrichment of the transitions among events is aimed to make histories more informative form the clinical point of view but also to compare the sub cohorts, and understand if a patient belongs to one of the identified group because a specific condition or a followed process of care. Clinical enrichment matches the similarity of the medical condition underlying the sequence of events across time. As also discussed in (Frey et al. 2014), retrieving from process data the knowledge of the path that a subject underwent allows precise descriptions of patients conditions and treatments. A further step is to describe the similarity among the detected paths to understand if they are a reliable driver for phenotypes.

Histories Similarity. The developed CFM algorithm mines the most frequent histories in the way they occur in reality. Each history represents the sub cohort of subjects that follow, not a similar, but exactly the same sequence of events during a precise period in the disease history. Differently from other approaches where the temporal similarity is exploited to mine clinical workflows (Combi et al. 2009), this approach represents the exact timing of evens' occurrence and identifies sub cohorts on the basis of frequent temporal patterns. The computation of the similarity measure among the mined histories allows to compare the temporal phenotypes detected through the algorithm, or to associate a new sequence of events to the previously detected careflows, on the basis of a score of similarity that takes into account only the executed procedures and not their temporal constraints.

The ***Jaccard similarity*** coefficient (Jaccard 1901) is an index used to compare the similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. The result is a value between 0 and 1 representing a degree of similarity, where 0 means that the sets are completely dissimilar and 1 that are identical. One of the most common applications of the Jaccard index is in bioinformatics to compare genes and metabolic

pathways (Zhou et al. 2016; Wang et al. 2015; Wang et al. 2013; Kaneko et al. 2013; Lewis et al. 2011). In this application, once the frequent histories are mined through the CFM algorithm, the Jaccard similarity coefficient is computed for each pair of sequences of events that build the histories associated to a detected phenotype. The measure of the similarity is done in terms of the information about events derived from process data. For each pair of mined histories H_i and H_j, J_{ij} is computed as

$$J_{ij}(H_i, H_j) = \frac{|\, H_i \cap H_j \,|}{|\, H_i \cup H_j \,|}$$

The result is a symmetric matrix of the dimension NxN, where N is the number of the mined frequent histories. Intuitively, the similarity matrix allows understanding if the disease evolution of a group of patients, which underwent to a certain history H_i, will be the similar or not to another group of patients who underwent H_j.

The CFM algorithm allows discovering phenotypes within which clinical procedures have similar evolving patterns. The disease progression in time can be better comprehended through the identification of the most meaningful paths in the DAG, which are the basis of phenotypes. The exploitation of the CFM results to convey phenotypes can be used for patients' stratification. The identified sub cohorts can be compared, assuming that ***similar careflows*** intuitively have comparable responses to treatments and are likely to encompass the same type of complications. Clinical decisions can be therefore tailored on the basis of the medical procedures expressed in the mined histories, enhancing precision medicine.

Beside the already discussed capability of leveraging on longitudinal instead of static data, the use of the CFM algorithm to automatically retrieve phenotypes from careflows detected in a specific clinical setting allows to dynamically reassess risk profiles and to characterize the complexity of chronic patients. The application of the CFM to perform electronic phenotyping on the collected data set has to respond to the very first question of the research: How to perform meaningful analytics to derive the right knowledge for novel insights on the disease? This question has been articulated into specific issues and choices on which data consider and how to process information.

i) What characterizes the complexity of T2DM chronic patients?

T2DM disease careflows can last for long time (within Mosaic data have been retrospectively collected for 10 years), the disease progression pace is slow and characterized by frequent modifications (e.g changes in therapies, changes in the frequency of follow ups, changes in health care providers) for each subject as the disease evolve. Sequences of events are very heterogeneous and affected by multiple conditions. Moreover, events can be related not only to the diabetic disease, but to any other event that can happen in a 10 years to a subject (e.g. national screening procedures, seasonal flu or traumatic events).

The CFM can be applied to the process data collected since the diagnosis of the disease. Instead of using a collection of defined ***ICD9-CM codes***, phenotypes can be extracted from all the procedures (ambulatory visit, follow ups and hospitalizations) executed during the disease evolution, as long as accurately preprocessed. In the following section the algorithm has been tested on the data of T2DM patients collected within the MOSAIC project to retrieve their careflow in term of process data derived from ***ambulatory procedures and hospitalizations*** before and after

the first visit near the hospital facilities. This first set of analysis was meant to prove the efficacy of the algorithm when applied in a real setting.

ii) Which are the latent characteristics of these phenotypes?

As automatically retrieved from retrospective collected data, data driven phenotype has to be proved to have clinical relevance. Clinical time series can be used to enrich the phenotypes so to detect if transitions among events are due to any meaningful medical episode (e.g. changes in metabolic control) or if different careflow can be associated to different clinical responses and lead to changes in risk profiles (e.g. increasing value in the CVR score). Instead of static thresholds on ***laboratory tests*** values, qualitative representations of ***clinical time series*** (Hba1c and CVR time series) are used to characterize phenotypes.

The third domain used to define T2DM phenotypes is ***medication data***. Within the developed system the information retrieved from drug purchase of patients have been used to characterize subjects' ***behaviors*** in time. A detailed description of the approach is in the *Mining Drug Exposure* Patterns paragraph.

The choice of mining careflows through ***process data*** has its first, functional, reason in the fact that such data are already well-structured to be represented as event logs, which enable their exploitation for workflow modeling while avoiding costly procedures of preprocessing the entire corpus of clinical data. This practical reason involves further advantages of the adopted two steps approach. Beyond facilitating the setup of the discovery step, process data trigger the phenotype definition through the segmentation of the entire population and the search space reduction. The patterns mined in the discovery step are the substrate for the additional inferences performed in the second step, where appropriate ***clinical data*** are used to enrich and compare specific transitions. This distinction and different use of process and clinical data is also necessary to prove the clinical relevance of the automatically extracted phenotypes.

5.1.4 Preliminary Results and Evaluation of the CFM Algorithm

In this section the potential of the developed methods for ***exploratory analysis*** and ***hypothesis generation*** on temporal phenotypes are evaluated. The goal is to leverage on process data, usually not available to physicians and decision makers, to identify sub-cohorts in the basis of the CFM algorithm and to detect potential interesting temporal patterns in the clinical histories of the patients. The relevance of the findings is evaluated from the clinical and disease management point of view, mainly through descriptive analysis. The goal is to evaluate if it is possible to in depth study the characteristics of the identified sub-cohorts or specific patients, to detect potentials treatments issues, to compare them with other patients' groups and check guide line compliance. Some of the limitations of the approach, due both to some intrinsic features of the algorithm and the data analyzed, are illustrated.

To accomplish this task, the process data, gathered from the Pavia Local Health Care Agency (ASL), were taken into account. In particular, on the basis of clinicians' interests, the algorithm was applied to the events reflecting patients' encounters within other health care facilities, or GPs follow-ups, since the diagnosis of T2DM before the first visit near the Pavia FSM hospital. Due to

the ASL data availability, only patients diagnosed date after 2004 were included in the analysis, obtaining a cohort of 424 subjects. The observation time window from the diagnosis to the first visit near the hospital was, in median of 5.2 months, its distribution is shown in the histogram in Figure 24

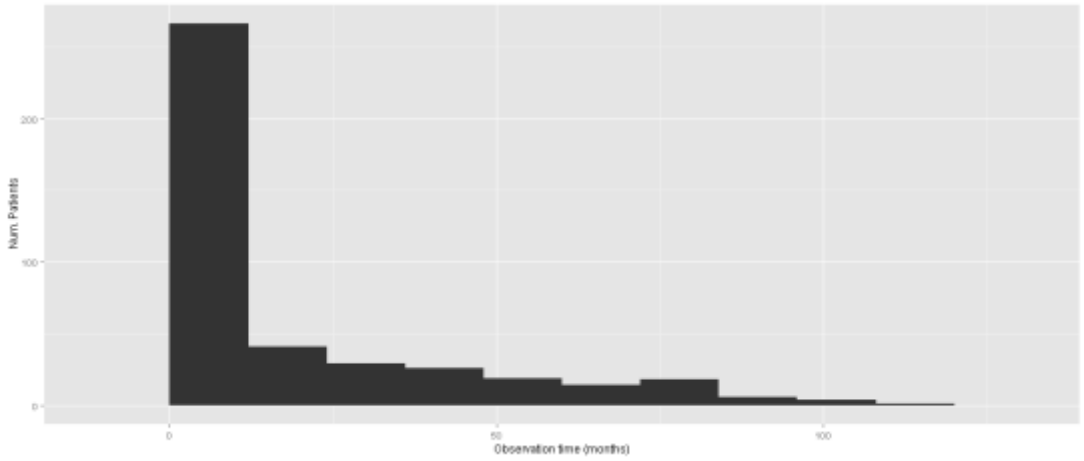


Figure 24 Observation time distribution

5.1.4.1 Preprocessing and Mining

Events were defined as ambulatory procedures, inpatient hospitalizations short procedure unit (SPU) visits and drug purchases, with the aim of including the most inclusive set of exposure data patients' underwent. To create the final event set, each data streams were pre-processed:

- ***Ambulatory Procedures*** were grouped on the basis of the taxonomy used near the ASL that indicates the ward where procedures have been performed. These kind of events had been preprocessed in order to aggregate consecutive identical events. This means that, for example, if a subject underwent more than one laboratory exam, without any other events in between, this is represented as a unique event with start date equal to the first exam and end data equal to the last one;
- ***Hospitalization*** and SPU Procedures ICD9-CM codes contained in the discharge summary were mapped into the first levels of the Clinical Classifications Software (CCS) for ICD-9-CM (Elixhauser et al. 2014);
- ***Drug Purchases***, already classified on the basis of ATC levels were filtered in order to take into account only treatments for diabetes and further aggregated on the basis of clinicians' suggestions (Metformin with Pioglitazone, Sulfonamides with Repaglinide). Consecutive purchases of the same drugs have been aggregated in unique events, like ambulatory services;
- ***Diagnosis*** and ***First Visit*** near the FSM hospital, which are used as censoring events, were extracted from the hospital EHR.

The choices of which data include in the event logs show how administrative data can be used as a proxy for pathological conditions. In the diagnosis process the overlap of billing and clinical information is clear: when clinicians apply a diagnosis code to a patient they are applying a

procedure, but this action also constitutes an event as it changes the patient state and it is a meaningful event in the disease evolution from the clinical point of view.

For each of the considered events, Table 6 shows the type, the name and the frequency of the event in the whole dataset.

TYPE	EVENT	Events Count	% on the event type	% on the total of event
Ambulatory Procedures	Laboratory Exam	1511	32.68%	28.58%
	Radiology	534	11.55%	10.10%
	Other Visit	499	10.79%	9.44%
	Cardiology	294	6.36%	5.56%
	Ophthalmology	239	5.17%	4.52%
	Rehabilitation	162	3.50%	3.06%
	ER	160	3.46%	3.03%
	Screening Exam	136	2.94%	2.57%
	Orthotics	128	2.77%	2.42%
	Surgery	117	2.53%	2.21%
	Other Ambulatory	91	1.97%	1.72%
	Neurology	77	1.67%	1.46%
	Pneumology	75	1.62%	1.42%
	Oncology	73	1.58%	1.38%
	Gastroenterology	72	1.56%	1.36%
	Urology	59	1.28%	1.12%
	Nephrology	57	1.23%	1.08%
	Dermatology	51	1.10%	0.96%
	Otorhino	44	0.95%	0.83%
	Angiology	39	0.84%	0.74%
	Transfusion	37	0.80%	0.70%
	Nuclear Medicine	34	0.74%	0.64%
	Radiotherapy	29	0.63%	0.55%
	Psychiatry	28	0.61%	0.53%
	Infectious Disease	27	0.58%	0.51%
	Hemodialysis	22	0.48%	0.42%
	ICU	15	0.32%	0.28%
	Rheumatology	13	0.28%	0.25%
	TOTAL	**4623**	**100.00%**	**87.46%**
Hospitalization	Circulatory System	73	23.55%	1.38%
	Other Hospitalizations	47	15.16%	0.89%
	Neoplasms	40	12.90%	0.76%
	Musculoskeletal System	34	10.97%	0.64%
	Injury and poisoning	24	7.74%	0.45%

	Genitourinary System	18	5.81%	0.34%
	Respiratory System	18	5.81%	0.34%
	Digestive System	17	5.48%	0.32%
	Endocrine Metabolic	16	5.16%	0.30%
	Nervous System	16	5.16%	0.30%
	Mental Illness	4	1.29%	0.08%
	Skin Subcutaneous Tissue	2	0.65%	0.04%
	Infectiousa and parasitic diseases	1	0.32%	0.02%
	TOTAL	**310**	**100.00%**	**5.86%**
Drug Purchases	Metformin and Pioglitazone	138	30.80%	2.61%
	Sulfonamides and Repaglinide	63	14.06%	1.19%
	Incretine via OS	54	12.05%	1.02%
	Acarbose	49	10.94%	0.93%
	Combination	23	5.13%	0.44%
	Incretine INJECTION	13	2.90%	0.25%
	Glicosurici	3	0.67%	0.06%
	TOTAL	**343**	**100.00%**	**6.49%**
TOTAL		**5276**		**100.00%**

Table 6 – Events frequencies

In this case, the high variability in events, and the reduced number of patients, imposed the choice of a low support. The CFM algorithm was run after setting a support of 0.07 and without any restriction on maximum careflow length. The choice was made after several heuristic experiments, where also the meaningfulness and readability of the mined careflow was considered. The most frequent careflows mined by the algorithm are shown in Figure 25.

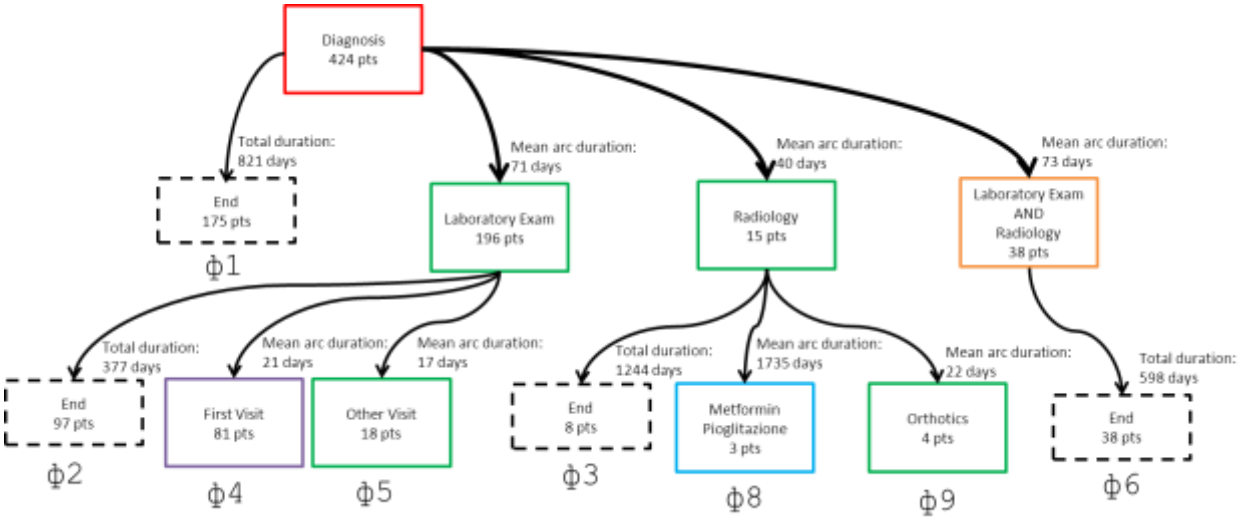


Figure 25 The mined careflow

On the basis if the previous figure is possible to identify nine phenotypes. Table 7shows each path of the careflow associated to such phenotypes, which are defined as the sub cohort of patients following one of the frequent clinical paths identified by the careflow mining algorithm. The table also includes the number of the patients associated to each path and the median time between the diagnosis and the first visit, which is the time span considered for this analysis.

Careflow number (Phenotype)	Careflow	Number of Patients (% on the total of 424 patients)	Time from Diagnosis to First Visit (Years)
Φ1	DIAGNOSIS → END	175 (41.27%)	2.25
Φ2	DIAGNOSIS → LABORATORY EXAM → END	97 (22.88%)	2.02
Φ3	DIAGNOSIS → RADIOLOGY → END	8 (1.89%)	3.41
Φ4	DIAGNOSIS → LABORATORY EXAM → FIRST VISIT	81 (19.10%)	0.56
Φ5	DIAGNOSIS → LABORATORY EXAM → OTHER VISIT	18 (4.25%)	2.93
Φ6	DIAGNOSIS → LABORATORY EXAM and RADIOLOGY → END	38 (8.96%)	1.64
Φ8	DIAGNOSIS → RADIOLOGY → METRFORMIN, PIOGLITAZIONE	3 (0.71%)	1.38
Φ9	DIAGNOSIS → RADIOLOGY → ORTHOTICS	4 (0.94%)	2.38
TOTAL	DIAGNOSIS → *	424 (100 %)	

Table 7 – Mined careflows

The main part of the patients are associated to Φ1 (41.27% of the total), these patients exit from the careflow after the first event, as they have too unfrequented events to be represented, according to the fixed support. The second more frequent, and more interesting, group of patients had one or more laboratory exams before the first visit near the hospital (Φ2) or a generic specialist visit into another clinical center (Φ5). These kinds of events ("Other visit") are classified as generic events and it was not possible to better define them, as they did not match any clinical event registered in the hospital EHR and the only available information derived from the administrative data source. Patients associated with Φ6 (8.96% of the total) follow either the careflow "DIAGNOSIS → LABORATORY EXAM → RADIOLOGY → END" or the careflow "DIAGNOSIS → RADIOLOGY → LABORATORY EXAM → END". In this case the algorithm merge the two careflows into the one "DIAGNOSIS → LABORATORY EXAM ***and*** RADIOLOGY → END" and the duration of the arc form the "DIAGNOSIS" to the event "LABORATORY EXAM ***and*** RADIOLOGY" is the mean value of the arcs to the first

"LABORATORY EXAM" and to the first "RADIOLOGY" procedures after the "DIAGNOSIS" event. Patients identified by the careflows Φ3, Φ8, Φ9 (3.53% of the total) experienced one or more radiology procedures after the diagnosis. Patients in Φ8 are the only ones for which the algorithm detect a treatment with anti-diabetics drugs. Although, due to the very small number of subject in Φ8 (3 patients), Φ9 (4 patients), it is difficult to retrieve any significant inferences on these patients.

Once the cohort is segmented through the algorithm, the ***Jaccard similarity*** can be exploited to compare the sub-cohorts. The Jaccard index is computed on the complete sequences of events that represent histories. This choice partially reduces the drawbacks derived from the imposed low support, due to the high variability of events. Figure 26 shows the 9 X 9 similarity matrix.

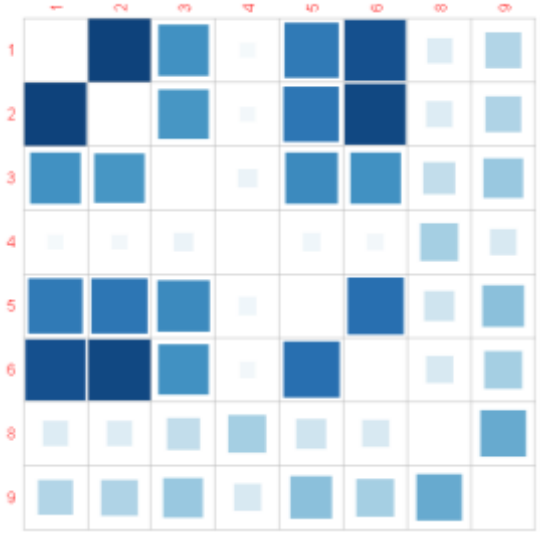


Figure 26 Careflows Similarity Matrix

The matrix shows that Φ1 has the highest similarity with Φ2, Φ5 and Φ6 (which careflows include the event LABORATORY EXAM) and also, although with lower value of the index, with Φ3 (which careflows include the event RADIOLOGY). As shown in Table 8, the frequency of LABORATORY EXAM (24.11% on the total of events) and RADIOLOGY (9.19% on the total of events) in Φ1 suggests that these patients are more monitored with laboratory test rather than with radiology procedures, like Φ2, Φ5 and Φ6. However Φ1 clinical histories are more similar to the ones of patients in Φ3, which experience radiology procedures, than to the ones of patients in Φ4, which do not undergo any radiology procedure.

Phenotype	*Φ 1*	*φ 2*	*φ 3*	*φ 4*	*φ 5*	*φ 6*
LABORATORY EXAM	***24.11%***	28.25%	17.83%	33.33%	27.76%	28.14%
RADIOLOGY	***8.19%***	9.10%	15.50%	0%	10.59%	14.14%

Table 8 – Frequency of Laboratory tests and radiology exams in the phenotypes

At this step of the analysis, it is already possible to point out some considerations and identify limitations that help to guide further analysis and, above all, to find and efficiently way to exploit the algorithm within the CDSS. For example, it is interesting to note how some frequent Ambulatory procedures, like visits near Cardiology units (accounting for the 5.56% of the total of

the events), are not included in the mined careflow. This happen also for Hospitalizations, which are important but infrequent (5.86% of the total events) events and are never shown in the careflow.

As already mentioned, T2DM patients experience a lot of different and heterogeneous procedures in long periods of time, often the registered clinical actions are not even related to the diabetic disease or its comorbidities (e.g. screening for other pathology or traumatic events). The assumptions for pruning the mined careflows close to the diagnosis is motivated by this high variability in addition to the necessity of choosing a support that could balances the clarity of the mined careflows and meaningfulness of the phenotypes. However, some clinical relevant events that occur later in patients' histories are not represented. To overcome these kinds of context related problems, several efforts were focused on better preprocessing events and in merging data into more informative, knowledge based, events. For example, while taking into account the general objective of the MOSAIC project on complications detection, to reduce the variability and obtain more consistent results the algorithm was applied to LOC (Level of Complexity) events, which have been described in 3.5.

5.1.4.2 Clinical Data Enrichment

The second step of the analysis framework is dedicated to enrich the history with clinical data, which can be considered as the outcome variables compared to the exposure ones used for mining the careflows.

This example also helps to illustrate the misalignment of process and clinical data in the available data set. In fact, clinical data are only available after the first visit near the FSM hospital, while process data are available from the diagnosis for the selected cohort. This data structure peculiarity leads to another important consideration. While developing the CFM algorithm, an alternative possible strategy was explored, where clinical time series were exploited to detect if transitions among events could lead to splits between paths in the careflows, thus explicitly taking into account the clinical information during the discovery process. Even though the potentials of an entirely data driven approach are interesting, establishing the clinical relevance of the obtained results would not be possible due the lack of such data, neither would be possible to apply this approach in the case clinical and process data were gathered in different observation windows.

To assess the informative value of the careflow mined from process data, the Glycated Hemoglobin (Hba1c) values were selected as the most meaningful biomarker from the perspective of diabetic patient care. Figure 27 shows the boxplot of the Hba1c values measured for two years after the first visit in each phenotype. When compared with Kruskal-Wallis chi-squared test, the Hba1c values results significantly different (p-value $<<$ 0.01) between different phenotypes. Excluding from the discussion $\Phi 8$ and $\Phi 9$, due to scarce number of patients involved, $\Phi 3$ shows the highest value of Hba1c in the period.

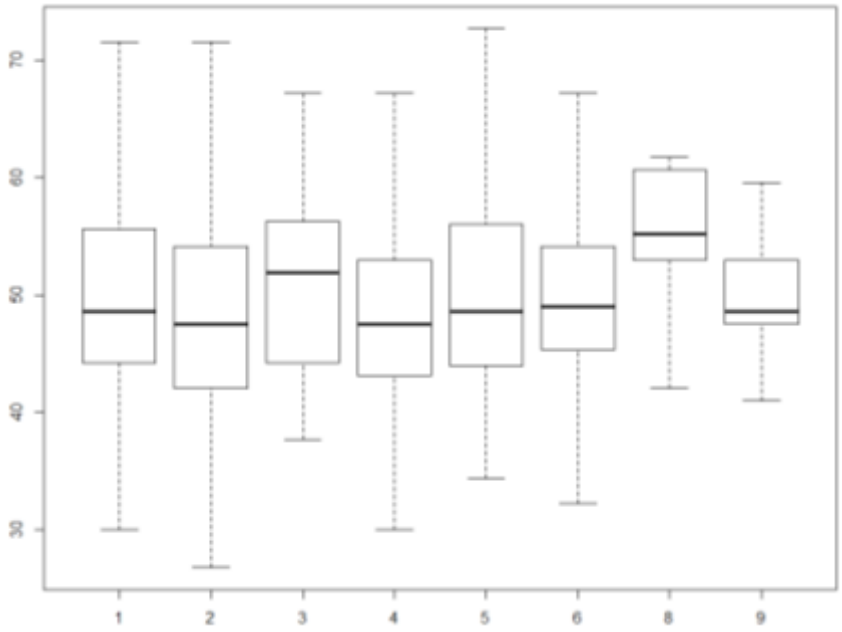


Figure 27 Hba1c Values in the two years after the first visit- Boxplot

To get a better insight into the clinical time series that characterize the phenotypes after the first visit, the so called vertical enrichment approach was exploited to compare patients' metabolic control in time. Only patients with at least 3 measures in the period were included in the analysis. Firstly, the Hba1c time series, measured in a two years' time period from the first visit, were tested with a the non-parametric Mann-Kendall test for monotonic trends (Hamed & Ramachandra Rao 1998), with time as independent variable X and Hba1c values as dependent variable Y. With the Mann-Kendall test "the null hypothesis of stability is rejected when the *tau* of Y versus X is significantly different from zero", in this case is possible to conclude that there is a trend in Hba1c over time. The *tau* range from 1 (in the case the agreement between Y and X is perfect, meaning an increasing of Hba1c in time) and -1 (in the case the disagreement between Y and X is perfect, meaning a decreasing of Hba1c in time). Results are shown in Figure 28. The results suggest that patients in Φ3 have the best (decreasing value of Hab1c) and patients in Φ6 have the worst (increasing value of Hab1c) metabolic control.

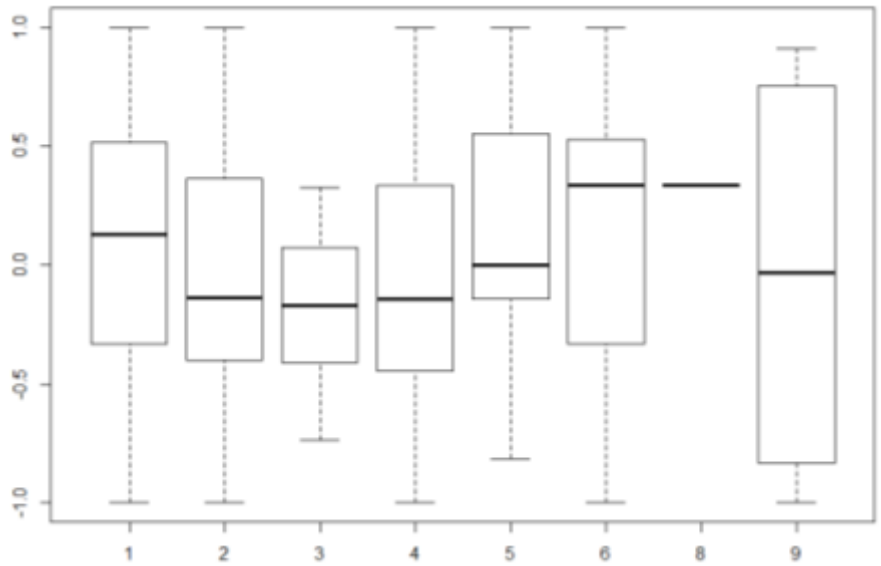


Figure 28 Mann - Kendell TAU values

In this case the most interesting phenotype to in depth investigate is Φ3. In fact when the Hba1c time series are further analyzed, and trend TAs are applied to detect the longest trend (with a slope equal to the 5% of the Hba1c mean value of each series in order to account for single patients variability), patients in Φ3, despite having the highest values of Hab1c, always have decreasing trends, as shown in Figure 29.

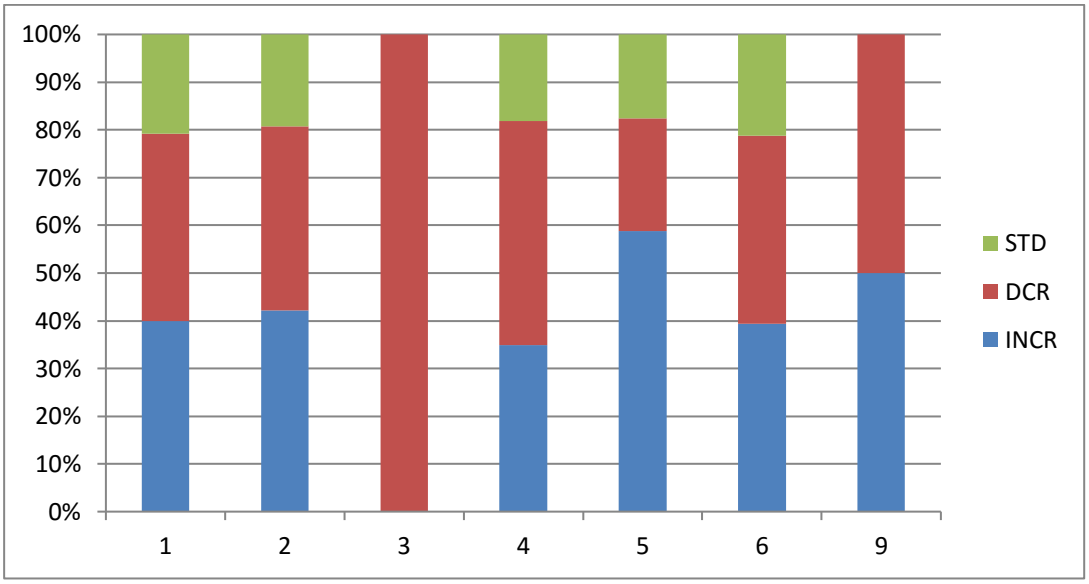


Figure 29 Hba1c Trends in the two years after the first visit

On the basis of these findings it is possible to assume that patients in Φ3 are subjects more complicated, and for this reason clinicians at the hospital spend more efforts in trying to stabilize their metabolic control. A first characteristic of these patients, for example, is that they have the longest period between the diagnosis end the first visit near the hospital (3.4 years, compared to 1.9 years in the rest of the cohort). Another trait of Φ3 patients that is possible to investigate is their ***treatments***, in particular if there was any evident change in their drug purchasing before and

after start to being treated near the FSM hospital. Table 9 shows the number of patients treated with a certain drug in the two considered time frames. Clearly Φ3 patients receive more complex and complete diabetic treatments once they are in charge of the hospital.

Treatments	*Patients treated BEFORE First Visit*	*Patients treated AFTER First Visit*
Lipid Lowering	2	6
Antihypertensive Beta Blocking	3	5
Antihypertensive Calcium Antagonists	3	3
Antihypertensive Other	2	2
Antihypertensive Renin Angiostin System antagonists	5	6
Antithrombotic	6	7
Diuretics	4	4
Diabetes		
Glp-1_Analogues		2
Insulin		1
Metformin	3	6
Metformin DPP-IV Inhibitors		1
Repaglinide		2
Sulfonamides	2	3
Alpha glucosidase inhibitors		2

Table 9 – Drugs treatments in Φ3

From the medical point of view, another interesting aspect to investigate is which kind of Lipid Lowering treatment was prescribed to these subjects. As this analysis, wants also to illustrate the potential of a drill down approach performed on heterogeneous set of clinical and process data, an incisive example is to show the zooming, from the entire considered cohort to a single patient, for example the one treated before the first visit with Simvastatin, Lovastatin and Rosuvastatin, and after the first visit only with Rosuvastatin. The hypothesis generations process in this case lacks a final step, clinicians wanted to have a more informative vision of the patient treatment, for example they want to know for how long he/she have been treated with each one of the listed Lipid Lowering or if there was any overlapping in the treatments. These aspects have been tackled and the methods applied to detect drug exposure patterns are illustrated in 5.2

The last aspects to include in the analysis, in line with the overall project objectives, is the study of the complications onset. In Figure 30 is shown the distribution of the ***complications*** that arise after the first visit in the different groups. In this case Φ8 and Φ9 are excluded from the analysis as only one patient in each group developed one complication. T2DM complications are grouped in Macrovascular (MAC), Microvascular (MIC) and Not Vascular (NV). Table 10 reports the number of patients, the proportion of patients in each group and the time from the first visit to the onset of the complication in each group for each type of complication.

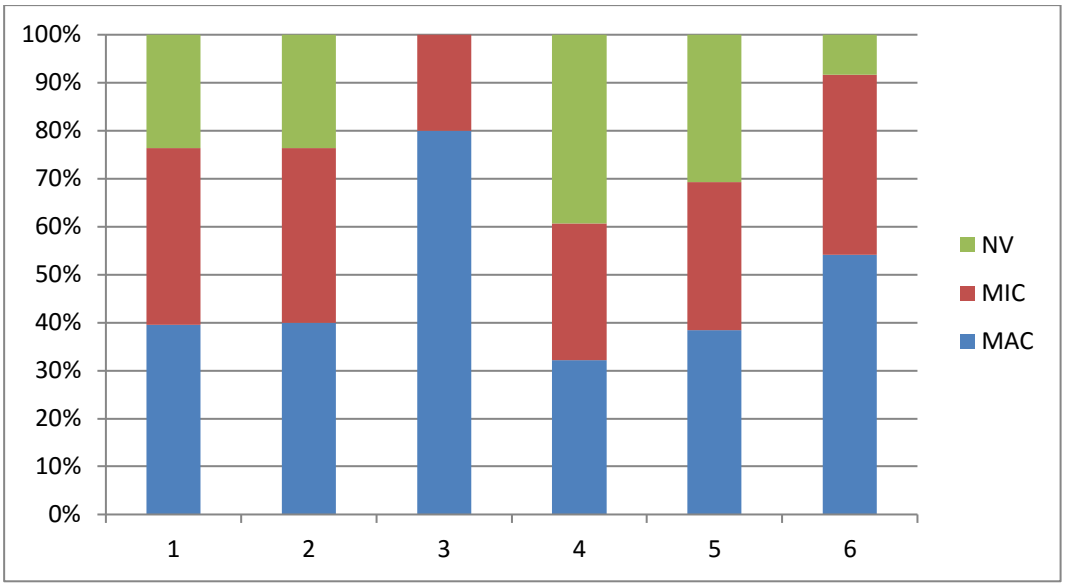


Figure 30 Complication type - After the First Visit

TYPE	Φ	Num. Pts with complication	% Pts. With Complication on the pts in the phenotype	Time (days) FV > COMPL
MAC	Φ1	45	25.71%	mean(758.93)sd 750.75
MAC	Φ2	22	22.68%	mean(725.77)sd 893.64
MAC	Φ3	4	50.00%	mean(891.25)sd 554.24
MAC	Φ4	18	22.22%	mean(666.56)sd 611.31
MAC	Φ5	5	27.78%	mean(781.4)sd 911.64
MAC	Φ6	13	34.21%	mean(654.92)sd 961.2
MIC	Φ1	42	24.00%	mean(825.05)sd 666.05
MIC	Φ2	20	20.62%	mean(543.35)sd 585.63
MIC	Φ3	1	12.50%	mean(404)sd NA
MIC	Φ4	16	19.75%	mean(487.5)sd 616.95
MIC	Φ5	4	22.22%	mean(647.75)sd 648.27
MIC	Φ6	9	23.68%	mean(524.44)sd 435.7
NV	Φ1	27	15.43%	mean(164.52)sd 345.7
NV	Φ2	13	13.40%	mean(416.85)sd 566.37
NV	Φ4	22	27.16%	mean(157.14)sd 331.5
NV	Φ5	4	22.22%	mean(774.75)sd 223.17
NV	Φ6	2	5.26%	mean(703.5)sd 415.07

Table 10 – Complications and onset time.

Also in this case, patients in Φ3 show an interesting pattern in the developing of complication. They experience more Macrovascular events than any other phenotype, while these events arise after a longer period (on average 891.25 days after the first visit, compared with the mean of the rest of the population that is 758 days). Although, for the purposes of hypothesis generation, Φ3

shows some interesting features, it is not possible to further investigate the group from a statistical point of view, due the very small number of patients (8 subjects). If Φ3 is grouped with the most similar phenotypes (Φ5, Φ6) on the basis of the Jaccard similarity, the Kaplan-Mayer survival curve for Macrovascular events after the first visit (Figure 31) suggests a faster worsening of the conditions of these patients. Although the computation of the total sample size needed to reach a Significance of 0.05 and a Power of 0.8 is of 1023 subjects, versus the 351 available in the data set.

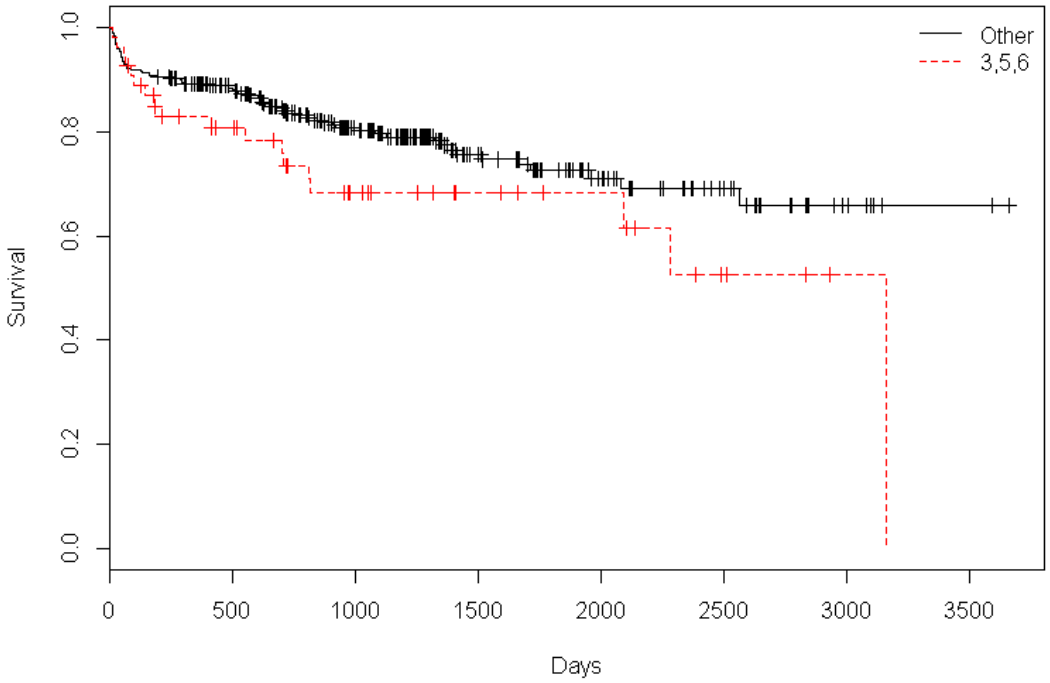


Figure 31 Kaplan Mayer survival curve – ф 3, ф 5 and ф 6 compared to the rest of the phenotypes.

5.1.5 Careflow mining for Electronic Phenotyping: Discussion

The previous paragraphs illustrate the development of an algorithm able to retrieve the most frequent careflows in a population and leverage on them to segment the cohort in temporal phenotypes. More specifically, once careflows are extracted, it is possible to rely on them to stratify the population by undirected dynamical phenotyping, described in (Albers et al. 2014). While the concept underlying this approach is similar, it has some distinguishing ***novelties***. The CFM algorithm takes into account the temporal nature of the data, explicitly including both process and clinical information. This is a novel feature, as most electronic phenotyping approaches have so far been focused on clinical data only (Hripcsak & Albers 2012; Hripcsak et al. 2015). While such data are quantitative in nature, and are thus more suitable to be analyzed using time series analysis methods, process data provide insight into the sequence of events that patients undergo during their care. The CFM algorithm differs both from sequential pattern mining and from process mining methods because of the central role we give to the specific timing of occurrence of the events in patient histories. As a matter of fact, the algorithms available in the literature usually consider frequent occurrences of sets of events within sequences or event logs, without taking into account the moment of occurrence of such events within the real temporal history of the subjects. This usually does not fit well the healthcare context, where the same type of event has a different

clinical meaning if it occurs at the beginning or the end of the treatment. As a consequence, we have considered as different the events of the same type (i.e. the same label) occurring at different time points of a clinical history, and we anchor the careflow mining algorithm to start from the first events of the sequences of all the patients.

From the ***knowledge discovery*** point of view, the developed mining algorithm can be exploited (i) to trace the careflows followed in any health center. This activity can be referred as process discovery (W. M. P. van der Aalst 2011) and it mainly depends by the data availability and the data model used to gather information. The algorithm findings can be used (ii) to compare the discovered careflows with the care processes derived from the guidelines adopted by the specific health center, supporting process conformance activities (W. M. P. van der Aalst 2011). Is clear that, while the transferability of the method itself can be addressed tackling some technical issues, like data format and available information, the knowledge extracted from event logs is strongly affected by a center effects. The potential generalizability of the results, that are the recognized phenotypes, has to take into account the center effect and a validation step is mandatory to assess the clinical relevance of the findings in another setting.

The proposed method has been applied in ***other clinical settings***, where highlighting the different clinical evolutions that may happen to similar patients represents a valuable element to extract patients' subgroups and to inform biomedical decision making, while assuming that similar careflows have comparable responses to treatments and are likely to encompass the same type of complications. The algorithm has been applied in in to obtain better insights on the health care processes of ***breast cancer*** patients (a manuscript is under final revision for publication on JBI) and in (Canavero et al. 2016) to detect the complaint use of anticoagulants treatment for the prevention of ***ischemic stroke*** in patients with atrial fibrillation. The obtained results offer a promising scenario for applications of the algorithm in order to involve the comparison of the extracted careflows with the existing clinical protocols followed at the hospital, to identify and study potential deviations, or to compare costs allocation for each phenotype.

The proposed approach has some ***limitations***. First, there are some clinical settings that are more suitable for its application than others. Since the algorithm considers all events as different, the methodology fits well to contexts where there is a relatively small number of events that can take place. For example, when the algorithm was applied in an oncology setting, this was particularly true, as the pattern of care follows specific standards and protocols that involve a limited number of actors and of possible type of events. In addition, the treatment of the disease is mainly carried out in the hospital setting, including outpatient services, and can thus be captured in its completeness by analyzing data included in the EHR. In T2DM, where there is a high variability of treatments due to the variety of comorbidities that are often involved and the non-hospital delivery of care, are less suitable for the application of the algorithm as it is. As suggested, this limitation can be addressed by carefully preprocessing the data, and properly selecting and aggregating similar events.

The discovery phase of the algorithm works by ordering the events on the basis of their starting time. This could represent a limitation in the case of events starting at the same time and having the same support. These events are currently handled by randomly selecting the first one to be shown in the careflow. Also in this case, the selection of the events to be included in the careflows could be a way to overcome this limitation, but this could be not possible in some situations. There

might also exist complex situations that imply one event frequently taking place during another one. In this case, the algorithm would represent the events in their starting order and the information related to the fact that the second event frequently ends before the first one would be visible as temporal enrichment on the nodes and edges. While this representation is anyway complete, its interpretation would require more experience by the user.

The understating of the algorithm limitations and potentials is essential to define the strategy for the algorithm implementation into the ***CDSS***. In particular, as described in 7.3.1,the approach was exploited to segment the cohort of patients treated near a specific center with a drill down approach able to guide health decision maker into a better understating of the characteristics of the patients, from demographic variables, to temporal phenotypes, to complications distributions.

5.2 Mining Drug Exposure Patterns

Data collected from administrative flows have the capability of enrich data from EHRs as they are collected outside the standard care facilities and they can represent not only billing and reimbursement procedures, but also several the activities that patients undergo or, or even actively perform, during their daily life. Interesting results were obtained as regards of the discovery of well-defined drug purchases patterns behaviors derived from administrative data streams.

Drug purchasing data encompass information never accessed before by clinicians which has the capability of showing actual behavior of their patients and enhance their monitoring. These data, once properly processed and combined with clinical time series derived from EHRs, provide new insight in the disease evolution of the patient. The discovery of temporal drug purchases patterns derived from administrative data streams supply hospital practitioners with better insight on the diabetes management and with several information about patients' behaviors otherwise not available to them.

Moreover, as described in the introduction, CDSS structured as dashboards can employ ***visual analytics*** techniques to present data in more informative way and are especially useful to represents temporal patterns. Moreover, dashboards have been already used in pharmacotherapy to enhance cognitive recognition of patterns in time and provide new insight on treatments.

Although the retrieved and shown patterns of drug purchases are not used to automatically detect specific phenotypes within the population, their use improves the characterization of patients. Drug purchases are collected as longitudinal data and the efforts for their display take into account the temporal dimension.

The analysis of drug purchases data and their exploitation into the CDSS has been organized in the following steps:

- Find and use ad hoc ***measures*** and methods to address exposure representation;
- Compute ***indicators*** that allow to compare the selected measure of a patient with the measures of the cohort of patients exposed to the same drugs;
- Detect temporal patterns and define dynamic ***tailored thresholds*** able to take into account patient and population variability;
- Exploit the results in the CDSS with proper ***visual analytics.***

5.2.1 A Measure of Drug Exposure.

The data coming from the Pavia ASL contains information about prescription-related drug purchases. Each drug purchase is described by its ***Defined Daily Dose (DDD)***, which allows computing the expected number of therapy days related to that purchase. As the information about the actual drug dosages a patient should take was not available in the current datasets, drug purchases were used as a proxy for estimating drug intake. Under this assumption, the DDD information represents the days' supply for a specific purchase and then DDDs has been used in the next steps of the analyses to evaluate patients' behavioral patterns related to drug treatment.

The activities performed in the data gathering phase allowed to collect drug purchases records on the basis of the analyses clinical significance. The first step was to classify the purchase records on the basis of their relevance to the diabetic disease and to reduce the variety in the data set to build simpler and meaningful scenarios based on treatment behaviors. To this end, the following drugs classes were selected:

- Drugs Used in Diabetes
- Antithrombotic Agents
- Antihypertensive
- Diuretics
- Lipid Modifying Agents

After this selection step, the number of distinct drug codes in the data stream was 87. With the collaboration of clinicians, the resulting codes where further grouped into 17 classes defined on the basis of medical knowledge and of the ***ATC-WHO system*** [http://www.whocc.no/atc_ddd_index/]. Drugs that are not specific for the treatment of diabetes have been considered at a higher level of the ATC taxonomy, whereas drugs used to treat Diabetes have been characterized at a deeper level.

As first attempt the exposure was measured relying on the CSA (Continuous, Single Interval Measure of Medication Acquisition) index (Steiner & Prochazka 1997), calculated as the ratio: sum of DDDs over an observation period / length of the observation period (discretized in semesters). Though some recent pharmaco-epidemiology studies (Cadarette & Burden 2010; de Vries et al. 2015; Nichols et al. 2015; Simon-Tuval et al. 2015) investigate the impact of drug adherence in the arising of T2DM acute events and in glycemic control. These works successfully exploit ***Proportion of Days Covered (PDC)*** measure to assess patient exposure and compliance to specific drug, in particular statins (Lin et al. 2015; de Vries et al. 2015) and metformin (Nichols et al. 2015). Moreover, within the EU funded project ABC (Vrijens et al. 2012), researchers proposed a new taxonomy, in which adherence to medications has been conceptualized on the basis of behavioral and pharmacological concepts. Consistently with this proposed framework, the analysis efforts were directed to take into consideration the "implementation of the dosing regimen", defined as the extent to which a patient's actual dosing corresponds to a prescribed dosing regimen. As mentioned, DDD values were used as proxy for the dosing regimen and PDC was exploited for the drug exposure analysis. PDC is calculated as the number of days with drug on hand divided by the number of days in the specified time interval. The PDC can be multiplied by 100 to yield to a percentage.

Leveraging on DDD measures, several preprocessing steps allowed building drug tailored exposures time series with 1 month granularity. As suggested by other works (Peterson et al. 2007), early refills of the same drug and dosages were considered additive (cumulative use), while switches between drugs or dosing regimens were considered as complete switches with no overlap granted. Figure 32 shows an example of how purchases raw data for each patient and for each ATC were preprocessed. In the example, the patient purchased a certain drug every two month and, within each purchase, the sum of the DDD for the active principle accounts for an exposure of 60 days (a). During the pre-processing phase we subtracted the DDD exceeding 30 in one month and

added them to the following one. In this way is possible to depict the actual exposure time, as defined by DDDs.

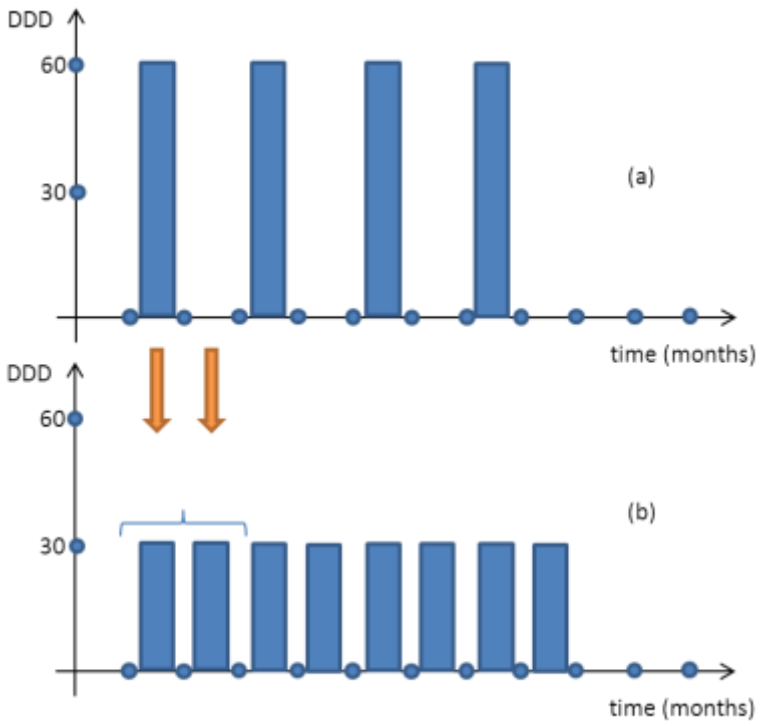
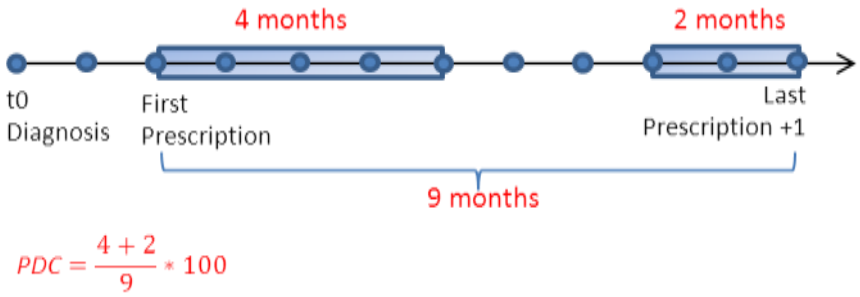


Figure 32 – DDD preprocessing

When in the entire observation period was detected only one purchase, the event was labelled as "single prescriptions" and it was not included these records in the PDC computation or in further analysis. In fact, in this case the PDC would be 100%, while is more appropriate to include those patients in a separate class. Once exposure time series had been built, it was possible to compute PDC to characterize exposure time intervals. Figure 33 shows graphically the computation of PDC through a simple example. In the example PDC is computed with a monthly time granularity. Choosing the right time granularity affects the way time series might be represented, making them more or less readable. However, as the PDC is an absolute value, it doesn't affect its final value. The example shows a patient that starts to purchase a drug two months after the diagnosis and continuously purchase it for 4 months. After a 3-month break, the patient purchases the drug for another two months. The exposure time is 6 months (Total covered month by the DDDs), while the denominator account for the period between the first prescription and the last one, also considering that, after the last purchase, there is an extra month to be added in the formula representing the time covered by the last purchase.

$$PDC = \frac{\text{Total Covered Months}}{(\text{last Prescription} + 1) - \text{first Prescription})} * 100$$



$$PDC = \frac{4+2}{9} * 100$$

Figure 33 PDC calculation

The i2b2 DW architecture permits to handling multivariate temporal data and to choose a customized ***level of detail*** in the observations, e.g. PDC can be calculated both for active principles (identified by the ATC code) or for higher-level observation, like groups of drugs (the ATC Class). Moreover, as the data are sorted with daily granularity, it was also possible to choose on which time interval compute the PCD. Figure 34 show the procedure to compute the PDC over semesters. The results of this procedure are then stored in the DW, accounting both the level of detail and the temporal granularity chosen in the computation. The procedures in Figure 34, Figure 36 and Figure 37 have been implemented via R scripts.

PRE-PROCESS
DATA GATHERING from i2b2: Drug Purchases (PATIENT.ID, ATC, ATC.CLASS, PURCHASE.DATE, DDD)
COMPUTE SEMESTERS DDD time series from 1 DAY to 182 DAYS GRANULARITY
SUM of DDD values, group by (PATIENT.ID, ATC, SEMESTER)

DATA.FRAME= {PATIENT.ID, SEMESTER , ATC, ATC.CLASS, DDD}

PDC COMPUTATION compute_ProporitionDaysCovered(DATA.FRAME)
%This procedure can be done both for ATC and ATC.CLASS

```
For each patient PATIENT.ID
        For each ATC or ATC CLASS
                For each SEMESTER
                Compute PDC = DDD/182 Days
        End
return(DATA.FRAME.PDC)
End
```

Figure 34 PDC computation

To evaluate the clinical relevance of the PDC as a measures of drug exposure within the research context and the opportunity of exploiting it for further analysis, the physicians involved in the project suggested to focus on a set of specific drugs: Metformin and Lipid lowering. Among Lipid Lowering, Atorvastatin and Simvastatin were considered as, in a recent publication (Cederberg et al. 2015; Arfè & Corrao 2015), has been indicated to have correlation with the onset and the development of T2DM. This shift from an approach based on the whole treatment to a more specific one led to a more precise drug exposure approach, too. Figure 35 shows PDC continuous values distributions when PDC is computed for each patient and each drug (Metformin and Lipid

Lowering divided in Atorvastatin, Simvastatin and Other) considering the whole observation period, from the diagnosis of T2DM until the last registered purchase. Differently from other exposure measure previously used (like CSA), when the PDC is calculated over the whole exposure period, it shows more homogeneous values among different ATC classes. This fact, together with several considerations of physicians about single patient cases, suggested that the use of PDC were a valid instrument to depict the exposure to different kind of drugs in a consistent and meaningful way.

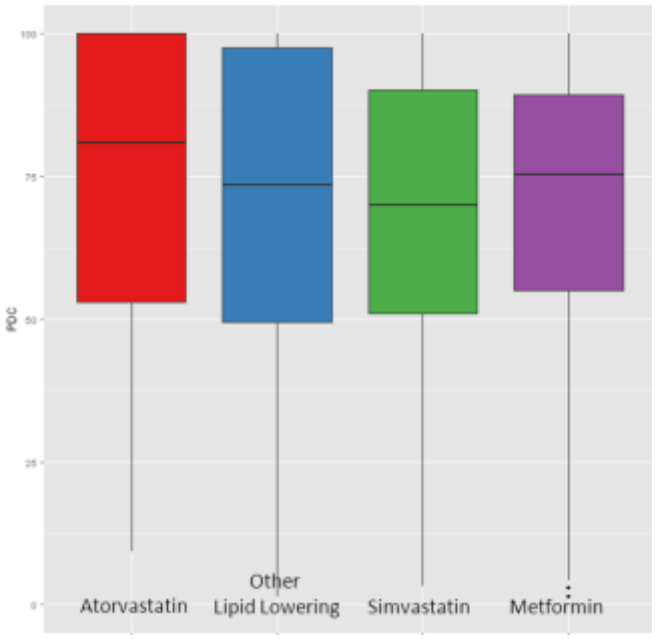


Figure 35 - PDC values calculated for the considered active principles over the entire population

5.2.2 Drug Purchasing Indicators and Tailored Thresholds.

Once the PDC values had been computed and stored in the i2b2 DW, it was possible to easily extract the data necessary to compute the indicators and thresholds that show, for each patient, his/her behavior in purchasing drugs during the evolution of the disease.

The first procedure, shown in Figure 36, was dedicated to detect if subjects have statistically significant diverse purchase patterns from the population exposed to the same drugs. In this case, the computed indicator represents a behavior kept for the entire observation period and the retrieved information account for the whole disease history not taking into account changes in time. The ***indicator*** is computed on the basis of a comparison of the median of PDC values between the patient and the population, performed via a Wilcoxon test. For each patient, and for each of the drug he/she purchased, the result of the procedure indicates if the patient behavior was in line with the rest of the population (when the p value < 0.05) or if he/she tend to purchase more or less of a certain drug (sign plus or minus).

```
Compare_PDC_values_in patient_and_population (DATA.FRAME.PDC)
%This procedure can be done both for ATC and ATC.CLASS

Compute          MEDIAN.PDC.ATC.PATIENT = median(PDC ~ PATIENT + ATC)
                 MEDIAN.PDC.ATC.POPULATION = median(PDC ~ ATC)

For each patient
                 For each ATC or ATC CLASS
                                  p.value = Wilcoxon.test(MEDIAN.PDC.PATIENT ~ MEDIAN.PDC.POPULATION)
                                  if MEDIAN.PDC.PATIENT  > MEDIAN.PDC.POPULATION sign = +
                                  else if MEDIAN.PDC.PATIENT  <= MEDIAN.PDC.POPULATION sign = -
                 End
return(DATA.FRAME.PDC.PVALUE.SIGN)
End
```

Figure 36 Patient behavior compared to population - indicator

The second procedure is shown in Figure 37 and is aimed at building, for each time interval where the PDC has been computed (for example semesters), a ***label*** that indicates if the quantity of the drug purchased in the period is under or over a certain ***threshold***. A frequently used threshold of PDC for defining a correct adherence to a certain therapy is 80% (Cheong et al. 2008; Colombo et al. 2012; Lo-Ciganic et al. 2015). This mean that a patient is covered by a certain active principle for at least 80% of the time in the observation period. Though, there is no evidence that the fixed threshold can be used in a context where the main focus in not the adherence to a prescription but a behavior in purchasing drugs. Furthermore, a clinically meaningful threshold should be adapted to a specific context, accounting for the studied disease, the analyzed drug class, the searched clinical outcomes and, above all, the characteristics of the patient (Steiner & Prochazka 1997). For these reasons, the procedure to calculate the thresholds was tailored on the single patient for each of the purchased class of drug. The 33rd and 66th percentiles of the PDC was computed for each drug (coded as ATC) purchased by the patients, the value of PDC in each semester was compared to these thresholds and associated to a label that indicates a behavior that diverge from the patient's standards (over or under purchasing a drug). Nevertheless, to give to physicians a general indication, when the PDC is in range, the label carries also the information about two fixed thresholds of 80% and 100%. The semesters during which there are no prescriptions are labelled as interruptions.

```
Tailored PDC thresholds and behavior in time (DATA.FRAME.PDC)
Compute          Under.Threshold.ATC.PATIENT =  33Percentile (PDC ~ PATIENT + ATC)
                 Over.Threshold.ATC.PATIENT = 66Percentile (PDC ~ PATIENT + ATC)

For each observation
%Define single patient thresholds
                 if PDC  <= Under.Threshold.ATC.PATIENT   label = UNDER
                 else if PDC  >= Over.Threshold.ATC.PATIENT   label = OVER

%Add information on population thresholds
                 else if Under.Threshold.ATC.PATIENT < PDC < Over.Threshold.ATC.PATIENT
                                  if PDC  <= 80  label = IN.RANGE.80
                                  else if PDC  >= 100 label = IN.RANGE.100
                                  else if 80 < PDC  < 100 label = IN.RANGE
End
return(DATA.FRAME. THRESHOLDS.LABELS)
End
```

Figure 37 Patient exposure pattern - thresholds and label

Both the procedures to derive the behavior indicators and the exposure thresholds produce results which summarize complex and rich information. To be efficiently exploited to support clinical practice a further step had to be taken beside the analysis process. The ***visual representation*** of overloaded information turns the complexity into an opportunity to gain insight, conclusions, and to support make decision making (Keim et al. 2008; Thomas & Kielman 2009).

5.2.3 The role of Drug Exposure for the Prediction of Microvascular Complication: Evaluation

To in depth assess the relevance of the selected measures and indicators of drug exposures, the possibility of including them in predictive models for the onset of complication was explored.

As advised by clinicians, it was evaluated the influence of treatment with Metformin on the onset of nephropathy at 3, 5 and 7 years after the first visit near the hospital. Specifically, logistic regression models were exploited and the analysis includes the available clinical variables measured at the first visit near the hospital plus the PDC for Metformin before such visit for each patient. PDC was computed between the first prescription after T2DM diagnosis and the first visit at hospital for each patient; PDC values were subsequently discretized in a binary variable adopting a threshold value of 90% (Figure 38).

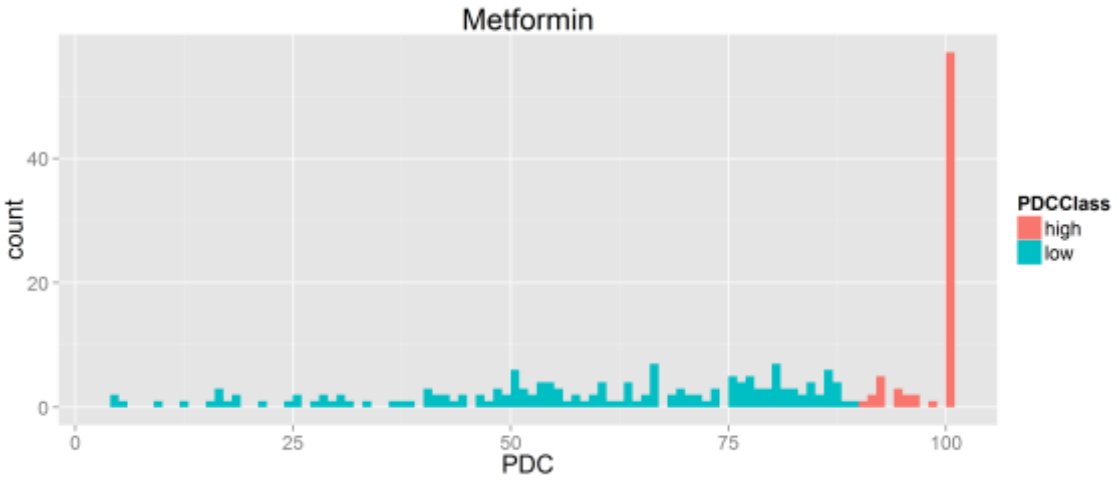


Figure 38 Distribution of the PDC values for metformin in our population before the first visit

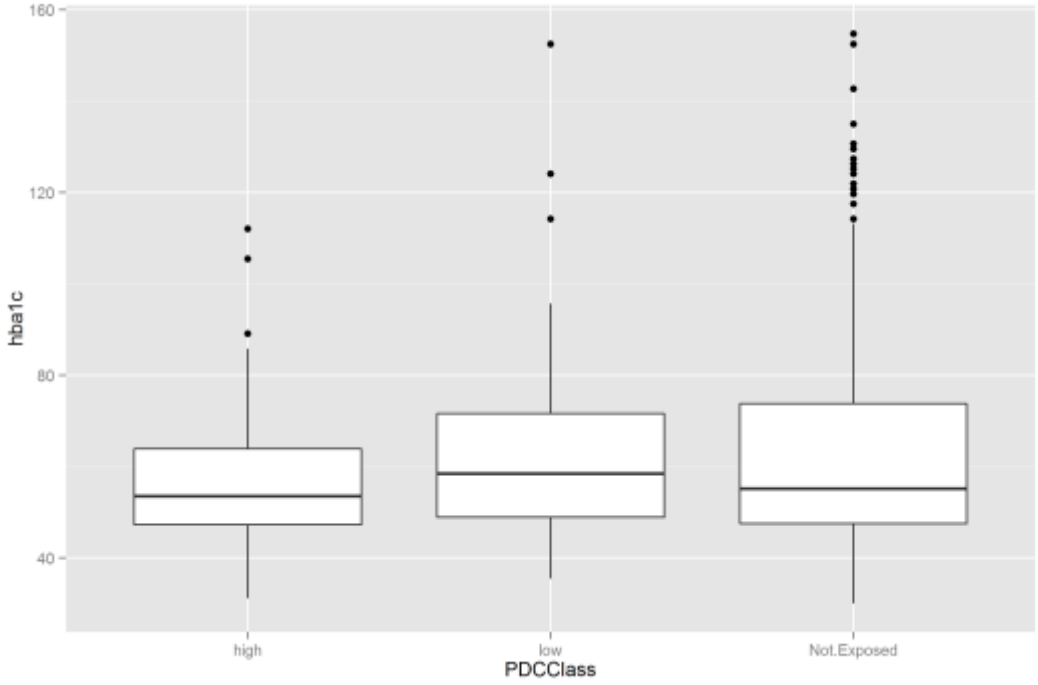


Figure 39 Values of Hba1c at first visit in our dataset for patients in 3 groups: low PDC, high PDC, not exposed to Metformin

Shows the values of Hba1c at first visit in our dataset for patients in 3 groups:

1. "low": patients who have PDC < 90%
2. "high": patients who have PDC >= 90%
3. "Not.exposed": patients who were not exposed to Metformin before the first visit

Patients in the "low" group present at first visit with higher values of Hba1c.

To be added to the logistic regression models for nephropathy, a binary PDC Class was adopted by merging patients in the "high" and "Not.Exposed" group. This choice was motivated by

considering that blood glucose control for patients in classes 2 and 3 are more similar than the ones in class 2 (Figure 39) Table 11 shows the odds ratios for all variables calculated on the entire dataset within the multivariate logistic regression. The results suggest that an irregular behavior in enacting the treatment with Metformin, which might result in a worsened glycemic control, is associated to a higher risk of developing nephropathy.

years	gender =M	age	T2DM	bmi	hba1c	hypert =Y	smoke =Y	smoke =N	PDC Class =low
3	-	-	-	1.132***	1.022***	2.382**	2.815**	0.726	2.276**
5	-	-	-	1.099***	1.019***	3.942***	2.932**	0.885**	1.788
7	1.86	-	-	1.085**	1.013**	3.517***	2.634**	1.211	2.146 .

Table 11 - Odds ratios for the logistic regression models for nephropathy

The performances of the predictive models including the additional pharmacological predictor, validated with a leave-one-out strategy, are provided in Table 12.

years	acc	sens	spec	ppv	npv	auc	brier
3	0.680	0.608	0.688	0.181	0.939	0.683	0.091
5	0.682	0.576	0.705	0.303	0.882	0.684	0.143
7	0.603	0.569	0.617	0.366	0.786	0.654	0.193

Table 12 – Performances of the logistic regression models for nephropathy (LOO validation)

5.2.4 Drug Exposure Patterns: Discussion

These results are related to the importance of the behavior of a patient in purchasing drugs. In fact, irregular purchase patterns for specific drugs, such as Metformin, shown to be related to a higher risk of developing nephropathy. These findings support the hypothesis according to which patients with an irregular behavior in following their treatments are those who also show less controlled clinical variables and, ultimately, this poor control leads to a deterioration of the patients' conditions, increasing the risk of developing a complication.

The interpretation of these results is not trivial, as drugs purchases most often reflect the result of an action of the physician based on the patient's condition. For this reason, using this variable in the calculation of the risk score could be misleading. In any case, these analyses show that this

information is worth taking into account, especially at the first visit, when the physician sees the patient for the first time after a period he/she has been treated by the GP.

Despite the complex scenarios that such results might hide, in any case it underlines the importance of showing to the user the patterns of purchase followed by a specific patient before one visit. This result justifies the inclusion of the drugs purchases visualization functionality in CDSS, as an instrument to identify potentially critical behaviors that might need closer control. The visual analytics approach exploited to integrate drug exposure measures and indicators in the CDSS is detailed in 7.2.3

CHAPTER 6

6 Risk Models for T2DM Complications and Metabolic Control Variations (Aim 2b)

T2DM is one of the main focuses of current health policies. In (Cichosz et al. 2015) authors discuss how the huge quantity of available data and information, derived from risk prediction and pattern recognition models, may be fused into new predictive models that combine patient information and analytical outcomes. The knowledge derived by such models can be used to improve disease management and patient care if integrated into the clinical decision support chain. Authors underline that, although regression analysis has been widely used for building risk prediction models, there is a notable scarcity of studies that evaluate their impact into clinical practice.

Within the MOSAIC project, the application of multivariate risk prediction models to the data gathered from the medical centers was aimed at gaining better insight on the T2DM management. Indeed, the continuous stratification of T2DM patients' risk may enhance the care processes, optimizing both clinical pathways and resources allocation. In particular, classification techniques, ***survival analysis and recalibration approaches*** have been applied on the datasets coming from the hospitals, towards the development of hospital-based, personalized T2DM ***complications risk prediction models***. The developed models were then evaluated in on the available datasets and on a new data set collected near the Pavia hospital FSM.

Further efforts focused on exploring additional models able to explicitly take into account the longitudinal nature of data, both in relation to T2DM evolution in terms of complications and to the ***assessment of blood glucose control***. These longitudinal models have been developed to analyze which factors increase the risk in having worsening conditions, especially in relation to glycated hemoglobin (Hba1c) variability. The results of the retrospective validation were carried out on the Pavia dataset. The two main questions that have been answered throughout this step of the work are the following:

i) Which are the factors that put patients at risk of adverse events during his/her disease evolution? To answer this question, the role of the risk factors through Cox regression analysis was assessed. Moreover, the impact of variations in metabolic control on the development of microvascular complications has been analyzed. In particular, it has been studied:

- How Hab1c variability in time is correlated with microvascular complications, especially with nephropathy? This question was explored also referring to a recent study on the Italian population.
- How Hab1c trajectories may impact on the evolution of cardiovascular risk? Continuous Time Bayesian Network were exploited to represent the disease evolution through clinical variables interactions.

ii) Which are the factors associated to a poor management of Hba1c? Is it possible to predict Hba1c variations between visits? To answer this question, the factors that trigger worsening metabolic control were identified. In particular:

- A model able to depict how population and patients' variability affects variation of Hba1c between follow-ups was implemented using Bayesian Hierarchical Models.
- It has been studied how environmental factors influence the variation of Hba1c during disease evolution on the basis of the integration of remote sensing and clinical data.

6.1 Complication Risk Model Development and Validation

This paragraph illustrates the results of the analysis pipeline applied to the data set collected near the FSM hospital of Pavia and the local healthcare agency of the area (ASL) to derive T2DM complications risks models that can be proficiently integrated in the CDSS. The analysis pipeline is made up of four sequential steps, as synthesized in Figure 40.

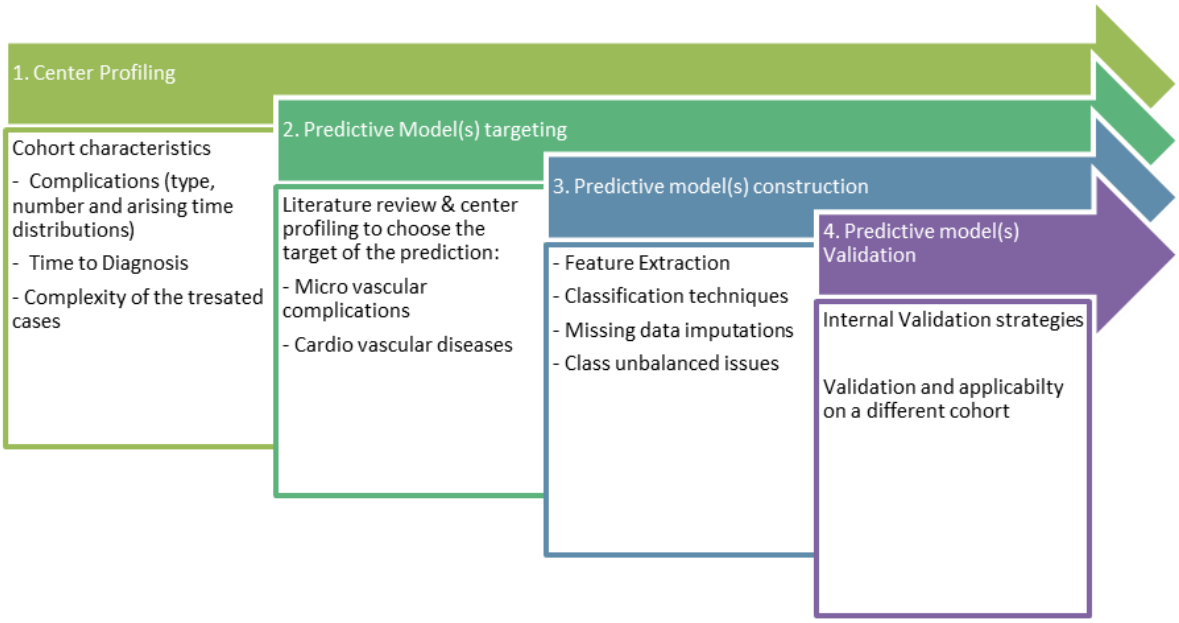


Figure 40 Risk model pipeline

The ***center profiling*** step is aimed at assessing the hospital characteristics in terms of population (number of patients with complications, time to diagnosis of the complications) and of patterns of care (e.g. centers that are used to deal with more complex cases, centers that perform an initial intensive diagnostic program to discover complication early after the first visit). The variables considered in the analysis include demographic data (age, gender, time to diagnosis), clinical data from the EHR (Body Mass Index, glycated hemoglobin, lipid profile, smoking habit) and administrative data (antihypertensive therapy) of a population of 943 T2DM patients in charge of the FSM hospital Fondazione Salvatore Maugeri and involved in the healthcare system of ASL, which data were stored in the i2b2 MOSAIC DW. As clinical data were only available after the first near the hospital, it was important to detect the proportion of cases occurred before and after the first visit at FSM for each complication.

On the basis both of center profiling and literature review it was possible to ***target*** the possible different modeling strategies to assessing the risk of developing complications to a specific center data set. The analysis focused on deriving predictive models for microvascular complications in the population: nephropathy, neuropathy, and retinopathy. In the studied population, microvascular complications account for a larger number of cases developed after the first visit as compared to macrovascular complications. This makes possible to build predictive models based on data collected at the first visit at FSM. Moreover, the validated Progetto Cuore score for cardiovascular risk was already in use in the clinical practice at FSM and included in the data set (Risk score and complexity index). The literature search targeted predictive models of the microvascular complications. The most interesting results of the search resulted in 3 papers, all performed under the United Kingdom Prospective Diabetes Study (Stratton et al. 2000) (Stratton et al. 2001) and (Retnakaran et al. 2006). Several drawbacks were though identified to their application. The main problem was that they all consider data taken at the diagnosis of diabetes, while given the nature of the studied dataset, clinical information is available only from the first visit at the hospital. An additional problem is that none of these studies present a validation of the models in terms of prediction accuracy. Taking into account all these observations and also following the advices of the clinicians taking part to the project, a new set of models on the available data were developed and evaluated.

Once the target of the modelling has been selected the final methods to ***construct*** the predictive models for the risk of complications were identified. Given the patient's health status at the first visit, the aim was to predict if the patient will develop nephropathy, neuropathy or retinopathy in the future. Distinct models were built for each complication, considering a temporal threshold for risk prediction of 3 years, 5 years, and 7 years. The binary class variable in the models corresponds to whether a patient develops the complications within a number of years after the first visit inferior or equal to the threshold value. The classification models considered for the analysis are Logistic Regression (LR) and Naïve Bayes (NB). Even if collected, some of the variables had many ***missing data***, which represent a critical problem to be addressed. Lipid-related data, in particular, are very prone to missing values. For the data imputing approach, two simple statistical methods (i.e. imputing the mean and median of each variable) and a Random Forest approach were considered. The latter method is based on the Random Forest imputation algorithm (Stekhoven & B??hlmann 2012). Imputation performances were compared by measuring the root mean squared error (RMSE) and the normalized root mean squared error ($RMSE_N$) on the artificial missing values of the data-complete set. Random Forest imputation algorithm outperformed imputation with mean or median, and therefore was chosen as our data imputing method. Given the uneven number of patients with and without complications, the resulting classification problems were characterized by an ***unbalanced distribution*** of the class variable. To try to rebalance the cases/controls ratio, the algorithms were trained on a new data set that balanced by oversampling the minority class. In addition, for NB, while the marginal probabilities are estimated on the training sets balanced with oversampling, the prior probability of the class was computed on the original unbalanced dataset. In the following the models resulting from the class balancing strategy are denoted as "LR balanced" and "NB balanced+ adjusted prior". Models built on the original dataset are denoted simply as "LR" and "NB".

The final step was devoted to the ***validation*** strategy to assess the performance of the selected methods was defined. Following the possible option described above, 4 models were built (LR, LR

balanced, NB, NB balanced + adjusted prior) for 3 complications using 3 temporal thresholds. For each model, for each complication, and for each temporal threshold, data with or without imputation and with or without considering lipid-related variables was considered, as shown in Table 13.

Models Scenario				
Complications	Nephropathy	Neuropathy		Retinopathy
Time horizon	3 years	5 years		7 years
Lipid-related data included	Yes		No	
Imputation	Yes		No	
Predictive Models				
Model	LR	LR (balanced)	NB	NB (balance + adjusted prior)

Table 13 - List of explored options for feature extraction and model design.

The performances of the models were evaluated with a leave-one-out validation strategy. For each model Sensitivity, Specificity, Accuracy, positive predictive value, negative predictive value, Brier score, and Area under the ROC and Matthews Correlation Coefficient were measured.

6.1.1 Preliminary Results and Validation on the existing cohort

For each of the developed models, the number of cases and controls included in the training and test phases are shown in Table 14. As the number of years set as the temporal threshold increases, the total number of patients decreases since more subjects in the dataset lack a clinical history long enough to meet the including criteria

Complication	Within years	# cases	# controls	# total
Retinopathy	3	55 (8.7%)	574 (91.3%)	629
	5	67 (15%)	379 (85%)	446
	7	72 (23%)	241 (77%)	313
Nephropathy	3	61 (9.4%)	591 (90.6%)	652
	5	79 (17.1%)	382 (82.9%)	461
	7	90 (27.4%)	239 (72.6%)	329
Neuropathy	3	66 (10.5%)	560 (89.5%)	626
	5	76 (17.5%)	359 (82.5%)	435

	7	87 (28.1%)	223 (71.9%)	310

Table 14 - Number of cases and controls for each model

As the performances of the models did not improve when the imputing strategy was adopted to handle missing values so to include lipid-related variables in the analysis, in the following are shown the results obtained on the non-imputed dataset without lipid-related data. The reported results have been obtained using the following predictors: gender, age, time since diabetes diagnosis (T2DM), BMI, HbA1c, hypertension and smoke.

The next tables describe the values of the odds ratios for the independent ***risk factors*** as obtained by means of the multivariate LR analysis. Reported results are obtained developing the model on the entire dataset and using a feature selection procedure based on the Akaike Information Criterion. Odds ratios for the continuous variables age, T2DM, bmi and hba1c correspond to how much the odds increase with an increment in the continuous variable of 1 year, 1 year, 1 Kg/m2 and 1 mmmol/mol respectively.

Retinopathy									
years	**gender =M**	**age**	**T2DM**	**bmi**	**hba1c**	**hypert= Y**	**smoke= Y**	**smoke= N**	**smoke= ex**
3	-	-	1.122***	-	1.030***	2.394*	-	-	-
5	-	-	1.098***	-	1.037***	4.379***	-	-	-
7	-	-	1.075***	-	1.044***	4.793***	0.471	1.528	-
Nephropathy									
years	**gender =M**	**age**	**T2DM**	**bmi**	**hba1c**	**hypert= Y**	**smoke= Y**	**smoke= N**	**smoke= ex**
3	1.913 .	-	-	1.118**	1.022***	3.574***	2.320*	0.855	-
5	1.678	-	-	1.097**	1.019**	4.905***	2.294*	0.924	-
7	2.014*	-	-	1.072*	1.018**	4.016***	-	-	-
Neuropathy									
years	**gender= M**	**age**	**T2DM**	**bmi**	**hba1c**	**hypert= Y**	**smoke= Y**	**smoke= N**	**smoke= ex**
3	-	-	1.054**	-	1.029***	-	0.909	0.433*	-
5	3.202**	1.035 .	1.059**	1.057	1.033***	-	-	-	-
7	2.506*	-	1.071***	1.071 .	1.027***	-	-	-	-

Table 15 – Odds ratios for the logistic regression models for the three complications (significance codes: 0 '*', 0.001 '**', 0.01 '*' 0.05 '.', >0.1 ' ')**

Selected variables for single complications across different temporal threshold show a high degree of consistency. As expected, Hba1c value is included as an independent risk factor in all models

(Stratton et al. 2000). In 2 instances, the obtained odds ratio for the smoke variable seems to contradict expected results: it is worth mentioning that in both cases the related coefficient is not significant in the logistic regression model.

Results of the ***leave-one-out validation*** procedure in terms of Area Under the ROC Curve (AUC) for each scenario and modelling strategy are shown in Table 16. The ROC curves obtained on the original dataset and the ones obtained on datasets balanced with respect to the distribution of the observations in the 2 classes appear very close in most scenarios.

Retinopathy				
Years	**LR**	**LR balanced**	**NB**	**NB balanced**
3	0.832 (0.771 – 0.894)	0.818 (0.749 – 0.879)	0.774 (0.708 – 0.839)	0.780 (0.711 – 0.840)
5	0.821 (0.758 – 0.884)	0.826 (0.758 – 0.881)	0.757 (0.691 – 0.823)	0.763 (0.691 – 0.823)
7	0.779 (0.707 – 0.851)	0.788 (0.712 – 0.859)	0.743 (0.676 – 0.811)	0.747 (0.679 – 0.815)
Nephropathy				
Years	**LR**	**LR balanced**	**NB**	**NB balanced**
3	0.703 (0.626 – 0.781)	0.730 (0.594 – 0.749)	0.693 (0.622 – 0.763)	0.697 (0.628 – 0.766)
5	0.701 (0.629 – 0.773)	0.725 (0.602 – 0.771)	0.694 (0.633 – 0.756)	0.693 (0.631 – 0.755)
7	0.674 (0.598 - 0.749)	0.707 (0.623 – 0.784)	0.692 (0.618 – 0.743)	0.691 (0.629 – 0.753)
Neuropathy				
Years	**LR**	**LR balanced**	**NB**	**NB balanced**
3	0.647 (0.557 - 0.736)	0.680 (0.602 – 0.771)	0.692 (0.617 – 0.767)	0.689 (0.614 – 0.767)
5	0.711 (0.633 – 0.788)	0.697 (0.623 – 0.784)	0.695 (0.625 – 0.765)	0.689 (0.619 – 0.760)
7	0.696 (0.615 – 0.777)	0.671 (0.622 – 0.782)	0.694 (0.624 – 0.763)	0.697 (0.627 – 0.763)

Table 16– Values of AUC for each model (leave-one-out validation)

On the basis of the previously shown results, the models selected were the LR with feature selection based on the Akaike Information Criterion as the standard models for ***use in the clinical practice***. LR models are easily understandable from a clinical point of view, as they provide an intuitive interpretation of the parameters. Moreover, they are probabilistic classifiers, and thus provide an insight in the inherent uncertainty associated to the predictions. The performances of the models as evaluated with the leave-one-out validation strategy are provided in Table 17.

Retinopathy								
Years	**acc**	**sens**	**spec**	**ppv**	**npv**	**AUC**	**brier**	**MCC**
3	0.838	0.435	0.872	0.223	0.948	0.833	0.063	0.241
5	0.8	0.479	0.851	0.343	0.910	0.821	0.097	0.296

7	0.75	0.538	0.809	0.444	0.861	0.779	0.142	0.332
Nephropathy								
3	0.838	0.36	0.887	0.253	0.928	0.703	0.083	0.210
5	0.753	0.453	0.817	0.349	0.873	0.701	0.136	0.244
7	0.688	0.591	0.725	0.456	0.819	0.674	0.187	0.290
Neuropathy								
3	0.798	0.352	0.852	0.223	0.916	0.647	0.092	0.164
5	0.778	0.5	0.839	0.408	0.884	0.711	0.133	0.310
7	0.696	0.549	0.757	0.487	0.8	0.696	0.184	0.296

Table 17 – Models performances

Logistic regression allows calculating the probability of developing a complication within a specific time period, providing a way to calculate a risk score for the patients. This could be particularly interesting at the first visit, to suggest to the doctors those patients that might need particular attention. On the basis of obtained results, several considerations have been done with the medical experts in order to include the risk score in the CDSS and provide them a meaningful tool to monitor the status of patients during follow ups. Considering the performances of the models in terms of Matthews Correlation Coefficient (MCC), which is particularly suitable in case of unbalanced distribution among classes (Ramyachitra & Manikandan 2014; Bekkar et al. 2013), and the pace of the disease evolution the 5 years' time horizon has been chosen to be included in the system.

6.1.2 Validation on a new cohort

The developed models were validated externally on 105 previously unseen patients diagnosed with T2DM selected by the doctors at FSM according to the following criteria: (i) Not in the MOSAIC cohort, (ii) At least 7 years of follow-up, (iii) included in the Local Healthcare Agency of Pavia registries, (iv) At first visit not all retinopathy, nephropathy and neuropathy should be present. The descriptive data of the patients can be found in Table 18

	FEMALE	MALE
Number of pts	40	65
AGE	mean(61.33)sd 9.9	mean(58.89)sd 8.57
DIABETES DURATION	mean(6.42)sd 9.14	mean(4.58)sd 6.41
SMOKING YES	7(17.5%)	22(33.85%)
BMI	mean(30.75)sd 5.66	mean(29.68)sd 4.88
HBA1C	mean(63.84)sd 23.09	mean(68.95)sd 26.42

HYPERTNS	23(57.5%)	39(60%)
Retinopathy	9(22.5%)	16(24.62%)
Nephropathy	7(17.5%)	15(23.08%)
Neuropaty	13(32.5%)	22(33.85%)

Table 18 Descriptive data on the 105 new FSM patients.

Table 19 shows the results of applying the CDSS selected LR models at 5 years to the external dataset made up of 105 patients. Models were evaluated using the MCC. The results show good generalization performance of the developed models, with AUCs close to the ones obtained for internal validation.

COMPL	acc	sens	spec	ppv	npv	auc	brier	MCC
RET	0.869	0.917	0.688	0.306	0.982	0.822	0.117	0.417
NEU	0.911	0.667	0.757	0.261	0.946	0.706	0.094	0.296
NEPH	0.965	0.500	0.952	0.333	0.975	0.714	0.091	0.373

Table 19 Performance of the models (AUC) among the 105 new FSM patients

An addition analysis was performed to further study the results on the external patient cohort. For each patient and for each complication, the selected predictive models were exploited to compute the probability of developing a complication within five years. Such values have then been plotted by dividing patients in two group based on the real onset of the complication (group 0: patients who did not develop the complication within 5 years, group 1: patients who developed the complication within 5 years). The probability values in the two groups were compared using the Wilcoxon test. Significant results were obtained (p-value < 0.01) for all the complications, and higher values were found of the probabilities for the group of patients that have the complication.

6.1.3 Cox regression analysis to predict the onset of microvascular complications

The Cox regression analysis was performed to take into account the longitudinal nature of the onset of the complication. This analysis allowed confirming the role of the risk factors identified with the logistic regression models also in a survival model.

The Cox regression model was developed on the basis of the population of patients used to train the Logistic Regression models and previously described. Thanks to the collection of new data during the last project months some information in the dataset was updated. The results are shown in Table 20. For each complication and for each predictor the hazard ratio and an indication on the statistical significance of each parameter are reported. The significant risk factors found using the Cox analysis are exactly the same found using logistic regression models.

	gender	age	T2DM	bmi	hba1c	hypert	smoke
Retinopathy	1.28	0.99	1.07***	1.03	1.03***	1.8*	0.58
Nephropathy	1.84*	0.99	1.01	1.08**	1.01*	2.14*	1.30
Neuropathy	2.1**	1.01	1.06***	1.02	1.02***	0.68	1.34

Table 20 - Cox regression analysis results: Hazard ratios

6.1.4 HbA1c variability impact on Neuropathy

A recent Italian study, the RIACE study (Penno et al. 2013), examined the correlation between hemoglobin HbA1c variability and microvascular complications. Authors found that HbA1c variability was independently correlated to nephropathy, but not to retinopathy. The logistic regression analysis performed within the Riace study shows that in T2DM patients, what affects nephropathy is HbA1c standard deviation (SD) rather than HbA1c mean value. The Riace study represents one of the most recent and complete studies about T2DM microvascular complication prediction based on the data of an Italian cohort. The applied models take into account the longitudinal nature of data while depicting metabolic control variability, in order to allow microvascular complications prediction.

The difficulties of the application of other microvascular models to the Mosaic population have been already discussed. These problems were mainly related to the nature of the exploited data, as the main part of these previous studies were cross sectional and they considered data taken at the diagnosis of diabetes. As the Riace study takes into account the data of patients 2 years before a nonspecific visit during the history of the disease and not from the diagnosis, it was possible to apply a similar approach to the Mosaic data.

In order to have complete time series of clinical variables, it has been selected a random visit at least 2 years after the first visit at the hospital and considered the data before that visit. In this way it was possible to apply the same analysis framework exploited in the Riace study to assess the replicability of its results on the MOSAIC population. Models show that values of Hba1c - SD higher than 9.22mmol/mol influence the risk of developing nephropathy (OR 3.49, pvalue <0.001). No correlation was found between Hba1c SD and retinopathy or neuropathy.

On the basis of these results, the analysis of how Hba1c variability may affect the onset of nephropathy, especially when compared to Haba1c mean values were refined. The main objective was to have a cohort were cases and controls had a similar distribution of Hba1c mean values and then assess if there were significant differences in Hba1c Standard Deviation, which it has been used to assess HbA1c variability in the observed period. To create matched groups of cases and controls, Hierarchical Clustering and Bootstrap was used. The exploitation of Hierarchical Clustering allows to stratify the whole cohort and find groups of patients with similar profiles in terms of average Hab1c values, treatment and demographic variables. Figure 41 synthesizes the steps to obtain the paired cohort. From the original unbalanced cohort, the hierarchical clustering was used to stratify the population. Once the groups were identified, it has been sampled a number of controls from each group equal to the number of cases in the same group. In this way, not only

the final cohort was balanced from the point of view case/control rate, but cases and controls have also comparable distributions of the variable used to create the clusters, in particular Hba1c mean values.

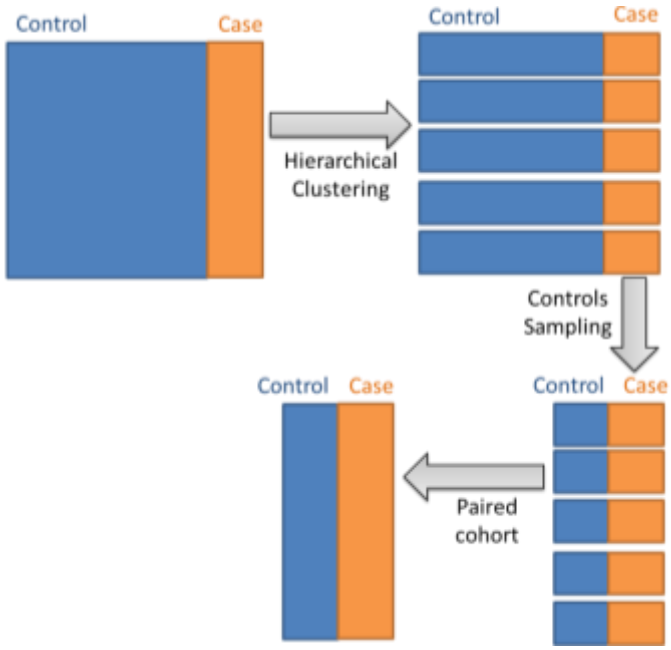


Figure 41 Steps performed to obtain a balanced dataset where cases and controls are matched on HbA1c, age, treatment and follow-up time

The next step was to assess if there was any significant difference in terms of Hba1c Variability between cases and controls. After assessing Hba1c SD distribution as normal (Shapiro Test, W =0.7, p <<0.01), it has been applied t-test on Hba1c SD and found that patients affected by nephropathy had higher variability (p-value = 0.022) in their metabolic control. To double-check the matching procedure, Hba1c mean values in the two groups were compared, obtaining a not significant difference (p-value = 0.21), as expected. As last step of the analysis a logistic regression on the paired cohort was performed, including Hba1c MEAN, SD and CV continuous values as additional predictors of nephropathy. The obtained model shows that the only variable selected to predict nephropathy was Hab1c variability in term of SD. In a population where cases and controls have comparable values of average Hab1c, the Hab1c variability significantly impacts ($p < 0.01$) on the presence of the complication.

6.2 Continuous Time Bayesian Networks, study of T2DM disease trajectories

Clinical data coming from EHR are sampled with a frequency that is variable among patients (days to years), and that depends on the frequency of follow-up visits at the hospital. The application of time-discrete models to this type of data requires discretizing the sampling process to a user-defined pace, with possible loss of information. Continuous Time Bayesian Networks (CTBNs) are a modeling technique that embeds this irregular recording pace in their structure (Nodelman et al. 2002; Stella & Amer 2012).

In CTBNs, variables are modeled as discrete nodes that evolve over continuous time as functions of a conditional Markov process. Each variable is represented as a node and its values discretized on the basis of a suitable threshold. This means that Discrete nodes can describe whether vital signs are (or are not) within safe boundaries suggested by medical literature (e.g. triglycerides < 150). In addition, CTBNs handle nodes changing state at different temporal granularities (i.e. fast evolving variables such as triglycerides, and slow evolving such as health complications). For these reasons, the CTBN framework is well-suited to describe patient vitals trajectories for the purpose of disease assessment and prediction.

CTBNs are typically designed to consider each node as dynamic, i.e. a node that evolves according to its conditional intensity matrix (CIM). Intuitively, we noticed that this cannot be applied to all the considered variables, since some of them, such as age, sex or time after disease diagnosis, have a deterministic, look-up table driven evolution. Admitting deterministic variables makes the model flexible enough to explore prognostic/screening/treatment scenarios, where we can simulate what would have happened if some external conditions were verified. For example, we can answer questions such as "How many cardio/vascular complications do we have to expect in the coming years, if all patients will keep their pressure under the risk threshold?"

6.2.1 Description of the models, analysis strategy and results

The CTBN approach has been modified to include deterministic nodes, and utilized our model to describe the trajectories of disease in the Mosaic cohort. Using this model, it was possible to learn a data-driven network. After learning the network structure on the basis of the data, the model is used to perform simulations over a cohort with the same characteristics. Starting from the network learned from training data, a test data set was used to build a new, simulated, cohort on the basis of the learned parameters. On this simulated cohort, errors were measured as the percentage of wrongly predicted states, obtaining an error of about 10% on a seven-years horizon on a test data set, not used in the learning phase. It is interesting to note that the errors shown by simulated states tend not to increase with the simulation time. This means that the model is able to describe the main features of a Diabetes patients' cohort with good generalization performances within the considered time frame.

The introduction of ***deterministic variables*** not only makes the model viable to describe medical conditions, but allows us to tweak it to explore possible scenarios. As an example, to investigate the importance of Hba1c control on the patients' cardiovascular risk, systolic blood pressure, and cholesterol, we studied best (or worst) case scenarios where Hba1c was artificially kept under (or over) the risk threshold. Variables trajectories in best, normal and worst case scenarios resulted significantly different from a statistical point of view, confirming the pivotal role of hba1c in diabetes and, in particular, its effect on complications onset. At the same time, the best case scenario allows to explore the benefits, for example, of an efficient treatment able to keep blood glucose under control. In other words, we can use our model to simulate the consequences of a medical procedure affecting one node (such as hba1c) on all the other nodes (such as the cardiovascular risk).

CTBNs require nodes values to be discretized. Hba1c and CVR were made binary on the basis of clinical knowledge, while other variables were discretized using their median values as a cutoff (Table 21).

Hba1c (mmol/mol) – Thresholds taken from Literature
• <=58 → low • 58 → high
CVR (computed using the Progetto Cuore thresholds)
• 1,2,3 → low • 4,5,6 → high
Age, diabetes duration, BMI, triglycerides, cholesterol, Systolic blood pressure (Thresholds computed as Median Values)
• Age → 66 Years • Diabetes duration (T2DM)→ 7.15 Years • BMI → 28.8 • Triglycerides → 116 mg/dl • Cholesterol → 185 mg/dl • Systolic Blood pressure (SBP) → 131 mmHg

Table 21 – Variables discretization for CTBN models

The ***CTBN network learned*** from the data set is shown in Figure 42. Deterministic nodes are shown in green. In summary, the direct relationships that were found in the data are the following:

- Age, gender, systolic blood pressure → Cardiovascular risk
- Hba1c, Total Cholesterol → Systolic blood pressure
- Triglycerides → Total cholesterol
- Time from diagnosis → Hba1c

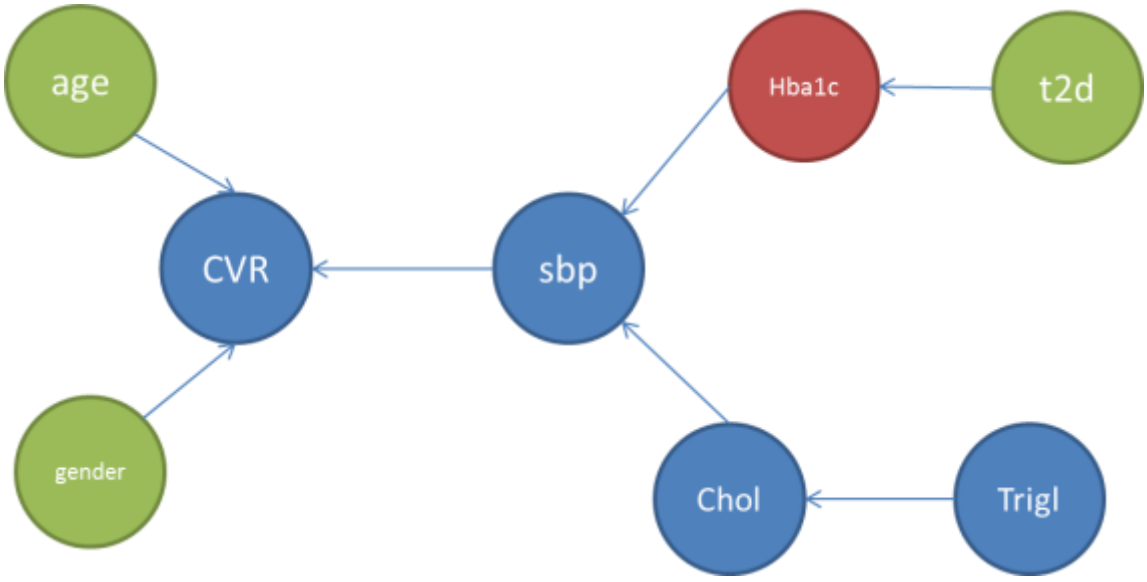


Figure 42 CTBN learned from training data

Once the CTBN structure was derived from training data, it was exploited to ***simulate*** a new cohort from test data. In the following are shown the error rates derived from the comparison of patients characterized by specific states (high value of SBP, CHOLESTEROL or CVR) in the test set (real patients' data) and in the data simulated from the test set (for each real patient we simulated 50 patients). Errors were computed as the absolute difference between the test set and the simulated test set of the % of patients with high values using three different granularities: days (as in the original data), or averaging over weeks and months. Figure 43 shows the errors computed using month granularity. Errors stay in general under the 13%. For CVR, the error stays under the 11% until the 6th year of simulation. This means that the percentage of patients simulated by the model to have high CVR values differs by less than the 11% from the percentage of patients having "real" high CVR values in the test set.

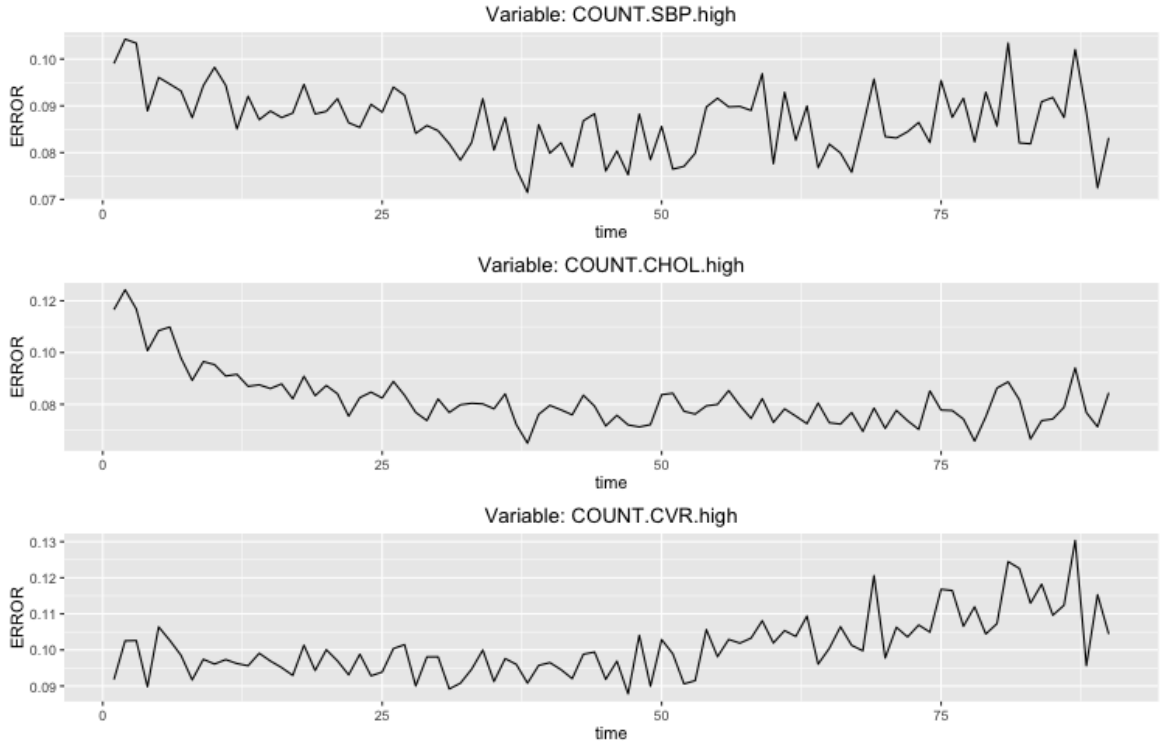


Figure 43 Error rates derived from the comparison of patients in specific states (high value of SBP, CHOLESTEROL or CVR) in the test set (real patients data) and in the data simulated from the test set (for each real patient we simulated 50 patients).

In order to compare the learned network with a base line, it was decided to run a simulation with a network without any arc among nodes, called Network Zero, as its dynamic structure is empty. We compute RMSE (as the % of patients with high Systolic Blood Pressure, Cholesterol and CVR) in the LEARNED and ZERO network, for each year in the simulation. In the Learned network RMSE is always under 0.1 for Systolic Blood Pressure and Cholesterol, while for CVR is under 0.1 until the 5th year of simulation. Errors in the Zero network are slightly higher than in the learned network for Systolic Blood Pressure, Cholesterol, while the lack of the combined effect of these variables with Hba1c has a consistent effect on the CVR errors, which is always above 0.1 in the Zero Network.

When CVR errors were compared with month granularity using a Wilcoxon test, errors in the Learned network were significantly ($p << 0.01$) lower than in the Network Zero (Figure 44).

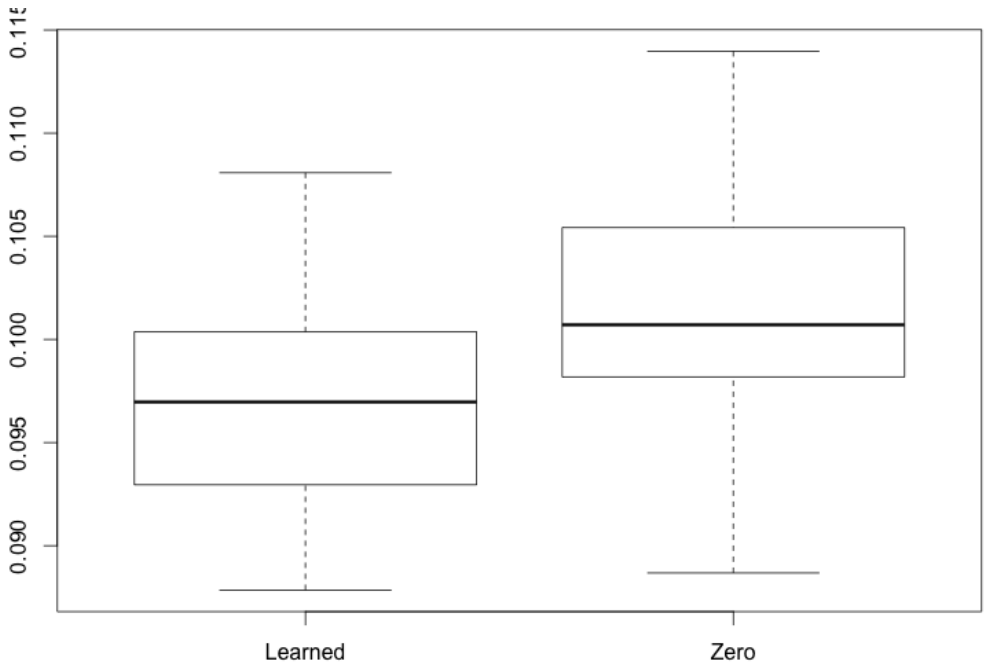


Figure 44 Error distribution in the learned network and in the Zero Network.

6.2.2 "What if" scenario, effects of Hba1c control on CVR.

Once the simulation error was assessed and the Learned network compared with the baseline, it was possible to explore a specific scenario, as defined by the question: "*How many patients with high risk of cardiovascular events do we have to expect in the coming years, if all patients will keep their Hba1c under the risk threshold or if all the patients will be over out of metabolic control?*"

It's important to highlight that this kind of simulation is not possible with the Zero Network, as it is not possible to study the effect of a specific variable on another one in continuous time without a learned dynamic structure. In the following, the dynamic components (CIM) for the variables involved in the simulation are shown. Besides the previously described simulation (here referred as A1c.real, in blue), other two simulations were run on the basis of the learned network and its components on a 10-years horizon.

- A1c.high, where we set all the values of Hba1c in the original test data set to 1 (meaning high, out of control value)
- A1c.low, where we set all the values of Hba1c in the original test data set to 0 (meaning low, controlled value)

Figure 45 shows the rate of patients with high CVR values per year in the three different simulated scenarios. Percentages of patients with high CVR are statistically different ($p<<0.01$).

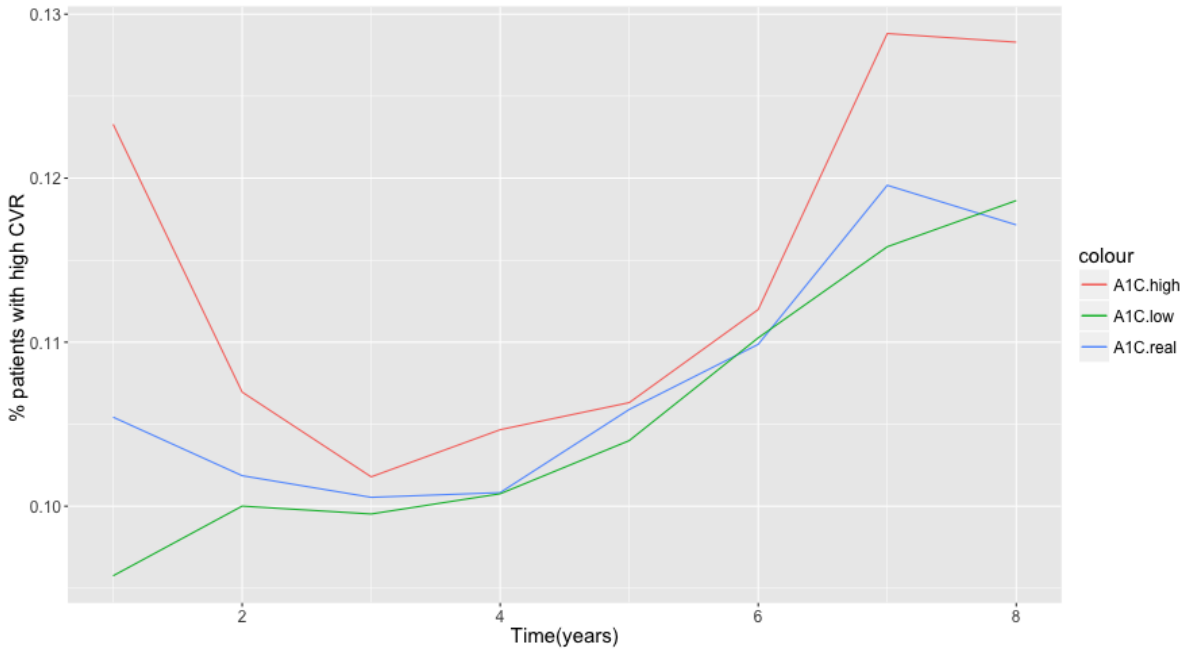


Figure 45 Percentage of patients with high CVR values in the three simulated scenarios: real HbA1c value, HbA1c value set to high for all the patients, and HbA1c set to low for all the patients

It is possible to notice that a bad control in Hba1c (red curve) may influence the risk of cardiovascular disease, especially at a certain point of the trajectory: after the 6th year, the CVR trajectory that correspond to a High Level of Hba1c diverges from the other two, showing an increasing number of patients with higher risk. Thanks to CTBN trajectories simulations, it is possible to understand how a certain scenario is evolving in time and detect the moments, in the population history, when is more likely to have deviations from clinical control.

6.3 Hierarchical Bayesian LR, prediction of HbA1c variability from one visit to the next one

The contents and the results of this section have been recently accepted for oral presentation and published in the AMIA conference proceedings [https://amia2016.zerista.com/event/member?item_id=4935184].

As discussed previously, the initial studied problem regarded the selection of suitable calculators for the risk of complications to be used in our patient group. A validated score, the Progetto Cuore score, was already available for cardiovascular risk estimation. Such score is derived from the Framingham study and adapted to the Italian population. On the other hand, given the very nature of the data available in Pavia, it was not possible to identify in the literature any microvascular risk model that could be applied without a consistent calibration drift. Therefore, it has been decided to build new predictive models based on data collected at the first visit for microvascular complications: Retinopathy, Nephropathy and Neuropathy. The outline of the above described analyses follows a cross-sectional approach and the risk calculators could be particularly useful when employed during the first visit, to alert the clinicians in case a patient might need particular attention. The developed models are suitable for the prediction of chronic complications onset.

However, they cannot be used when monitoring patients during disease progression, especially when the goal is to predict variations in complication risk biomarkers, like Hba1c, during consecutive follow-ups.

Another limit in deploying risk calculators for patients monitoring is to cope with external and internal heterogeneity, where external heterogeneity is the one represented by the differences among patients within a population, and internal heterogeneity is the one produced by variations in single patient's state over time. A Bayesian hierarchical model was applied, able to deal with both kinds of heterogeneity, in order to predict Hab1c variations from one encounter to the next 12 months.

6.3.1 Description of the models, analysis strategy and results

Hierarchical Bayesian Logistic Regression models are of peculiar interest when data are characterized by repeated measures, i.e. follow-up, by units of observation, i.e. patients. In this case two models are jointly applied: one model is used for "within unit" analysis, dealing with internal heterogeneity, and another model for "across units" analysis, dealing with external heterogeneity. The Bayes theorem is used to integrate the two models and to properly account for the uncertainty in the data, allowing individual learning by borrowing strengths from population data.

Given $h=1, \ldots, N$ patients, and $i=1, \ldots, n_h$ measurements available on the h-th patient, collected on a feature vector x of m "monitoring" variables, and on an outcome binary measure y, the probability that the outcome occurs (say metabolic control worsening) is described by the logistic model:

$$P(y_{hi} = 1|x_{hi}) = \frac{\exp(x_{hi}^t \beta_h)}{1 + \exp(x_{hi}^t \beta_h)}$$

The parameter vector of the h-th patient, β_h, is assumed to be a stochastic variable described by a "population" linear model:

$$\beta_h \sim N(\Delta^t z_h, V_\beta)$$

where $N(.,.)$ is the Gaussian p.d.f., z_h is a vector of s "static" covariates, such as gender, Δ is a $s \times m$ matrix of population parameters that associates static and monitoring variables and V is a $m \times m$ covariance matrix.

In order to perform Bayesian inference from the data, the population parameters are typically provided with a suitable prior choice. In our case, we will consider:

$$vec(\Delta|V_\beta) \sim N(vec(\overline{\Delta}), V_\beta \otimes A^{-1})$$

$$V_\beta \sim IW(v, V)$$

where $vec(.)$ is the vector representation of the elements of a matrix, A is a suitable prior precision matrix of size $s \times s$, $\otimes$ is the Kronecker product (the output is the covariance block matrix for vec(.) of size $m \cdot s \times m \cdot s$.), IW is the Inverse Wishart distribution, and the prior hyperparameters ($\overline{\Delta}$, A, v, V) are typically selected to generate diffuse priors.

Figure 46 shows a representation of the Hierarchical Bayesian Logistic Regression model through a Bayesian Network with plates. In this figure, rectangles represent observations and circles represent probabilistic variables (parameters and outcome). The inner plate is referred to temporal variables, whereas the external plate is referred to the static variables. Nodes outside the plates are the population parameters.

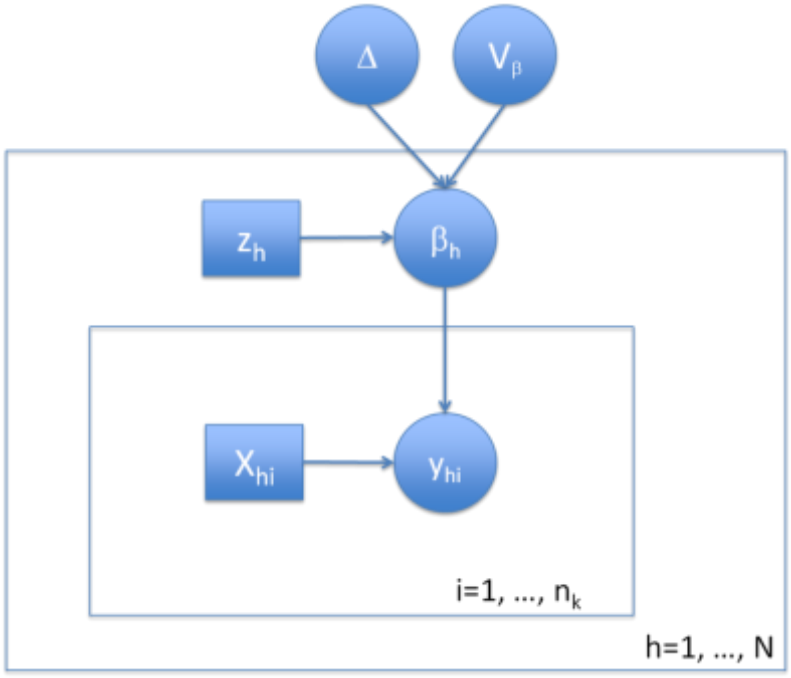


Figure 46 Bayesian Network with plates describing the Hierarchical Bayesian Logistic Regression models.

The estimate of such models is usually performed by resorting to Markov Chain Monte Carlo (MCMC) methods. Instead of deriving the analytic form of the posterior distribution, the idea behind the MCMC approach is to create a Markov chain able to generate draws from the posterior distribution of the model parameters. These Monte Carlo draws are then used to calculate statistics of interest such as parameter estimates and confidence intervals. Despite the idea behind MCMC methods is simple, its implementation requires the derivations of the appropriate (conditional) distributions in order to produce the draws. Many tools exist to efficiently implement MCMC strategies. The Hierarchical Bayes logit model is implemented in the R environment (www.r-project.org) package called bayesM

The model has been built to estimate the risk to have an increase in terms of HbA1c >0.5% within a time frame of 12 months starting at each visit. Demographical and behavioral characteristics (Gender, Age at follow-up visit, Time between follow-up visit and Time from T2DM diagnosis, Smoking habit), clinical measurements (HbA1c, BMI, Triglycerides, Systolic Blood Pressure (SBP), Total Cholesterol) and presence of pharmacological treatments (Insulin, other Anti-Diabetic treatments) were considered as independent variables to be included in the analysis.

In order to have a minimum number of data to learn individual models, data deriving from the first three visits available for each patient were used to train the model. These models are iteratively trained on the first set of visit and then tested on the consecutive one, while continuing to analyze consecutive follow-ups in time. At the first iteration, data of the 4th visit are used as test set, and

the event to predict was the occurrence of an increase HbA1c of at least 0.5% with respect to HbA1c % measured at 3rd visit in a time-frame of 12 months. At the second iteration, measurements deriving from the 4th visit were included in the training set and the prediction was performed on data from the 5th interval and so on until six iterations were completed (at the last iteration, predictions were made on visit number 9).

The extracted models allowed estimating population and individual-level regression coefficients for each time-varying covariate included in the analysis (age at follow-up visit, time between follow-up visit and T2DM diagnosis, HbA1c, BMI, triglycerides, SBP, total cholesterol, insulin, other treatments) conditioned on the considered static variables (smoking habits, gender). Table 22 shows the estimated population regression coefficients. Each column represents a ***static*** variable, each row a ***time-varying*** variable discretized using the 33th and 66th percentiles of the distribution. Both ***time-varying*** and ***static*** variables are binary. We used dummy variables to represent the smoking habit, which had originally three values (Yes, Ex, No).

Variable	*Level*	*Intercept*	Gender (M)	Smoke (Yes)	Smoke (Ex)
Age	[61-70)	-0.77	1.74	-2.56	-0.12
	≥ 70	-0.98	2.10	-1.36	-0.56
Time from T2DM diagnosis (years)	[4.32-10.9)	1.04	0.39	2.46	1.00
	≥ 10.9	2.25	-1.68	2.18	2.05
HbA1c %	[47.5-56.1)	-2.70	-0.93	0.64	2.86
	≥ 56.1	-4.15	0.22	-1.29	0.99
BMI (kg/m^2)	(26.7,31]	-0.68	-1.32	1.17	0.30
	≥ 31	-0.72	-0.20	1.25	0.63
Systolic Blood Pressure (mm Hg)	(127,140]	-0.09	0.07	-0.82	-0.98
	≥ 140	0.08	-1.07	-0.46	-1.31
Total Cholesterol (mg/dl)	(170,196]	-0.89	0.41	-0.25	-0.28
	≥ 196	-1.12	-0.46	-0.10	0.50
Triglycerides (mg/dl)	(103,140]	-0.25	0.75	1.79	0.79
	≥ 140	0.00	0.58	0.54	-0.87

Insulin	Yes	1.37	-1.97	-0.42	-2.55
Other drugs	Yes	-0.91	-0.24	-0.20	0.24

Table 22 – Population parameters estimated by the Hierarchical Logistic Regression model

Individual-level regression coefficients learned on the training sets were used to estimate the probability of experiencing an increase in terms of HbA1c % ≥ 0.5 within a time frame of 12 months from each visit. The regression coefficients reported in the previous table can be interpreted as follows: assume we want to estimate the impact of having triglycerides ≥ 140 mg/dl on the probability of a clinically relevant increase of HbA1c for a male smoker subject. The OR corresponding to triglycerides ≥ 140 in male subjects is exp(0.58) = 1.79, while it corresponds to exp (0.54) = 1.71 in smokers. The probability of an increase of HbA1c % ≥ 0.5 given the fact that an individual has triglycerides ≥ 140 mg/dl can be therefore computed by combining the regression coefficients (Beta) as follows: log-odds = Intercept (0) + Beta male (0.58) + Beta smoker (0.54) = 1.12; thus, the log-odds can be converted into probability by the ratio exp(1.12) / (1 + exp(1.12)) = 0.75. The OR for having triglycerides ≥ 140 mg/dl in this subgroup of patients is therefore exp(1.12) = 3.06, showing a high risk of increased HbA1c in the next year.

Logistic regression was also applied, as a term of comparison. When applied to the problem of predicting variations in HbA1c between two consecutive visits of T2DM patients, the Hierarchical model proved to outperform standard logistic regression. Results show that the MCC reached by Hierarchical Bayesian Model was always significantly higher than the MCC obtained by logistic regression on the same data (t-test on MCC values, p-value <0.05), except for the prediction after the 9th visit.

6.4 Integration of environmental data, exposure factors associated with HbA1c control

The methods and the results described in the following have been published in (Dagliati et al. 2015).

HbA1c and air pollution are both time-dependent variables. On the one hand, chronic patients are likely to evolve through several disease complexity levels, on the other hand, pollution levels may change due to land transformation, like building new residential areas or changes in industrial strategies. Thanks to the exploitation of time series analysis and satellite images processing, it was possible to study whether metabolic control showed seasonal variations and assessed if these had a spatiotemporal correlation with air pollution. The Pavia data was used to improve the studies currently available in the literature (Rajagopalan & Brook 2012; Thiering & Heinrich 2015; Janghorbani & Momeni 2014; Chuang et al. 2011; Tamayo et al. 2014; Park et al. 2015) by:

- Retrieving more evidence from longitudinal studies, taking into account temporal aspects and fluctuations in chronic populations followed for several years during disease arise and progression
- Defining finer-scale models of air pollution

The Pavia data was used to analyze satellites images, moves along the direction of improving the quality of clinical and environmental data analysis. Heterogeneous and diverse dimensionalities of the data streams were explored, in order to use them in a common analysis framework.

6.4.1 Description of the analysis and results.

The objective of the analysis was twofold: first, it was examined whether HbA1c levels of the studied population showed seasonal fluctuations and, in the case these variations were significant, if they showed a relation to air pollution measurements (air quality maps) as derived from satellites data. In order to achieve air quality maps from satellite, the data acquired by Landsat L8 mission have been considered. Pollution plays a key-role in the thermal pattern of a remotely sensed scene. The implemented spectral analysis took into account the overall pattern of the raw counts thermal signal as a function of black particulate concentration. The specific aim was to detect local relationship in Hab1c and pollution variation, which were not influenced by other well-known seasonal factors. The next section illustrates the performed analysis steps.

1) Spatio-temporal Scales Identification. The first step in the analysis was to define the level of detail through which derive meaningful patterns and observe events of interest. To this end, the temporal and spatial aspects of clinical and satellite data were evaluated to find the scales that better balance coarse data (over the whole geographic area; for the entire observation period) and fine data (a pixel value; a single HbA1c measure taken in a certain day). The analysis was based on seasonal variations (temporal) within counties (spatial).

2) HbA1c Longitudinal Analysis. Once geographic boundaries and temporal scales were defined, the average HbA1c values in each county was calculated, so to assess glycemic control seasonal variations in each year. Seasonal Hab1c levels of the nine geographic areas were defined for each year from 2009 to 2014.To retrieve significant fluctuations of HbA1c, considered as relevant differences in Hab1c values among seasons, and select the patterns to further investigate, we performed a mixed effect analysis (implemented in the R package; http://cran.r-project.org/web/packages/lme4/index.html). The applied methods allowed selecting significant (P value < .05) HbA1c seasonal geo-localized variations during the observation period. Table 23 shows the counties and years were there were significant HbA1c fluctuations among seasons.

District	County	Year	*P* value
Pavese	Certosa	2011	.00088
Pavese	Certosa	2012	.00034
Pavese	Pavia	2011	.00005
Pavese	Pavia	2012	.00009
Lomellina	Garlasco	2011	.00046
Oltrepo	Casteggio	2012	.00921

Table 23 - Counties and years where we found significant HbA1c fluctuations among seasons

3) Time series decomposition. To separate local effect in Hab1c variations, which were the effects to study, from other general HbA1c seasonal components, the additive model of seasonal decomposition procedure was applied. From time series of the whole studied area, monthly variations of HbA1c profiles were extracted. The extraction of seasonal adjustment factors allowed identifying months with peaks. Consequently, original time series of each county were adjusted and smoothed by removing these factors.

4) Air Pollution Satellites Maps. Air quality maps were derived from Landsat 8 satellite, which was launched in February of 2013. The satellite collects images of the Earth with a 16-day repeat cycle, referenced to the Worldwide Reference System-2. Thanks to the assumption that pollution plays a key-role in the thermal pattern of a remotely sensed scene, the correlation between the presence of black particulate and the recorded temperature can be thoroughly characterized. In order to match the region of the remotely sensed data with the Pavia area, spectral analysis was implemented while taking into account the overall pattern of the raw counts thermal signal as a function of black particulate concentration. For each pixel in every temporal series, a polynomial fitting model has been implemented to estimate the air quality of the scene. Air quality has been quantized on five levels over the black particulate concentration estimate. Figure 47 shows that air pollution levels are highly correlated to the seasons.

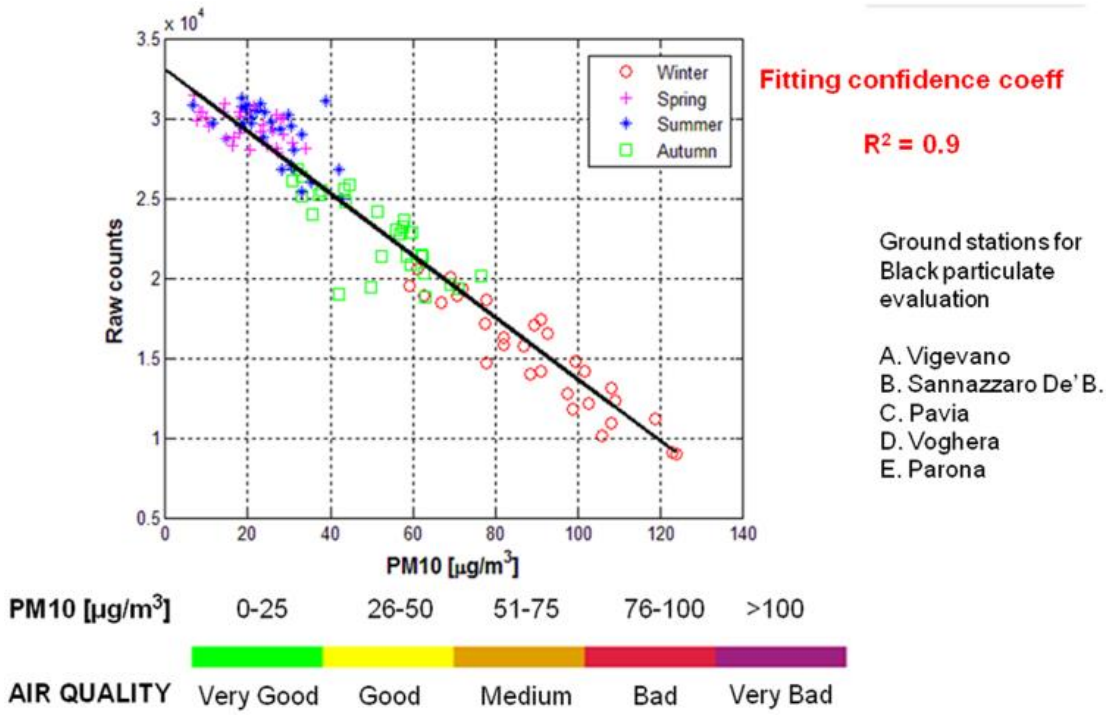


Figure 47 Air quality estimation in different seasons

1. Correlation between HbA1c and Air Pollution Variations.

The core phase of the analysis was devoted to merge and compare metabolic control and pollution values in each area and period of interest to detect possible correlations between HbA1c variations and changes in air quality.

As already mentioned, we wanted to evaluate if, in specific counties, notable seasonal fluctuations of HbA1c can be correlated with the air pollution pattern in that year. Table 24 shows Air Pollution values in the counties.

District	County	Mean HbA1c values (mmol/mol)	Mean air pollution values (PM10 □ g/m³)
Pavese	Certosa	53.26	25.17
Pavese	Corteolona	55.11	24.83
Pavese	Pavia	53.75	25.13
Lomellina	Garlasco	56.70	27.25
Lomellina	Mortara	50.93	23.78
Lomellina	Vigevano	62.66	20.77
Oltrepo	Broni	60.30	25.43
Oltrepo	Casteggio	54.56	25.71
Oltrepo	Voghera	57.70	20.91

Table 24 - HbA1c and air pollution values in different counties. Mean Hba1c and pollution values are computed for the whole observation period.

While detecting correlations between HbA1c and pollution trends, associations in all counties were Hab1c fluctuation were previously reported to be significant were founded. The analysis was deepened on the Pavia County, which is the one with the largest number of patients and the most homogenous geography. Table 25 shows the P values of the HbA1c patterns calculated through the mixed effect model test for each year as well as the correlation between the HbA1c mean seasonal values and seasonal air pollution for the Pavia area.

	Year	HbA1c patterns	HbA1c air quality correlation
Pavia	2009	n.s.	–.029
Pavia	2010	<.05	.66
Pavia	2011	<<.01	.94
Pavia	2012	<<.01	.81
Pavia	2013	<.01	.83
Pavia	2014	ns	.37

Table 25 - Correlation between HbA1c and air pollution in the Pavia county

As an example, Figure 48 shows that in the Pavia County in 2011 HbA1c and air pollution values followed the same trend:

- HbA1c, compared with the yearly mean value (54.05 mmol/mol), shows higher values during winter, lower values in spring and fall, and slightly higher values in the summer.

- Air pollution (on a 0-50 scale), compared with the yearly mean value (27.5 mmol/mol), shows the same trend.

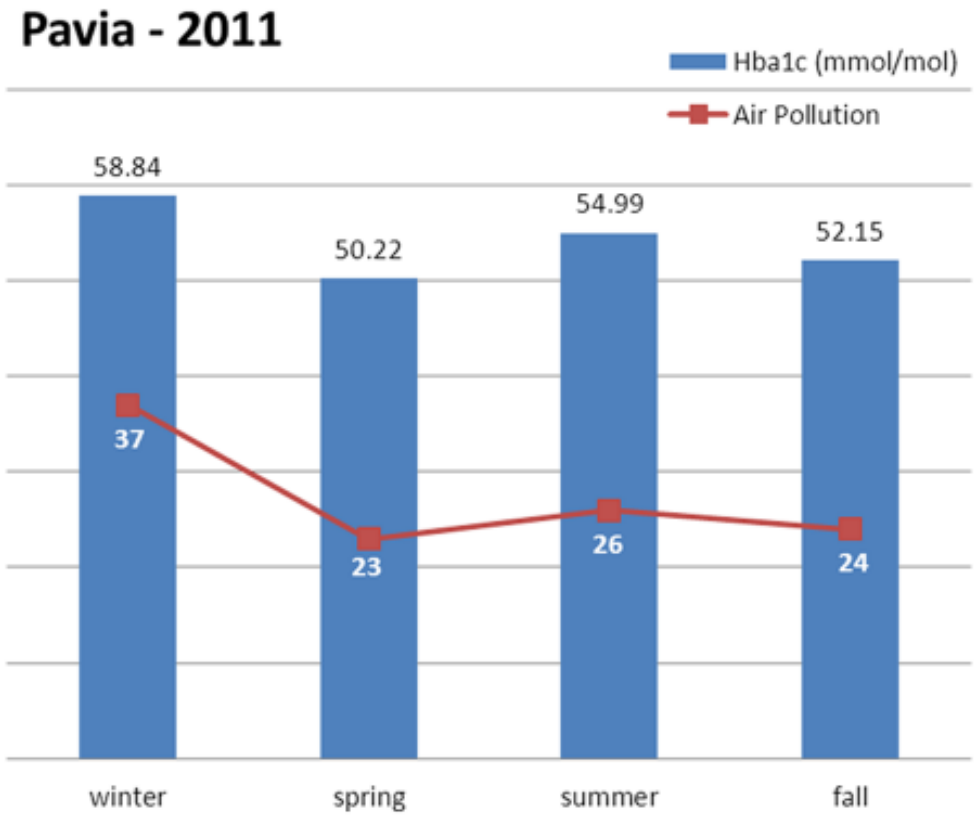


Figure 48 HbA1c and air pollution in the Pavia county in 2011

It is of course important to highlight that the presented analysis is a proof of concept. In particular, it has been shown that, thanks to data availability and big data technologies, it is now possible to jointly study heterogeneous data, such as healthcare and air pollution information extracted from satellites. This provides an unprecedented opportunity to improve our understanding of phenomena by extracting unseen temporal and spatial correlations.

CHAPTER 7

7 The clinical decision support system (Aim3)

To introduce the design and implementation of the clinical decision support system (CDSS), several considerations have to be recalled, in order to link the last Aim with the results achieved within Aim 1 and Aim 2 (Figure 49). Treatment and management of T2DM often takes place outside clinical settings and impacts on daily life of patients. As a consequence, clinicians depend on patient reports of symptoms, side effects, functional status and treatment adherence. Patients typically report at clinical visits that are months apart, and recall accuracy can be highly fluctuating.

The possibility to gather and analyze information from different sources, such as public healthcare systems, open data repositories or hospital information systems is necessary to perform valuable clinical decisions (Barbour et al. 2013). The ***data model*** defined and implemented to accomplish the objectives of ***Aim 1*** responds to this issue.

The activities performed in ***Aim 2*** have the objective of providing clinicians, health managers and researchers with a set of models able to predict the onset of complications and stratify the population on the basis of the temporal behavior of a set of variables significant to describe the evolution of the disease. More specifically, Aim 2 resulted in (i) the development of the careflow mining algorithm and the implementation of a Time Series Abstractor, which have to be integrated into the CDSS architecture as ***temporal data mining*** modules, (ii) the development and validation of a set of ***risk prediction models*** for microvascular complications and (iii) the detection of ***drug purchase patterns*** through pharmaco-epidemiology metrics.

The last ***objective*** of this work was focused on delivering the results of the analyses performed and methods developed to the final users. The developed models and solutions have been integrated into a software tool, the MOSAIC system, to support medical decision and clinical practice during management and control of the evolution of T2DM.

Although created with a continuous discussion with the MOSAIC project partners, the research design, the applied methods and the produced results of Aim 1 and Aim 2 are ***specific results of this research program***. The software solutions, in particular the final user interface, implemented to reach the goals of Aim 3 are the results of a more inclusive collaboration with the other project partners, while the main focus of the research program was to desing the best solution for the integration of developed algorithms and models in the tool.

The development of the CDSS completes the Learning Healthcare System cycle by enabling the reintroduction of research findings into care to better inform clinicians about patients' behavior and guide decision making. The algorithms to support the identification of custom careflows, the longitudinal data analytics, and the data aggregation strategies to support their effective visualization, made possible to turn the acquired knowledge into effective healthcare actions.

The focus of this chapter is to explain how the developed data model and analysis tools have been

integrated into the proposed CDSS system. The developed algorithms and models have been implemented as ***modular software tools*** to be integrated in the final system, which has been structured as a ***Dashboard***. The described activities concern the design of the application and the definition of the best solutions to provide an instrument that could be as effective as possible in supporting the management of T2DM patients. The technological choices behind the integration are briefly presented, as well as how the data model impacted on the integration process.

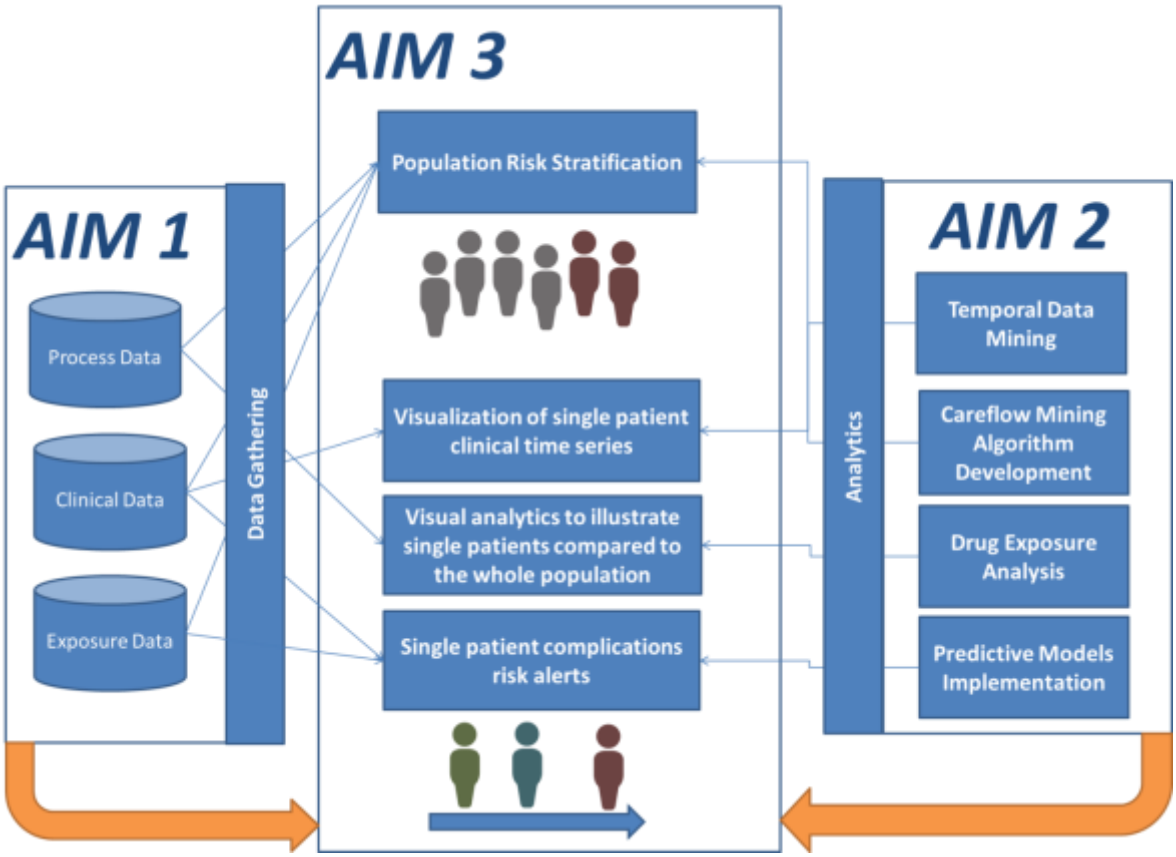


Figure 49 Aims links and Aim 3 sub tasks

7.1 System design and implementation

To design and evaluate the set of decision support tools to be implemented on top of the developed data model and methods, an evidence-based holistic framework, the CeHRes roadmap (van Gemert-Pijnen et al. 2011; Van Velsen et al. 2013) has been adapted and contextualized to design the CDSS. This approach has been suggested and the process has been structured with the collaboration of the Universidad Politécnica de Madrid partners.

Thanks to this approach, it was possible to understand user needs, behaviors and interactions with the developed solutions. The obtained results have given new insights in the definition of effective CDSS to tackle the issues correlated to the epidemics of T2DM. The process was able to finalize the development and implementation of the CDSS prototype and involved end-users and stakeholders in the whole analysis. This approach guided the definition of the operative solutions for developing the CDSS using data mining techniques on top of existing longitudinal electronic health records.

During the design phase of the CDSS two scenarios were identified, depending by the final users of the system and current needs in the T2DM management. Nowadays there are plenty of guidelines, risk models and recommendations on how to manage T2DM patients. The challenge was to produce reliable, meaningful and easy to access information that promote a better coordination between primary care, specialist medical doctors, healthcare managers of hospitals and local healthcare agencies.

The developed CDSS implements several functionalities within two use cases. One use case is devoted to the consolidation of clinical decisions during ***patients follow up***, providing support at an individual patient level. It has been designed for medical doctors, allowing them to better specify their risk profiles and behaviors of individual patients. While the single patient use case can be associated to more classic approach for delivering support during encounters and it can be seen as a tool built on top of actions routinely performed in patients' care (follow up visits), the other use case tackles a more ambitious challenge. This second use case is devoted to introduce a new approach in T2DM ***population management***, with the objective of giving a more complete vision on T2DM patients care and segmenting the population across center specific, temporal phenotypes.

In both use cases the decision support is delivered through a user interface, structured as dashboard, which implements the analysis layer, based upon the temporal data mining modules, through visual analytics solutions.

Within the use case devoted to deliver decision support for ***clinicians***, several strategies were developed to assist the care of a patient treated in a specific clinical context on the basis of his/her temporal clinical history. This solution supports clinicians during follow up visits in findings risk of developing complications or worsening conditions, having a whole picture of the patient clinical history and disease evolution in time. Specific visual analytics solutions implement the results of ***Risk Prediction models, Temporal Data Mining algorithms and Drug Exposure patterns*** detection methods, to assist medical doctors in the actions to be performed during encounters (e.g. changes in therapies).

The second use case allows to assess the risk of developing complications within patients' sub cohorts characterized by similar temporal patterns (temporal phenotypes). It implements a top-down analysis solution for ***decision makers***, who can identify subgroups with similar health care trajectories, detect the most critical pathways in terms of severity and use of resources, and plan suitable clinical and organizational actions. The use case leverages on the ***Careflow Mining algorithm*** to segment the population, recognize phenotypes on process data and compare their clinical characteristics in terms of complications. The implemented solution supports in a graphical way the stratification of patients with a drill down approach. Healthcare managers or head of hospital medical divisions are provided with specific criteria for a more efficient management of the center population (e.g. identify cohorts with higher frequency of complication).

7.1.1 System architecture and technical solutions

The four architecture components involved in the implementation of the use cases are the following:

- **i2b2 DW (Data Storage Module).** As described in details in Data Gathering and Integration (Aim 1) chapter, this component collects heterogeneous data coming from the hospital EMRs, administrative data from local healthcare agencies, and environmental data from regional databases.
- **DB Query Engine.** This is a Big Data oriented Java backend service that provides a logical layer between the users and the data. It is a backend query module that retrieves data from the i2b2 DW and sends results back to the GUI.
- **Data Mining Module.** This module implements all the developed analytics algorithms through which it is possible to retrieve meaningful patterns in patients' follow-up and the distribution of diabetes-related complications in specific groups. Due to the large quantity of data needed to run this module, some performance lacks have been found during the development phase (responses taking over 10 seconds), therefore, another no-SQL storage module has been introduced to act as a cache for this specific sub-module.
- **Interface (GUI).** This component allows the interaction between end users and the whole system. This interaction allows stratifying the chronic population and showing the results of the different mining modules performed on the selected subset of patients.

In the MOSAIC system the Data Mining Module components have been developed using R and Matlab. The communication between the DB Query Engine and the Data Mining Module layers occurs through messages exchange in JSON format. The GUI is a software component that stands as the interface of the user towards the system. The main purpose of this component is to carry out a viewer-controller model by splitting the core functionality from the viewer. The technologies exploited to develop the GUI are based on JavaScript, alongside HTML and CSS. Communication between modules occurs exchanging data in JSON (JavaScript Object Notation) format. The technology used to create all the charts in the GUI is provided by Google Charts.

7.2 Patient Use Case

This use case has been implemented to support clinicians in better managing T2DM chronic patients during follow ups visits. Starting from patient's EMR records, the system retrieves heterogeneous data from the DW and displays relevant clinical information and behavioral patterns in a graphic way. Clinicians are provided with a more exhaustive picture of what happened to patients while exchanging information with them during periodic visits and are able to better understand what needs to be improved in the patients' treatments.

The integration of information coming from billing data stream with hospital EMRs clinical data, the joint analysis of these sources of information and the implementation of graphic methods to represent longitudinal and sparse data, are the core tasks performed in building this use case.

The importance of ***data integration*** for a complete and coordinate care of T2DM patients is evident in the case of the implementation of the CDSS at the FSM Pavia hospital. In fact, the clinical information related to the period s from the onset to the first encounter at the hospital is missing. Usually, the longer is this period, the more information are unavailable or incomplete. It is evident that this is a limitation, as the events that might be early determinants of a complication are missing. The clinical and process data integration helps in tackling this issue. The secondary

reuse of public healthcare data provides structured information related to a previously unknown, or partially known, time window.

Another fundamental component of this use case is the integration in the CDSS rules engine layer of the ***analysis methods and models*** developed to transform data into clinical relevant knowledge. For each patient, the detected temporal patterns are shown, so that his/her disease path can be compared with references care standards (e.g. having metabolic control in target for a certain period). Moreover, these patterns have the objective of inspect the patient's disease evolution taking into consideration his/her internal, individual, variability in terms of multiple exposure events he/she underwent.

Although several efforts have been done in the analyses step to process data to automatically extract concise and easy to understand information (e.g. the drug exposure patterns), the use of proper ***visual analytics*** solutions can further enhance the decision process showing this information in a way to facilitate assumptions and inferences of medical doctors.

The following paragraphs are shown the implementation in the patients use case of:

- Complications risk models
- Temporal Abstractions of clinical time series
- Drug Exposure Patterns

7.2.1 Complications Risk models

As illustrated in chapter 6, a set of models to predict the onset of microvascular complications were defined and validated. These models have been integrated into the CDSS system.

On the basis of validation results, it was chosen to integrate in the system the microvascular models based on Logistic Regression for a 5-year time horizon prediction. Since the clinicians are already used to exploit a calculator for computing the cardiovascular risk (CVR), this score was included as well. Both CVR and microvascular complications probabilities are computed via R scripts when the system is synchronized. Raw data are extracted from the i2b2 DW. These data are the input of the R scripts that implement the models and return risk prediction values. An Extraction Transformation and Loading (ETL) procedure stores risk prediction results in the i2b2 DW. \Patient_Data\Cardiovascular Risk\

To foster the acceptance and the effective use of risk prediction models in clinical practice, it is important for physicians to understand how these model works and easily interpret their results (Van Belle & Van Calster 2015). As for the information derived from drug exposure, the developed models have been integrated in the CDSS leveraging on specific ***visual analytics*** that summarizes the complication risk predictions and facilitate their interpretation. Predictive risk models have been exploited in the Patient use case in a "***Traffic Light***" section to indicate the risk of developing macro or micro vascular complication. Traffic lights shows the risk calculated at the last follow up. Arrows and equal symbols indicate risk variations (increasing, decreasing or stable) with respect to the previous follow up (Figure 50). In case a complication has already been diagnosed for the patient, a red traffic light is shown and the onset date is displayed (retinopathy and neuropathy in the example).

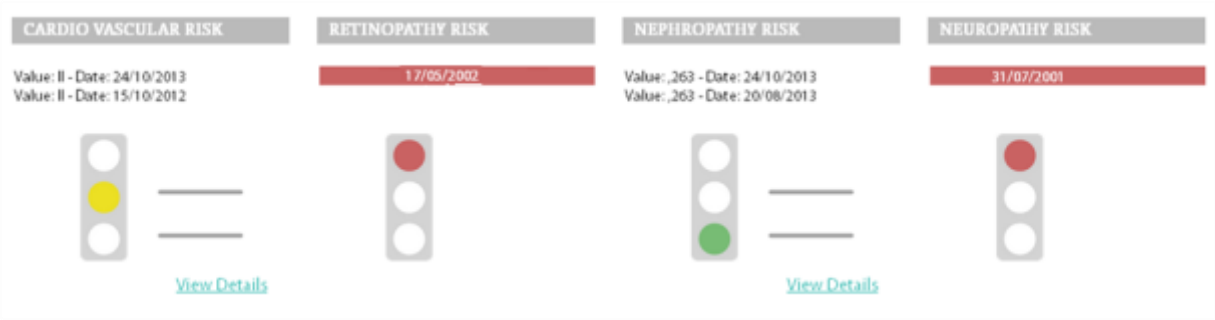


Figure 50 prediction models in the use case

7.2.2 Temporal Abstraction

Within the project, TAs are extracted by using a specific tool, JTSA (***Java Time Series Abstractor***), developed to perform this kind of task (Lucia Sacchi et al. 2015a). JTSA framework allows time series pre-processing and abstraction through a library of algorithms and an engine. The JTSA framework is grounded on a comprehensive ontology that models temporal data processing both from the data storage and the abstraction computation perspective. The JTSA framework is designed to allow users to build their own analysis workflows by combining different algorithms. Simple to highly complex patterns can be detected thanks to the modular structure of the system. The JTSA framework allows managing events time series (TS), represented by the raw data collected in the DW and episodes (abstractions) time series (A-TS), which are the output of any of the JTSA algorithms. The JTSA library includes several algorithms, that differ on the type of input and output data that are needed to execute them. These algorithms are listed in Table 26.

Algorithm	Input	Output
Pre processing	TS	TS
Basic	TS	A-TS
Aggregation	A-TS	A-TS
Complex	Pair of A-TS	A-TS

Table 26 – Input and output data for the TA detection algorithms available in JTSA

Within the MOSAIC project, TAs of interest have been defined thanks to a collaboration with the physicians, who provided the domain knowledge on the most important and clinically meaningful temporal patterns. This is reflected in the CDSS tool, where TAs can be visualized together with the clinical data of the patients. The defined abstractions are listed in the following:

- ***Abstractions on diet***. Regarding diet, physicians are interested in evaluating intervals where a patient shows good eating habits. The extraction of intervals of good eating habits is a Basic abstraction task. Diet is a categorical variable assuming "good" or "bad" values. The qualitative TS needs to be adapted to the suitable JTSA data structure as A-TS and, second, an Aggregation

TA algorithm merges consecutive "good" episodes to create clinically meaningful patterns. Besides intervals where eating habits have been good, we chose to show also the intervals where those habits have been bad, for completeness and clarity of visualization.

- ***Abstractions on weight***. As regards weight, physicians are interested in identifying both the time intervals where patients lose weight and the time that takes to the weight to reach a specific target. Finding intervals of decreasing weight requires the same workflow as the one for Diet, except for the detection of trends instead of states abstractions. The procedure that leads to the definition of the JTSA workflow to extract the time needed to reach a specific weight target (time-to-target) is a complex temporal abstraction task. The target weight loss is patient-specific and has been set to the 10% of the baseline patient's weight. The "TIME-TO-TARGET" pattern can be seen as the interval lasting from the first out-of-target value to the moment a patient reaches a specific weight target. The steps that JTSA performs o extract such pattern are the following:

  - Find "OUT-OF-TARGET" episodes
  - Find "IN-TARGET" episodes.
  - Find "OUT-OF-TARGET" episodes followed by "IN-TARGET" episodes
  - Extract the output "TIME-TO-TARGET" episodes.

- ***Abstractions on HbA1c***. As regard HbA1c, Clinicians were interested in investigating the amount of time that patients take to reach a specific target HbA1c level. The workflow components to extract the time taken to a patient to reach a specific HbA1c level are the same as the ones described for weight in the previous paragraph. In the case of HbA1c, the target level is defined as a value that does not depend on the baseline HbA1c value recorded for each patient.

Within the system, JTSA is run every time the i2b2 DW is updated and the results are stored in the DW through suitable ETL procedures A first ETL procedure extracts raw data from the i2b2 DW. These data are processed through the JTSA framework and uploaded, through a second ETL procedure, in the i2b2 DW.

TAs are used to show specific temporal patterns of clinical variables for each patient. In the CDSS system, the visualization of data for a single patient is performed as a step of the Patient Use Case. The results are presented in the "***Clinical data***" section, where the user can, for example, check the evolution of the patient Hba1c and compare it with the weight trend (Figure 51). Scatter plots show the quantitative measures made during follow-ups, whereas JTSA results are represented using time line plots.

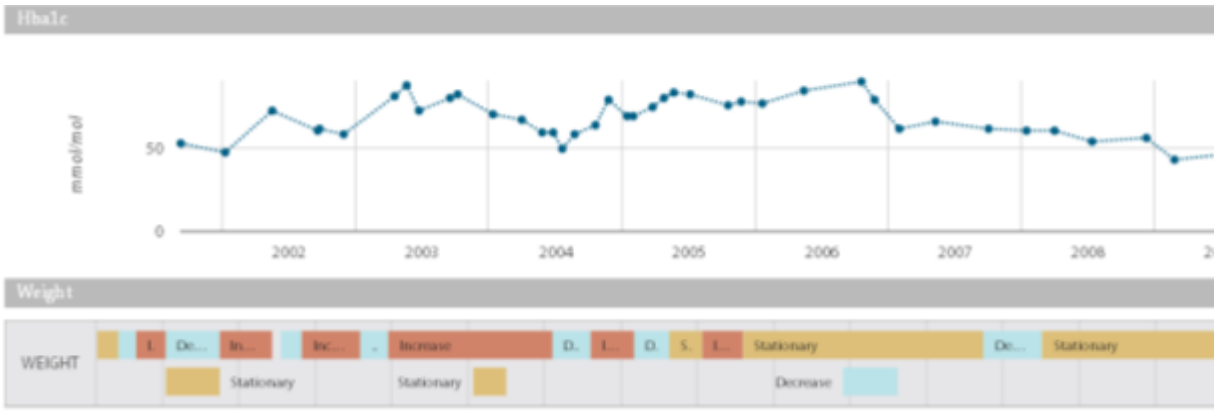


Figure 51 Hba1c time series and weight TAs, as calculated via JTSA module

7.2.3 Drug Exposure Patterns

Drug purchasing patterns can be extracted from administrative data recording the purchases made by patients in the territory pharmacies. As described in chapter 5.2, the indicator we have selected for representing purchasing patterns is the proportion of days covered (PDC). PDC can be calculated on the basis of the data stored in the i2b2 DW, both for active principles (identified by the ATC code) and for higher-level observations, like groups of drugs (the ATC Class). A first ETL procedure extracts drug purchases observations from i2b2 as input for a R Script that implements all the procedures described in chapter 5.2 (it calculates PDC, detects if subjects have statistically significant different purchase patterns from the population and finds patients tailored threshold for drug compliance in time). A second ETL procedure stores the R script results in the i2b2 DW. PDC values for each semester starting from the first purchase and a label indicating if the value of the PDC is under or over the patient specific computed threshold, are stored as observations. The p-values resulting from the Wilcoxon test are stored in JSON format in the Blob column of the observation fact table (Table 27).

Patient ID	ATC Code	Start Date	End Date	PDC Label	PDC Value	Blob
182	A10BA02	16-NOV-06	17-MAG-07	OVER	297	"grouping_atc": {"ddd_sem":"540","csa_obs":"296.703","csa_atc_med":"219.78","csa_atc_paz_med":"120.33"}, "grouping_class": {"atc_class":"***Metformin***","ddd_sem_atc_class":"296.703","csa_obs_atc_class":"296.703","csa_atc_class_paz_med":"219.78","csa_atc_class_med":"120.33", "p_value_atc_class":"0.112574","sign_med_atc_class":"-1"}
182	A10BA02	17-MAG-07	15-NOV-07	OVER	132	"grouping_atc": {"ddd_sem":"240","csa_obs":"131.868","csa_atc_med":"219.78","csa_atc_paz_med":"120.33"}, "grouping_class": {"atc_class":"***Metformin***","ddd_sem_atc_class":"131.868","csa_obs_atc_class":"131.868","csa_a

						tc_class_paz_med":"219.78","csa_atc_class_med":"120.33", "p_value_atc_class":"0.112574","sign_med_atc_class":"-1"}

Table 27 – Drug purchasing indicators concepts in i2b2

In the CDSS system, the "***Therapies section***" exploits the drug purchasing indicators to present several information about patient's drug exposure during the disease evolution. This is done through two sets of charts, that are displayed one below the other. The first set of charts (Figure 52) shows:

- DDD raw value (presented as time series in the scatter plot chart) associated to the purchases of specific active principles, also grouped by ATC Class;
- whether the patient purchases larger or smaller quantities of the drug, compared to other patients that are treated with the same drug (grey box below DDD charts). The arrow indicates that the patient purchased a larger (pointing up) or a smaller (pointing down) quantity compared to the population, the symbol indicates if the difference was or not statistically significant (red if the p value of the Wilcoxon test was < 0.05, green otherwise). This is the exploitation of the indicator in depth described in 5.2.2.

Figure 52 Visualization of drug purchases in the Patient Use case interface. a box including the comparison of the patients to all the other patients taking the same medications is also shown

Therapy purchasing behavior graphs indicate the PDC discretized values, with a semester granularity since the first purchase registered by the Local Health Care agency (Figure 53). The user can select which drug class analyze. For each semester, the represented time line indicates through colors and labels the compliance of the patients to a certain therapy, on the basis of specific thresholds computed on the basis of his/her behavior. This is the exploitation of the thresholds in depth described in 5.2.2.

Figure 53 Discretized PDC values calculated for each semester, as shown in the GUI

7.3 Population Use Case

In this use case, ***careflow mining*** (CFM) is exploited to understand which are the most interesting and informative variables to use within the novel perspectives offered by longitudinal data integration. In this use case, the Dashboard illustrates in a graphical way the stratification of patients following specific criteria enabling more efficient management of the center population (e.g. identify cohorts with higher frequency of complication). As widely discussed in chapter 5.1, the central efforts while developing a new careflow mining algorithm were dedicated to add a new ***temporal*** dimension in ***electronic phenotyping***. The developed algorithm mines careflows from heterogeneous records to identify different temporal phenotypes across the studied population. Once mined, careflows identify and allow selecting subpopulations that underwent specific sequences of events. Clusters of individual undergoing the same path can be identified and the clinical characteristics of these subpopulations described. For example, the retrieved groups of subjects may show differences in terms of patients' complexity, disease stage or resources utilization. This approach is aimed at providing healthcare stakeholders with detailed, but at the same time clear visualization of mined patterns, derived from process data, that can be compared and enriched with clinical data regarding the disease progression.

The integration of the novel CFM algorithm in the CDSS system provides a detailed visualization of mined patterns, which that can be compared and drilled-down to reach specific sub-cohorts or single patient details. As the central component of the ***drill down*** approach, the CFM algorithm extracts the most frequent careflows from i2b2 data logs. The studied population could be the whole patients' sample included in the center data repository or a subset of patients selected through demographic and clinical criteria available in the starting page of the dashboard (e.g. age class, gender, BMI defined classes to indicate patients overweight or obese).

The developed CFM algorithm detects the most frequent clinical patterns that are experienced by the selected patients' population, focused on taking explicitly into account the temporal dimension that strongly characterizes the evolution of T2DM chronic diseases. The mined careflows are used to stratify the patients' cohort in tailored groups on the basis of its dynamic characteristics. These subpopulations are able to identify group of patients with specific conditions or events relevant to describe specific medical cases. For example, the most critical pathways in terms of severity or use of hospital resources could be highlighted.

This is the step of the drill-down approach where the CFM algorithm is applied and its results are shown. To efficiently integrate the algorithm in the CDSS, it was necessary to find a solution that tackles some of the limitations of the algorithm, especially those derived by the high variability in T2DM patients' events. To overcome these issues, instead of using raw process events, clinical histories are mined on exposure and behavioral ***patterns*** of prescription-related drug purchases, frequent clinical temporal patterns, cardiovascular risk (CVR) profiles and level of complexity (LOC).

At this point of the process, the user has defined a set of temporal phenotypes, which are dynamically computed each time by the CFM algorithm. Therefore, the user can select a specific temporal pattern (which identify a phenotype) to drill down to the patients' population experiencing that behavior.

In the last step of the drill-down, the extracted temporal phenotypes are exploited to assess the risk of T2DM complications that possibly arise during the disease time course. For each of the so identified phenotype, the complications distributions are shown.

7.3.1 CFM Algorithm System Integration

The CFM algorithm is implemented in Matlab. The code has been made available as a standalone application using Matlab Compiler. The MOSAIC system runs the executable file passing the correct parameters and the set of patients and variables selected by the user. To communicate with the MOSAIC system, data are exchanged using a specific JSON format. The JSON result is stored in the field results and it is used to create the Google Charts that the user visualizes in the GUI. The following example shows the structure of the JSON that is taken as input by the CFM algorithm to shows Cardiovascular Risk (CVR) careflows.

JSON - Input


```
{ "concept": "CVR",
 "patients": [{"patient_num": 780,
                "observations": [      {"obs_label": "IV", "end_date": "2015-02-24", "value": 16, "start_date": "2013-09-25"},
                                       {"obs_label": "IV", "end_date": "2013-09-25", "value": 16, "start_date": "2013-09-16"},
                                       {"obs_label": "V", "end_date": "2013-09-16", "value": 21, "start_date": "2012-09-21"},
```


The JSON includes the name of the specific i2b2 concept (CVR in the example) and all the observations available for each patient (identified by the i2b2 ID). Each observation has a start date, an end date, a value (the raw value of the variable, if it exists), and a label (in this example, the discretized value of the CVR).

The output JSON is structured as in the following example:

JSON - Output


```
{"histories": [ {"label": "story_I_III ",
            "steps": [
            {"label": "I", "id": "event", "n_pts": 1,
                        "time": 239, "prctile25": 239, "prctile75": 239,
                        "min": 239, "max": 239,
                        "h": 0.000000,
                        "num_classes": 1.000000,
                        "patients": [{"idcod": 12487, "duration": 239}]},
            {"label": "III","id": "event","n_pts": 1,
                        "time": 804,  "prctile25": 804, prctile75": 804,
                        "min": 804,"max": 804,
                        "h": 0.000000,
                        "num_classes": 1.000000,
                        "patients": [{"idcod": 12487,"duration": 804}]
```

In the output JSON, all the extracted careflows (in this case named "histories") are listed. Each careflow is identified by its label, which is made up by the list of events characterizing the temporal history. For each step of the history the algorithm specifies:

- the name of the event characterizing the step (label),
- the number of patients included in that step (n_pts)
- the median (time), 25th (prctile25), 75th (prctile75), minimum and maximum duration of the step for all the patients verifying the history,
- some information for the construction of the events durations (h and num_classes)
- the list of the identifiers of the patients belonging to the group (idcod), together with the duration of the single step for each patient (duration).

Within the CDSS the CFM algorithm is applied on several variables, uploaded in the i2b2 instance while the system is synchronized. I2b2 aims at handling complex multivariate temporal data, which are gathered on the form of observations, and easily managed as input for the algorithm in the form of event logs. In this implementation, the CFM algorithm has been applied on i2b2 concepts related to:

- Administrative process data: Hospitalization and Day Hospitals, Drug Purchases
- Heterogeneous Clinical data: CVR values represented as time intervals
- Mixture of clinical and process data: LOC

Careflows	Variable – Events to build the Careflow	Origin	i2b2 Concepts
Complications	Arising of T2D related complications	Hospital EHR	..\Complications\
Hospitalizations	Hospitalization and Day Hospitals	Administrative Data from Local Health Care Agency	..\Hospitalization\Course\Day Hospital\ ..\Hospitalization\Course\In Hospital\
Drugs	Drug Purchases	Administrative Data from Local Health Care Agency	..\Drugs\Prescription\
CVR	CVR calculated through the "Progetto Cuore" Algorithm	Hospital EHR	..\Patient_Data\Cardiovascular Risk\

LOC	Level of Complexity	Hospital EHR + Administrative Data from Local Health Care Agency	..\Patient_Data\Level of Complexity\

Table 28 CFM and data stream input

7.3.2 CFM Results Visualization in the Drill-down approach.

In the following are described the results of the integration of the CFM algorithm in the CDSS system via the drill down approach and the choices to visually present the CFM results.

The first page of the Population Use case (in Figure 54), named "***Selection***", presents pie charts that show patient counts grouped by Demographic variables (gender and age class), BMI and Risk indexes at the last visit. The careflows are extracted when the user selects a group of patients in this starting page. Clicking on a chart section the CFM algorithm extracts the careflows associated to the selected group of patients (for example the patients currently in the age class 40-50 years). The user has also the possibility to run the algorithm on the entire population.

In this implantation the user cannot select the algorithm parameters (i.e. the support) but only the initial stratification of the population and, then, which temporal phenotypes to investigate. The decision of including or not the possibility to choose support parameter is still ongoing, as its main drawback is to add an additional complexity to the tool. Moreover, in other applications, the CFM support was chosen after several heuristic tests to balance the meaningfulness and readability of the results.

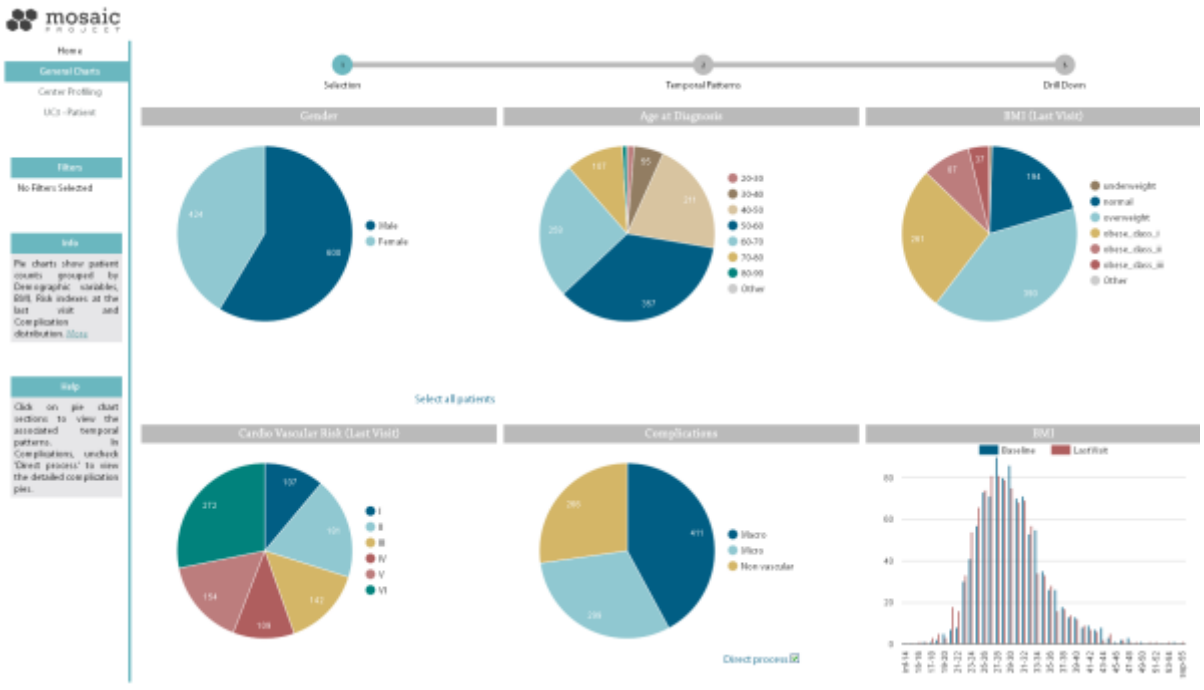


Figure 54 Starting page of the CDSS system

Figure 55, Figure 56 and Figure 57 show the most frequent careflows of medication purchases, cardiovascular risk and level of complexity events, extracted by the CFM and visualized as timelines in the "***Temporal Pattern***" page. Tooltips are used to display information on the histories duration, number of patients, and basic statistics. The selected patient sample is shown on the left in the Filter section (patients in the age class 40-50 in this example).

To facilitate the selection of specific sub cohorts, the extracted temporal patterns are presented to the user as ***timelines***, instead of directed acyclic graph. Timelines charts well fits the representation of time interval type of events. For this reason, within the CDSS interface, the events on which the careflows are mined are preprocessed to represent the time interval where a certain value holds, basically following the procedure described in 5.1.2.2. This procedure is straightforward for LOC events, which are already computed as time interval (e.g. the stable event is the period between the diagnosis and the first complication). Other events, like cardiovascular risk (CVR) profiles, are based on timestamp measures. In this case events are represented as the interval-based description of the period where a certain value remains the same. For example, if a patient has measures of "high CVR" at t0 and at t1, "low CVR" at t2 and t3, the CVR event will be "high CVR" [t0, t2], "low CVR" [t2, t3].

In timelines arcs are not represented, and the ***temporal enrichment*** (visualized in the tooltips) is performed only on events, which fill the representation in time of the whole patient history. This visualization of the CFM results was discussed and chosen with clinicians and usability experts as it simplifies careflows visualization and the detection of temporal phenotypes.

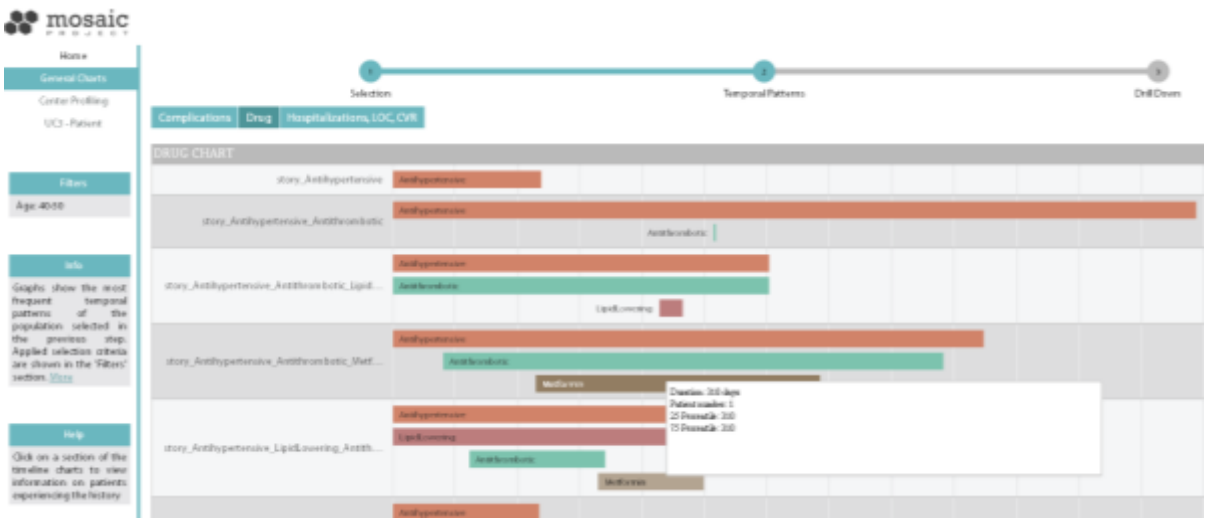


Figure 55 Drug purchases careflows are extracted from drug purchasing data stream. Careflows represent the most common exposure to groups of active principles since the T2DM diagnosis

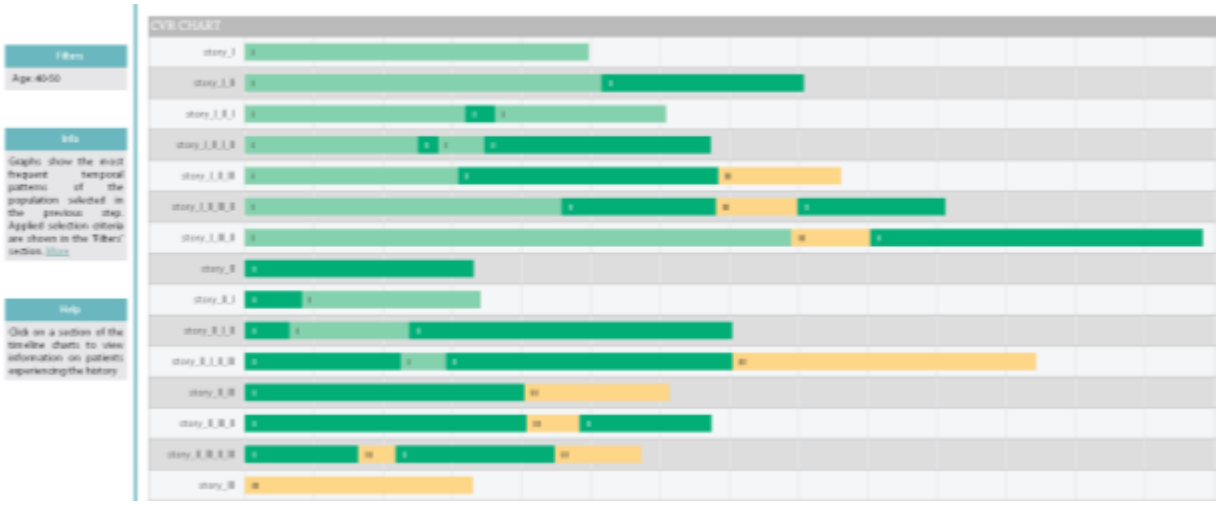


Figure 56 CVR careflows indicate the evolution of the CVR at 10 years, calculated with the 'Progetto Cuore' risk model. Careflows represent sequences of risk score intervals, stratified on the basis of fixed thresholds. From Risk I: less than 5% to Risk VI more than 30%.

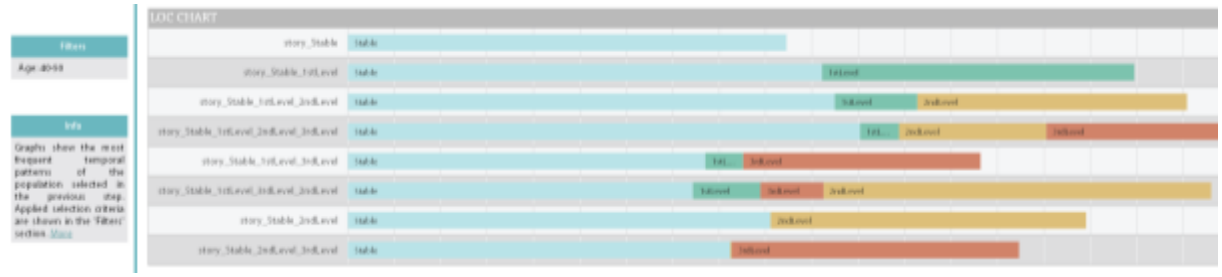


Figure 57 Level of Complexity careflows. LOC stays for Level of Complexity. LOC careflows represent the evolution of the disease from the diagnosis: Stable: no complication, 1stLevel: rise of the first complication, 2ndLevel: rise of multiple complications, 3rdLevel: hospitalization due to previous complication.

The last step of the process is triggered when the user selects a specific phenotype from one of the visualized careflows. For example, the user can choose to investigate one of the eight sub cohorts defined by the mining of LOC events (Figure 57), in particular those patients who went through all level complexity and are actually in the 3rd stage. The last page of the use case shows the "***Drill Down***" results as the complications distribution for the selected phenotype. Within this page the user can also select a specific complication and retrieve the list of the patients associated. This step is aimed at implementing the so called ***clinical enrichment*** of the careflow, showing the probability of the patients' complications given the belonging to the same phenotype. Although, in the CDSS implementation, the enrichment of the mined careflows with clinical information is performed on the whole history (without any temporal constraint) and, in this case, only with complications data, and not with other possibly relevant clinical variables, like Hba1c.

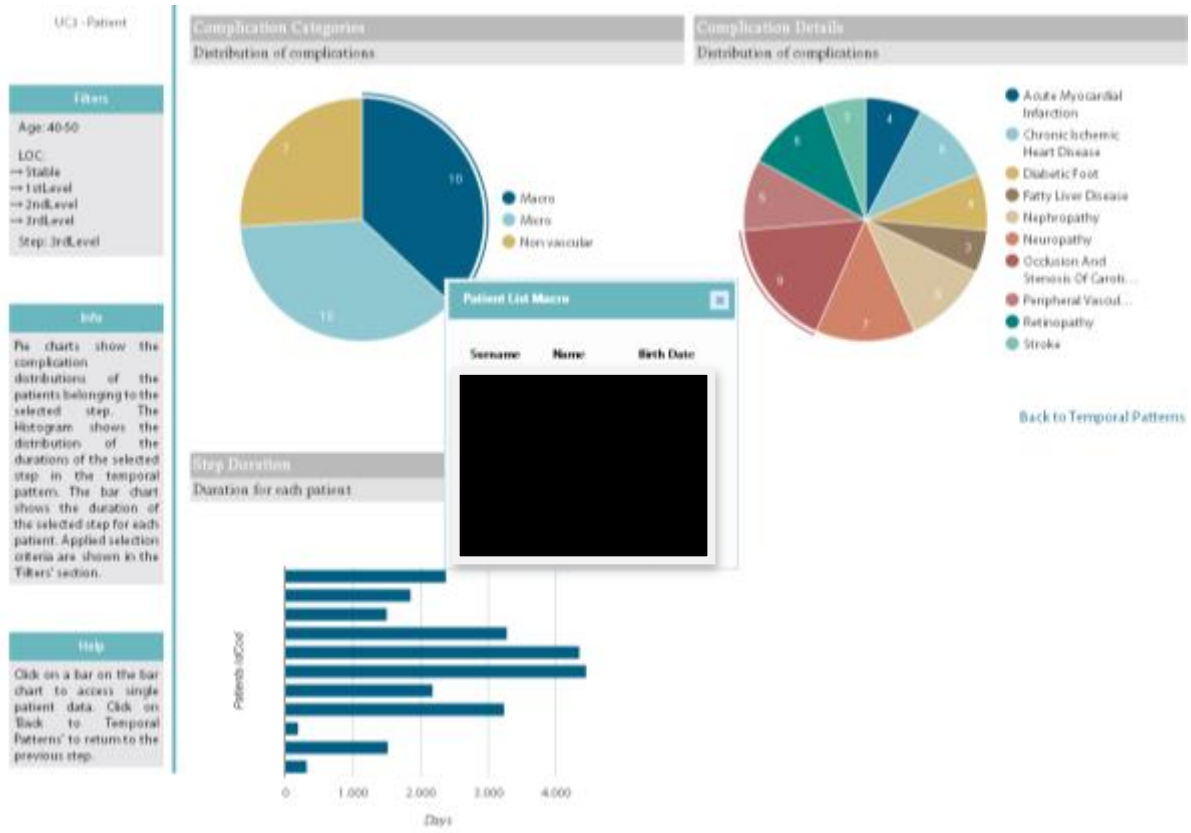


Figure 58 Complication distribution and patients selection

CHAPTER 8

8 System Evaluation

The developed CDSS was evaluated at the Pavia FSM hospital from both the use cases perspectives. The focus of this evaluation is on describing the most relevant information about user reactions, behaviors and needs.

The main tasks performed within this research program were dedicated to plan and manage the Pavia validation activities, and to perform the statistical analysis for the delivery of final results. The candidate also planned and write the IRB protocol, as required for the evaluation of the tool in the clinical practice.

The two use cases have different settings, employments and users. The implemented evaluation strategies, while following a general schema, required specific adaptations.

The Patient Use Case was evaluated following a pre-post approach to assess its impact on clinical activities. Clinicians were monitored for three months with and without the CDSS. In the first three months (between 15th of Sept – 1st of Dec 2015) 352 T2DM patients were visited without the tool; in the second period (between 1st of Dec – 30th of March 2016) 353 T2DM patients were visited near hospital with the CDSS dashboard. As the tool was used during medical practice, the study was approved by the competent Medical Ethical Committees of the FSM hospital.

The Population Use case evaluated the introduction of a new process that was not possible earlier for diabetes patient management, i.e. the comparison of a snapshot of the T2DM patient population at two different time points.

The profiles of the healthcare professionals who participated to the study are summarized in Table 29.

Gender	**Male = 3; Female = 6**
Age	40,6 ± 15
Years of Professional Experience	13 ± 12
Information Technology Literacy (Self-evaluation)	High = 2; Medium = 5; Low = 2

Table 29 Profile of the Users of the CDSS

8.1 Evaluation Results for the Patients Use Case

To perform a comparative evaluation between the pre and post evaluation phases, at each visit, T2DM specialist clinicians were asked to fill in a report with information related to the duration of the visit, the actions performed during the visit (including information about whether they referred the patient to further examinations and/or specialist visits), the changes in treatment (medications, physical activity, and diet), and the time to the next follow-up.

At the end of the evaluation period, the results obtained in the pre and post phases were compared. In the following, are reported the differences that were found significant by applying a chi-squared test.

Visit duration was found significantly lower when the CDSS was used. The number of visits lasting more than 25 minutes has been found to be significantly lower when the MOSAIC tool was used in clinical practice (p-value << 0.01, Figure 59). The developed CDSS gives the opportunity to have a snapshot of the current patient status the temporal evolution of the disease. The results suggest that this functionality allows reducing the visit duration, avoiding EHR consultation to retrieve the needed information, activity that can be time consuming especially for long follow-ups.

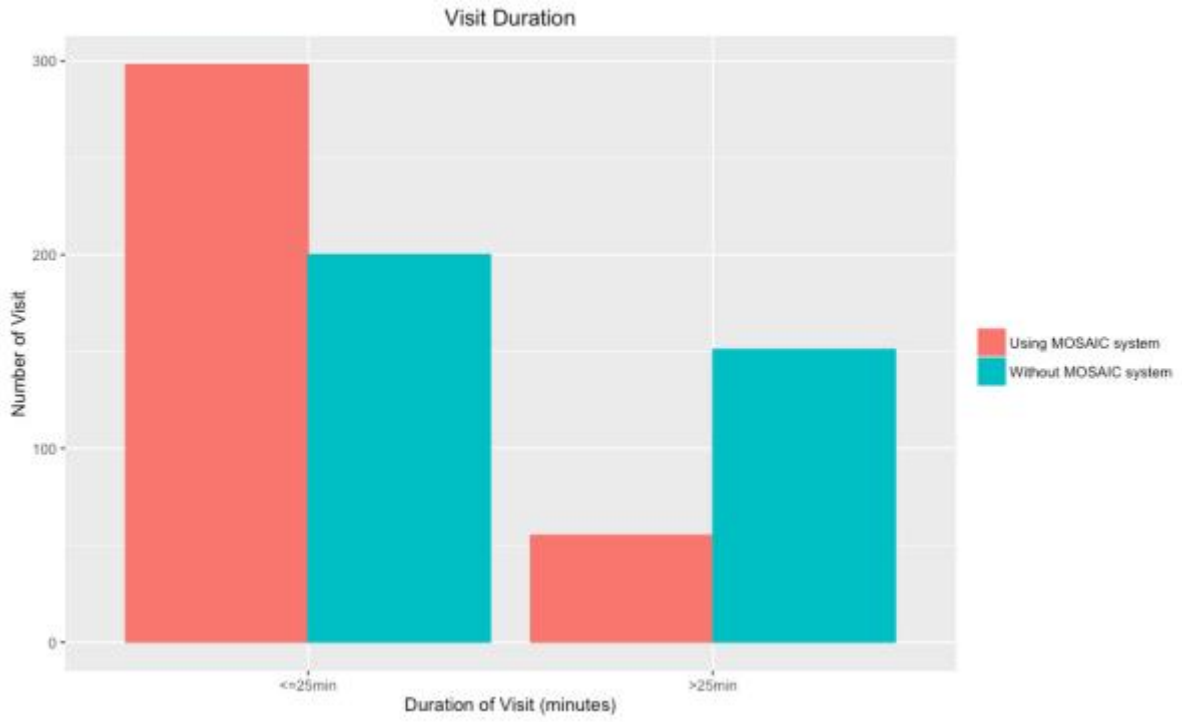


Figure 59 Duration of the visits with and without the CDSS.

Also, the ***interval between follow-up*** visits seemed to be more often longer than 6 months with the tool compared to that without (p-value = 0.07, Figure 60).

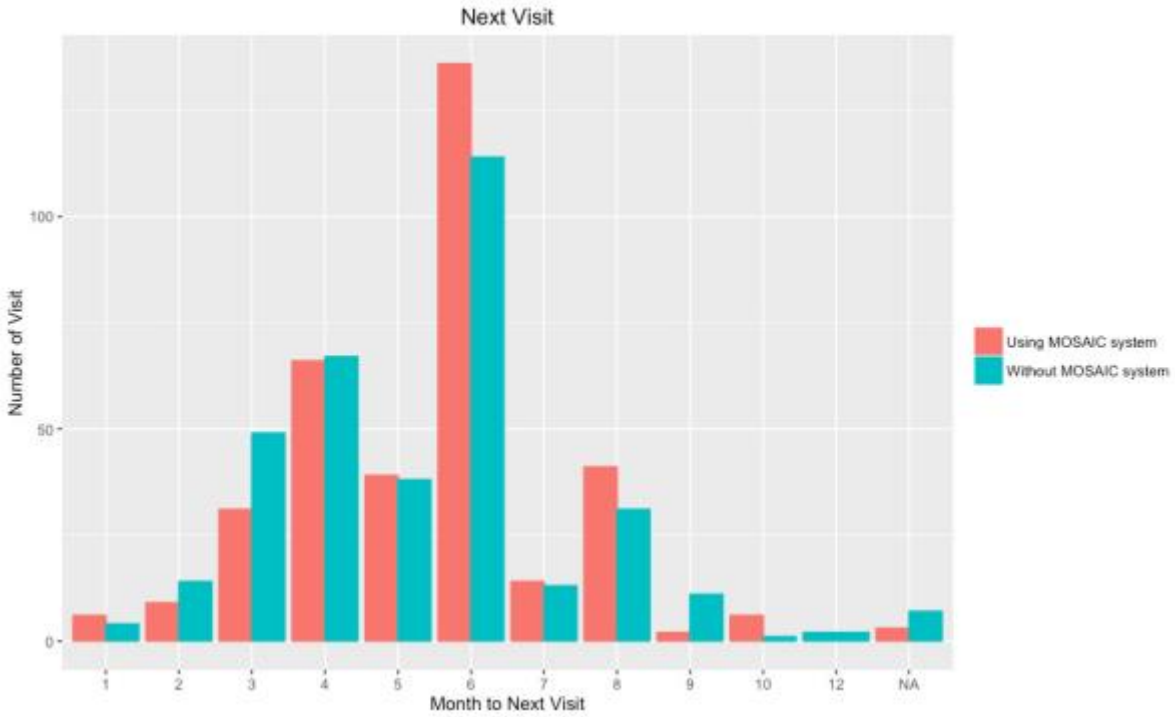


Figure 60 Time to the next visit with and without the CDSS.

With the use of the tool, the number of ***screening exams*** prescribed during visits performed with the tool is higher than that prescribed during visit performed without the tool (p-value < 0.01, Figure 61).

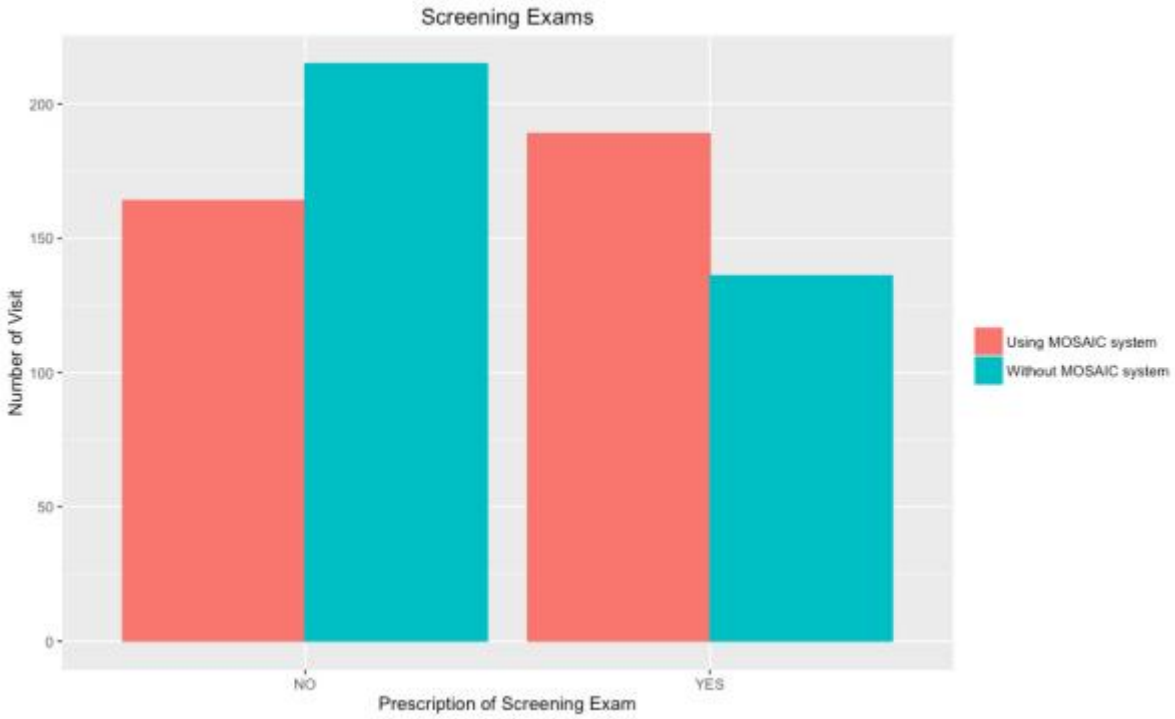


Figure 61 Screening exams performed with and without the CDSS.

Moreover, more interventions on the physical activity were prescribed during the visits performed with the tool (p-value = 0.05, Figure 62). This last result is particularly interesting, as the information on ***physical activity*** was often disregarded before the introduction of the MOSAIC

CDSS. Thanks to the traffic light screen that is automatically presented to the user when a patient is selected by the system (Figure 63), the information on physical activity can be immediately identified and considered by the clinicians.

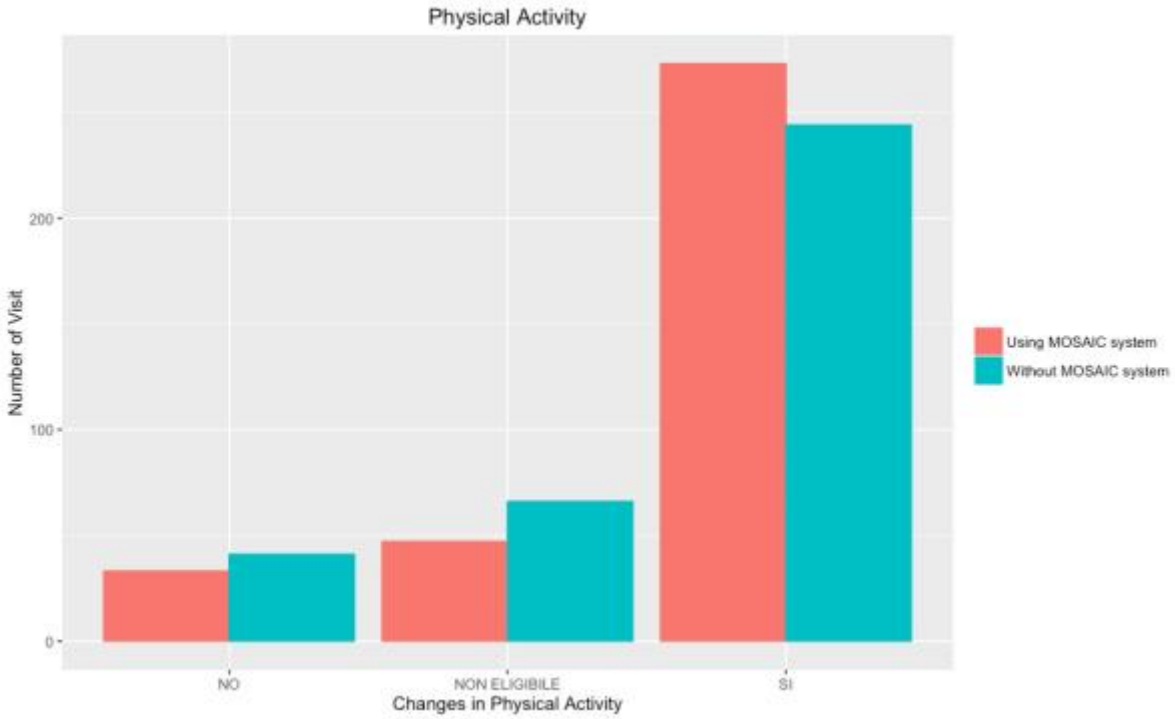


Figure 62 Physical activity interventions suggested with and without the CDSS.

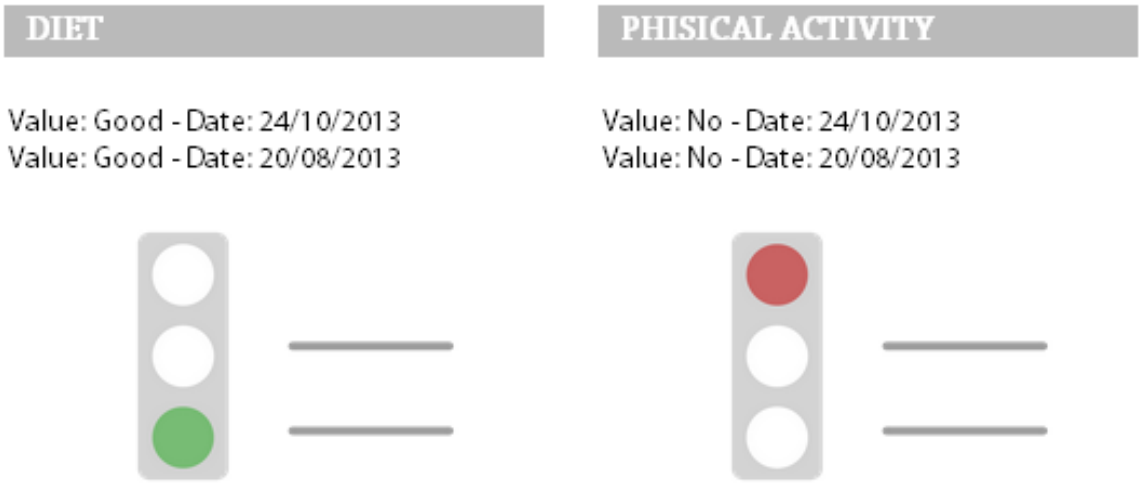


Figure 63 Traffic light view related to lifestyle data.

8.2 Evaluation Results for Population Use Case

To evaluate the Population Use Case, two evaluation sessions were performed involving three clinicians from the Pavia's FSM hospital and one health policy maker from the Pavia's Local Healthcare Agency. During these sessions, clinicians were asked to use the tool and compare snapshots of the center population across different time periods: at the beginning and at the end of the validation study (January 2016 and April 2016). The meetings were recorded. An analysis of

the recordings allowed the extraction of the most interesting results, which are reported in the following.

One of the most interesting results of these meetings was related to the interest of users in the possibility of using the CDSS functionalities to ***inspect specific clinical questions***.
One example of such questions was related to the analysis of patients who experienced a myocardial infarction (starting from their slice on the general charts page of the tool). According to clinical guidelines, such patients should all be treated with lipid lowering drugs, but the analysis of the treatment histories revealed that some of them didn't. The drill-down functionality allowed identifying those subjects and to inspect individual data using the patient use case, in order to find the reason for a possible noncompliance to the guideline. Another example question is related to the analysis of patients who take a specific combination of drugs. During the meetings, the users tried to answer this question for patients in the 50-60 years age group. In particular, the users tried to understand whether patients taking antihypertensive and antithrombotic drugs from the start of their treatment (13 patients) are more complicated than those who do not take the same treatment in the same age class. Also for these patients, the drill down to single patients was considered particularly interesting to motivate the behavior found in the population.

Another analysis that was performed by clinicians during the meetings was related to the ***level of complexity careflows***, which allows the analysis of ***disease progression.*** First, they considered the question of what distinguishes patients who remained stable from those who instead didn't. Then they considered in more detail patients who reach the highest level of complexity (those who underwent hospitalizations due to complications), to understand how much time passes between the second complication and a hospitalization. They found this functionality very important to understand the population of patients treated at the center.

A ***comparison*** of the center population ***between the start and the end of the validation*** period highlighted a significant increase in the number of patients with high cardiovascular risk (30.8% versus 28.2%), while there was an increase in the percentage of patients who are in the normal weight range (from 19.9% to 21%). The distribution of the other variables remained almost constant over time. This result stimulated clinicians to inspect in more detail the careflows mined from cardiovascular risk profiles, to understand which groups of patients have evolved to reach the maximum risk level. One hypothesis formulated by clinicians during the meeting was that it is possible that, using the CDSS, patients who seemed to be more critical were prescribed additional exams, which could confirm an increased risk level, which was then included in the data.

The last interesting observation that emerged during the meetings was related to the impact of the Population Use case on ***individual patient management***. In fact, when a patient comes to the visit, it is possible to use the Population Use case to see if he/she is behaving like the patients in his/her category or if a different behavior, possibly implying a different treatment strategy, can be noticed.

CHAPTER 9

9 Final Discussion

9.1 Accomplishments and contributions to the context of medical informatics

As stated in the introduction of this manuscript, the general research question of this dissertation was how to effectively perform longitudinal data analytics to find new insights and tools to improve management and care delivery of the T2DM disease. It is possible to revise the achieved results using as benchmark the data management lifecycle (Möller 2013; IBM 2013), also positioning this work in a practical context. This work started with the ***collection*** of secondary data and the implementation of suitable tools to collect this data from heterogeneous sources and to organize them in order to enable the comprehension of possible correlations between events in time. To facilitate this objective, data were ***described*** through an ad-hoc concepts and metadata, which considered also a temporal qualitative description of data. The core methodological part of this work can be identified in the development of novel mining methods and in the implementation of state of the art analytics algorithms to ***discover*** new insights in the collected database. Novel T2DM phenotypes were discovered from process data and disease complexity stages. These phenotypes are novel because, in the application context, they have never been defined and explored in terms of temporal sequences of events. Moreover, the mined careflows not only occur inside the healthcare facilities (i.e. the hospital), but they trace patients' behavior in a bigger picture, considering environmental information and their actions in everyday life (e.g. purchasing drugs). These new insights, or more precisely, the new methods to continuously assist professionals to find new insights on T2DM patients, were ***integrated*** in an informatics tool that facilitates the visualization of the collected information, and triggers novel inferences from both clinical and public healthcare management point of views. By the end of this work, the ***exploitation*** of the results can be performed in two ways. On the one hand, T2DM care professionals could use them as a support of their routine activities to ***plan*** novel actions from which gather new data. On the other hand, the clinical and administrative activities continue to feed the database with new information, on which advanced data mining techniques can be applied to continuously perform new research activities, also enabling the learning healthcare system cycle.

To define how informatics solutions empowers precision medicine, the learning healthcare system cycle leverages on the concepts of data, information, knowledge and it adds the fourth dimension of action (Tenenbaum et al. 2016). The aims of this dissertation have been positioned in this cycle through current key areas of medical informatics research. To detail the accomplishments of this work, the defined aims are linked to the obtained results following the information spectrum (Ackoff 1989).

The achievement of Aim 1 "Gather and integrate of data from heterogeneous sources" turned ***data*** into ***information***. The tasks performed to realize the ***i2b2 data warehouse*** were guided by

the understanding of functional data relations, defined as information, in a longitudinal and heterogeneous context. This work gives its contribution to the informatics' field, demonstrating the feasibility of integrating a broad range of data that characterized T2DM chronic patient histories into a common and sharable data model. The gathered data originates from different institutions, where they are collected for different purposes, at different time points and with different granularities.

The translation of ***information*** into ***knowledge*** is "conveyed by answers to how-to questions" (Ackoff 1989). The data analysis methods, developed and applied to achieve Aim 2, contribute to the discovery of novel hidden patterns in T2DM disease. The key research idea is that, in chronic diseases, these hidden patterns are embedded in the sequential order of events and, once discovered, they can be used to segment the population in temporal phenotypes. The main value of the reached objectives is represented by how effective the developed methods are to tackle the T2DM endemic problem. T2DM disease has been systematically studied in terms of longitudinal data and, leveraging on the available information, the most complete vision of the mechanisms underlying its evolution has been given. Each of the implemented data analytics methods addresses specific issues.

- The ***Careflow Mining algorithm*** detects the most frequent careflows from process events and enriches them with clinical data to define temporal phenotypes across the studied population;
- The mining of ***Drug Exposure patterns*** recognizes patients' behaviors taking into account population irregularities and patients fluctuations in drug purchases;
- The developed and validated ***Microvascular Complications Risk Models*** are profiled to take into consideration the characteristic of the studied cohort;
- ***Continuous Time Bayesian Networks*** allow understanding the evolution of a certain scenario and to detect the moments when it is more likely to have deviations from clinical control;
- ***Hierarchical Bayesian Regression Models*** allow estimating individual patients' parameters besides the usual population parameters, in order to deliver prediction models tailored on individuals' characteristics;
- The jointly study of ***Clinical and Remote Sensing data*** allow a new perspective on clinical events by extracting temporal and spatial correlations between metabolic control and environmental information.

The realization of a tool, which integrates an algorithm that dynamically detects temporal phenotypes, and multivariate longitudinal analytics, shows how the results of secondary use of data helps to support clinical decisions in chronic diseases. The achievement of Aim 3, that is the implementation of the CDSS, supports the definition of knowledge as the capability of understanding why the patterns, retrieved through analytics methods, impact on clinical outcomes. This definition fulfills the translation of the acquired ***knowledge*** into ***actions***.

The integrated models and algorithms enhance the CDSS to improve the characterization of T2DM patients and help to evaluate the risk of developing complications. The understanding of the mechanisms underlying the progression of diabetes is made through the analysis of individual

patient histories, temporal events and behavioral factors. The CDSS allows a better control of patient clinical condition over time, and personalized interventions.

The system is structured to provide two distinct solutions, formalized as Use Cases:

- The ***Patient Use Case*** helps clinicians in better managing their patients during follow-ups, improving coordination among different levels of care. Visual analytics solutions enable users to understand if certain actions (e.g. periodical tests, complications' screening or changes in therapies) are needed according to the level of complexity and behaviors of the patient.
- The ***Population Use Case*** allows to assess in a quantitative way the performance of a healthcare organization and to develop new strategies for stratifying patients treated in a specific clinical context on the basis of their temporal clinical history. The cohort segmentation reveals phenotypes based on the combination of health profiles of patients in the different progression stages of their disease, with the intrinsic needs of those disease's stages. The tool interface is structured in order to guide healthcare organization managers to understand the processes deployed by the facility: from the entire cohort, zooming to single patients. Users can detect nonconformities in the care processes. The tool allows the continuous optimization of clinical pathways, in terms of clinical outcomes and allocation of resources and costs.

9.2 Limitations and Possible improvements of the CDSS functionalities

The system evaluation assessed the initial hypothesis of this work, proving the effectiveness of the developed CDSS system to improve the management of T2DM chronic care. Nevertheless, the evaluation of the system has been performed in a specific context, whose peculiarities might have influenced the success of the project. The Fondazione S. Maugeri is a research hospital center; this fact might have triggered a better acceptance of a novel tool based on innovative methods in the clinical practice. The collaboration with the Local Health Care agency, which allowed to gather a large and useful amount of process data, was possible through an agreement signed within the MOSAIC project. To make this collaboration continuative further agreements have to be defined. In general, a possible drawback is that several functionalities (e.g. mining drug purchases behaviors) of the framework require a collaboration between healthcare institutions.

The system was evaluated to prove its capabilities in enhancing patients' management processes, but it was not possible to estimate its impact on patients' clinical conditions. This problem arises by the time pace of T2DM disease evolution, which is longer than the available evaluation period.

Some limitations of this works are related to data quality and sample size. Several efforts were spent to preprocess data and deal with issues like missing data, especially for the development and validation of the complications risk models. The use of advanced longitudinal models, like the Hierarchical Bayesian Logistic Regression models for metabolic variations, needed to define strict inclusion criteria, which results in a reduction of the sample size that reduced the statistical power for the methods validation.

One of the main limit of the collected data set was the missed opportunity to integrate self-monitored data through smart devices. This limit was partially due to the population demographic,

mainly composed by old patients not very prone to use these kinds of devices. Moreover, the project funds, as allocated by the consortium, did not include the purchase of any smart or personal device.

The main limitations of the careflow mining algorithm, especially for its application in electronic phenotyping, have been discussed in 5.1.5. As already mentioned in 7.3, some of the careflow mining algorithm functionalities were readapted to integrate the algorithm in the CDSS, whereas other functionalities were not implemented (Support tuning, AND events recognition and Jaccard similarity). The main reason behind these choices is to simplify the user interface. Anyway, the exploitation of more adaptable visual analytics tools (for example the Plotly, Shiny R libraries) could have mitigated some of these limitations.

During the system evaluation focus meetings, participants made several comments regarding possible improvement or addition of CDSS functionalities. These comments were mainly related to clinical actions. T2DM specialist doctors and managers expressed the interest in the possibilities of: (i) Refining the system tuning features for better customized filtering and temporal analyses functionalities. For example, adding the possibility to select specific time windows to analyze the population and to personalize the sorting of the careflows results; (ii) Improving the personalization of results display, like adding the possibility of filtering on the basis of a blood glucose control; (iii) Adding further information and statistics to answer specific research questions on the population, for example the statistics on the first complications occurring to the patients.

9.3 Range of applicability and Future works

The longitudinal data analytics methods have been implemented within the CDSS to tackle the specific issue of T2DM management. The CFM algorithm potential generalizability has been widely discussed in 5.1.5. The results obtained with the CFM algorithm offer a promising scenario for its application in the comparison of mined careflows with clinical protocols followed by clinical centers, to identify treatments potential deviations, and compare treatments costs of each phenotype mined from the hospital information systems. The CFM algorithm can be applied to trace administrative processes, as well as diagnostic and therapeutic plans. The computable phenotypes extracted via CFM can be seen as the exposure to clinical actions, and used to evaluate outcomes like changes in performances scales at the beginning and the end of different processes. The exploitation of the CFM algorithm gives the possibility to reply to scientific and managerial questions like:(i) is a careflow more willing to enhance patient's condition than another, given comparable health status at its beginning? (ii) is a careflow more willing to reduce health cost than another, given comparable health procedures?

All the methods proposed within Aim 2 could be applied in other clinical contexts, where there is an interest in discovering different clinical evolutions that may happen to similar patients. Drug Exposure patterns mining and Hierarchical Bayesian Regression models are based not only on population parameters, but they also consider single patient intra-variability during consecutive encounters. The implementation of these approaches opens the possibility to tailor, and to update during consecutive follow-ups, individual descriptive and predictive models, which can be exploited in any chronic care setting.

In the following are described the potential applications of whole decision support framework.

The ***Patient Use Case***, which in this work has been customized for T2DM Specialists, can be used to foster the role of Primary Care or Endocrinologists in the care of diabetic patients. The CDSS can enhance the capability of General Practitioners to be the pivoting point for the coordination of the whole clinical and social specialties involved in the chronic management of T2DM. The adaptation of the system for Primary Care would require the inclusion of specific T2DM clinical guidelines, consultations and laboratory tests to perform. This would trigger the shift of General Practitioners to have a central role in diabetic patients' management, enabling the comparison between the guideline recommendations and the actual clinical pathway that individual patients undergo. The integration of data coming from heterogeneous sources has been considered one of the most interesting features of this use case. The proposed solution to gather, jointly analyze and display clinical and administrative data, allows coordinating multiple care providers and possibly solving some issues, like badly scheduled follow-ups or exams repetitions, which are typical of chronic diseases.

From the point of view of the analysis methods, and their integration in the CDSS, further efforts will be dedicated to validate the Hierarchical Risk Models for personal risk calibration and to exploit the careflow mining algorithm, in particular the similarity measure, to compare single patient with guide lines or a reference cohort.

The ***Population Use Case*** can be defined as a business intelligence tool to support hospital and healthcare agencies managers to understand, in their hospitals and territories, the evolution of the disease at population level and the use of resources of the healthcare systems, both public for national health systems or private for insurance companies.

The CDSS use case has the potential to facilitate the strategic approach of T2DM socioeconomic issues. The provided information can be very useful to (i) improve the internal organization of the healthcare system, (ii) estimate the clinical specialties which might have higher workload in the future, (iii) detect different phenotypes in geographic regions, (iv) measure and compare the impact of care approaches in different healthcare organizations.

The developed tool leverages on the availability of the long-term monitoring of T2DM patients' data and uses them to facilitate the control and management of the chronic disease and its complications. The tool, or at least the core careflow mining and drill down functionality, can be potentially implemented in every center collecting data about lifestyle, physical parameters, clinical values, hospitalizations, and medication of chronic patients.

To enhance the CDSS potentials, future works will be addressed to integrate clinical guidelines and allow the comparison of the mined careflows with the suggested ones. Depending by the future availability of health services related costs, the detected clinical pathway will be compared and analyzed in terms of total expenditures.

Possible translations of the obtained results into other clinical domains might involve both chronic and acute disease.

Adding a temporal dimension in population stratification can give the opportunity to gain new insight especially in ***chronic diseases***, which are usually characterized by a slow evolution pace

and multiple actors. The developed algorithms and framework can be exploited in primary care for population monitoring, or in rehabilitation centers, possibly integrating data from biosensor and biomedical instrumentation.

Respiratory diseases represent an interesting field where the framework might be exported. These disease areas are relevant to current public health authorities and it is possible to foresee an engagement of the public health authorities in projects devoted to their prevention and enhanced management. Chronic Obstructive Pulmonary Disease (COPD) is the 4th mortality cause at global level, and its prevalence increases when data from GPs are considered. In Europe, where 6% of health expenditures is due to respiratory diseases, COPD accounts for the 56% of this amount (WHO 2016; Roisin RR 2016; Smith & Wrobel 2014). Numerous precision medicine issues are involved in COPD patients care, as the need of evaluating if current treatments can be enhanced by broader range of risk factors. It will be possible to leverage on the results described in this research program to implement a platform able to plan health care policies with longer time horizons, to stratifies the population through geographic regions and tailored indicators, and to forecasts complex scenarios for the disease prevalence. Asthma affects a younger cohort of patients, treated both by GPs and pediatricians. The disease is underestimates by public health, especially in its early stages or complication-free cases. Large part of the patients is affected by multiple comorbidities and the role of climate on the geographic variability of asthma and respiratory symptoms has been demonstrated (Zanolin et al. 2004). Actual Treatments has to be update to take into account other risk factors and re-calibrate treatment on the basis of other patients characteristics (Himes et al. 2008; Meystre et al. 2009; Tomasallo et al. 2014). The achievements of this research might be the pillars to build a dashboard based upon clinical, environmental and self-monitoring data. The framework could include novel risk models, based upon environmental risk factors and seasonal trends, and gives the opportunity to include self-monitoring data collected by patients during daily activities, from wearable devices or smartphone apps. The possibility to include in a CDSS novel diseases models based on of the evolution of seasonal changes in air quality, diagnosis of comorbidities, and monitored life style activities, is the first step to build innovative systems able to control a patients' cohorts in real time, and to plan health care policies with longer time horizons.

Other possibilities for the clinical translation of results might be represented by those settings characterized by limited actions in more defined time windows, for example actions performed during single hospitalizations or ***acute diseases***.

TraumAID (Webber et al. 1998) is a CDSS for injured patients care. It well describes possible applications of CDSS in complex and multivariate settings, and also their varying cases complexity. The evaluation of potential impacts in cognitive interactions and responses of users to these systems should be carefully studied. Fast processing of events, and their temporal relationship, might suggest actions on the basis of previously treated cases, compared through their evolving levels of complexity. The developed CDSS can be readapted to support clinical actions in these scenarios, to detect processes that happen, and might change, very quickly or to identify triggers for rapid deterioration of patients' conditions.

9.4 Conclusions

This dissertation illustrates how, in medical informatics, the implementation of data models that integrate clinical, environmental and behavioral data of a specific population can be the pillar of a holistic analysis framework, which integrates heterogeneous characteristics of a chronic disease and produces new inferences from multiple events.

The overall objective of the final system is to link exposure and outcome data to support clinical diagnostics and therapeutic interventions in T2DM. Nevertheless, the data gathering efforts produce a functional layer for further research activities, in particular for the approach of medical and management issues in chronic care through longitudinal analysis methods. The results of this work mainly depend on the temporal dimension of data to tackle T2DM related issues. The shift from a cross-sectional to a longitudinal approach also enhances the understanding of clinical phenomena and, when supported by suitable technologies, helps healthcare professionals to make better decisions. The detection of temporal patterns that define electronic phenotypes and their integration into clinical practice and healthcare management triggers novel inferences about event causality.

The developed framework create value from longitudinal data to enable a translation in the rational models for T2DM diagnosis, therapies planning, disease monitoring and, ultimately, in scientific discovery. When new information is properly organized and displayed, it illuminates assumptions not previously considered. This generates new cycles where explanatory assumptions can be formed and evaluated and users are supported to understand entailments among events, and establish new ways to evaluate procedures and medical complexity. Thus, healthcare providers and decision makers are enabled to revise the data of the T2DM patients' cohort to get new insights about the disease, and to discover and explain the most important clinical events within a learning health system.

Appendix A

Patient Data

Variable	CONCEPT PATH	VALUES	DESCRIPTION
Gender	\Patient_Data\Gender\	Female, Male , Unknown	
Anamnesis	\Patient_Data\Anamnesis\Diagnosis of Diabetes\Year of Diagnosis\	Year of Diagnosis	Onset year
	\Patient_Data\Anamnesis\Family History of Diabetes\	Yes, No	Cases of Diabetes within the family and degree of kinship
Vital Status	\Patient_Data\Vital Status\	Alive or Dead. In the case of dead additional inforamtion on regards of Year of Death and Cause (Acute Myocardial Infarction, Cardiac Event, Hypercalcemia, Infections, Other, Stroke)	
Cardiovascular risk	\Patient_Data\Cardiovascular Risk\	Six level of risk from I to VI	Cardiovascular risk calculated through the Progetto Cuore algorithm and defined thresholds
Education	\Patient_Data\Education\	Years studying	
Level of Complexity	\Patient_Data\Level of Complexity\	Stable, 1st_Level, 2th_Level, 3th_Level	Patient Status Evolution levels on the basis of Complications onset and related hospitalizations
Profession	\Patient_Data\Profession\	Clerical, Manual worker, Student, Housewife, Retired,	Selected type of profession, as defined in clinical studies

		Unemployed, Not known	
Marital Status	\Patient_Data\Marital Status\	Single, Married, Widowed, Divorced	Marital status, as defined in clinical studies
Ethnicity	\Patient_Data\Ethnicity\	Spanish, Finnish, Italian, Greek	Patients provenience on the basis of the data set source
Metabolic syndrome	\Patient_Data\Metabolic Syndrome\	Yes or No	In hospital data set all the patients are T2DM, in this case every subject had the value set to Yes

Contact Details

Variable	**CONCEPT_PATH**	**VALUES**	**DESCRIPTION**
Hospitalization	\Contact\Hospitalization\Admission Mode\	Emergency, Planned	Data retrieved from administrative flows, indicating patients contacts with health care structures on the territory
	\Contact\Hospitalization\Course\	Day Hospital, In Hospital	
	\Contact\Hospitalization\Discharge Mode\	Assigned to another Hospital, At Home with Assistant, At Home without Assistant, Death, Other	
Life Style	\Contact\Life Style\Alcohol Habit\	Alcohol serving per weeks: 0-4, 5-10 , 11-20, >20	Observations retrieved from clinicians observations during follow-up (hospital data set)
	\Contact\Life Style\Diet\	Bad, Good	

	\Contact\Life Style\Physical Activity\	Intense, Light, Moderate, No	or from clinical studies records Note Physical Activity → from Hospital data sets AND Physical Activity in Leisure Time → from Clinical studies
	\Contact\Life Style\Physical Activity in Leisure Time\	Physical Activity in Leisure Time, Regular physical activity in leisure time, Sporadic physical activity in leisure time	
	\Contact\Life Style\Smoking Habit\	Current, Ex, Never	
Physical Examination	\Contact\Physical Examination\Blood Pressure\	Systolic and Diastolic values	Observations retrieved from clinicians observations during follow-up visits
	\Contact\Physical Examination\BMI\	Numeric values	
	\Contact\Physical Examination\Height\	Numeric values	
	\Contact\Physical Examination\HIP\	Numeric values	
	\Contact\Physical Examination\Pulse\	Numeric values	
	\Contact\Physical Examination\Waist\	Numeric values	
	\Contact\Physical Examination\Weight\	Numeric values	

Laboratory Exams

Variable	CONCEPT_PATH	VALUES	DESCRIPTION

Laboratory	\Laboratory\Cholesterol\	Total, HDL and LDL values	Data retrieved from hospital flows from Laboratory exam ward, validated by clinicians
	\Laboratory\Glucose Fasting\		
	\Laboratory\Glucose Fasting 2H_OGTT\		
	\Laboratory\Fasting Insulin\		
	\Laboratory\Hba1c\	Values both in % and converted in mmol/mol	
	\Laboratory\Trygliceride\		
	\Laboratory\Uric Acid\		

Complications and Comorbidities

Variable	**CONCEPT_PATH**	**VALUES**	**DESCRIPTION**
Complications	\Complications\Macro\	Coronary Artery Disease, Acute Myocardial Infarction, Angina, Chronic Ischemic Heart Disease, Occlusion and Stenosis of Carotid Artery, Peripheral Artery Occlusive Disease, Stroke (*390-459.99*)	Complications are recorded by clinicians during encounters visits. Each observation is associated to a list of ICD9 codes and with the information about the onset date.
	\Complications\Micro\	Diabetic Foot, Nephropathy (*580 - 589.99*),	

		Retinopathy (diabetic: *362.0*)	
	\Complications\Not Vascular\	Fat Liver Disease, Neuropathy (*580 - 589.99*)	

Drugs

Variable	CONCEPT_PATH	VALUES	DESCRIPTION
Therapy Prescription	\Drugs\ Prescription \Anti-Thrombotic\		Data about patients therapies plans as stated by practitioners.
	\Drugs\ Prescription \Anti-Hypertensive\		
	\Drugs\ Prescription \Lipid-Lowering\		
	\Drugs\ Prescription \Diuretics\		
	\Drugs\ Prescription \ Therapy for Diabetes\		
Pharmaceutical Data	\Drugs\Pharma\Anti-Thrombotic\	DDD values indicating the period cover by the prescription	Data about patients purchases over the whole territory. Data retrieved from Administrative data bases.
	\Drugs\Pharma \Anti-Hypertensive\		
	\Drugs\Pharma \Lipid-Lowering\		
	\Drugs\Pharma \Diuretics\		
	\Drugs\ Pharma \ Therapy for Diabetes\		
Drugs Adherence	\Drugs\Adherence\Anti-Thrombotic\	Adherence index	A sub-folder in the Drug metadata table to detect observations related to the adherence of a certain drug during fixed periods. The
	\Drugs\Adherence\Anti-Hypertensive\		
	\Drugs\Adherence\Lipid-Lowering\		

	\Drugs\Adherence\Diuretics\		values are pre-processed and shows therapy adherence through the CSA index.
	\Drugs\Adherence\ Therapy for Diabetes\		

ICD 9 –CM - Hospital Admission

Variable	CONCEPT_PATH	VALUES	DESCRIPTION
ICD9 –CM	\icd9\Diagnoses\	Principal Diagnosis or Secondary Diagnosis and realted code	Registered ICD9-CM codes from patients in-hospital admissions
	\icd9\Procedures\		

Visits and Follow up

Variable	CONCEPT_PATH	VALUES	DESCRIPTION
Visits and Follow up	\Visit Detail\	Indication about the type of encounter: 1st Visit, Follow Up, Other Visit	Visits and Follow up concepts observations have been inserted in order to build patients histories, as in the UC2

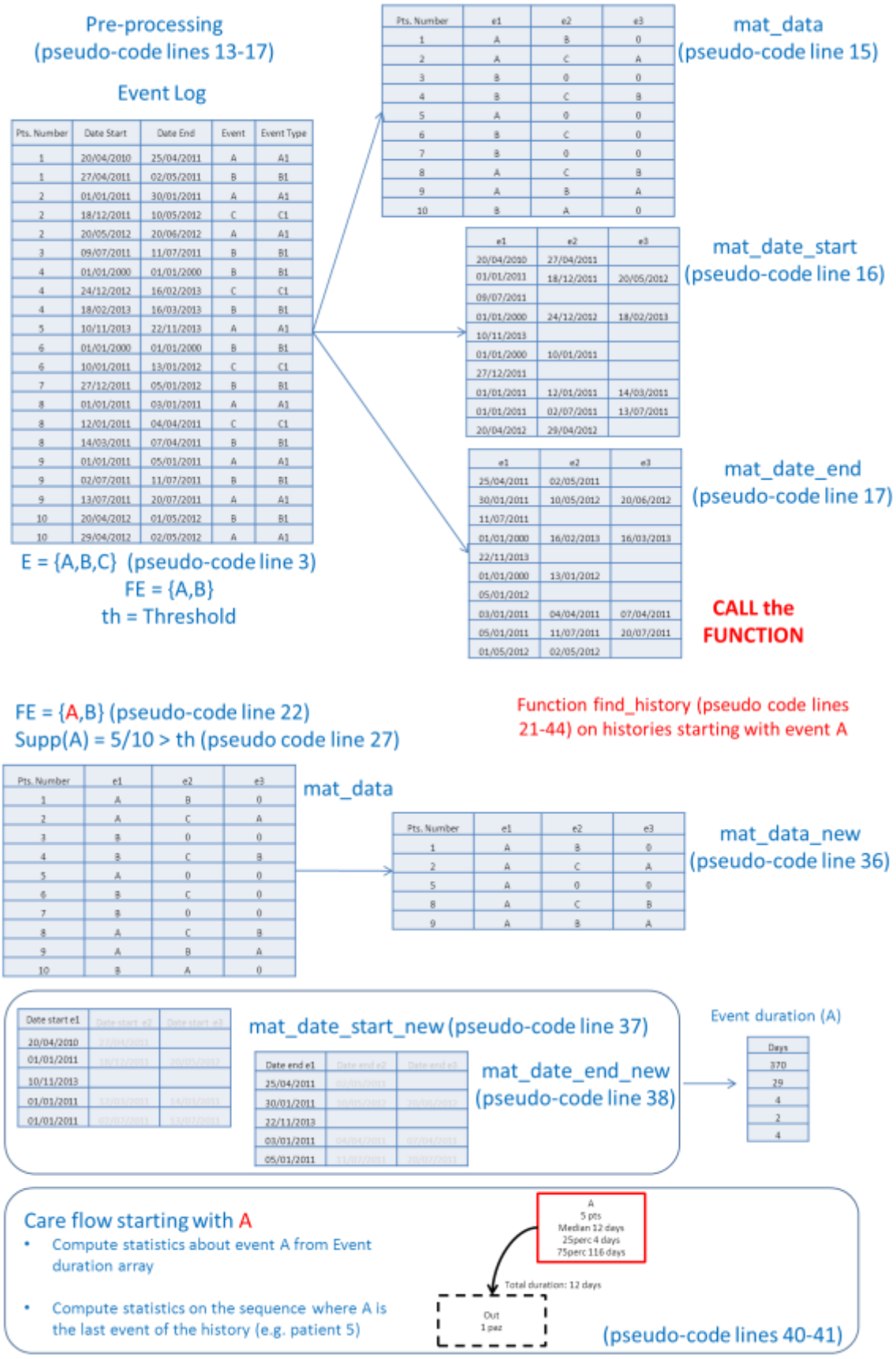
Living Area

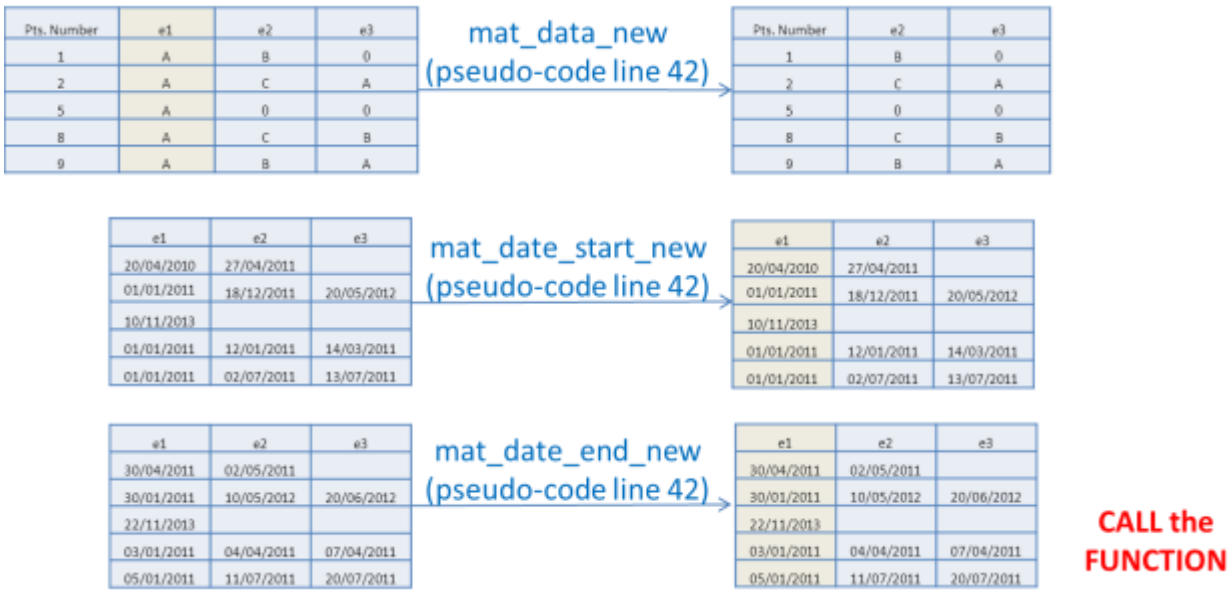
Variable	CONCEPT_PATH	VALUES	DESCRIPTION
Living Area (Environmental)	\Living Area\	In the case of the Pavia Area the Ontology has two level accounting for three Districts and nine Sub-Districts	Geo-referenced information about patients' provenience and living area. Observation retrieved from Administrative flows.

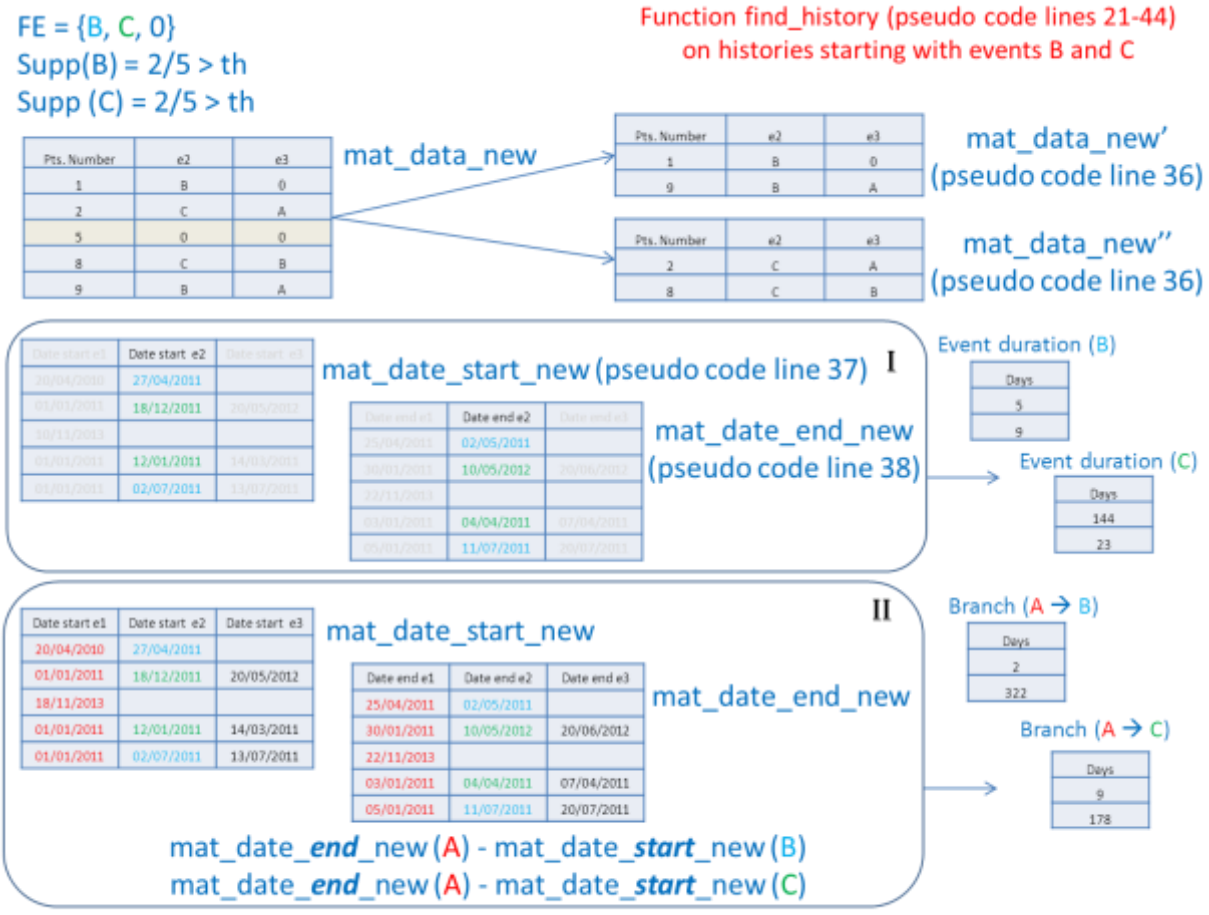
Temporal Abstraction

Variable	CONCEPT_PATH	VALUES	DESCRIPTION
Temporal Abstraction	\Temporal\Diet\	Bad or Good for at leat 6 months	Temporal abstractions are generated by a dedicated module integrated in the Dashboard. The JTSA module automatically retrieve time point data and compute interval based qualitative observations.
	\Temporal\Weight\	Decreasing, Time to Target (in decrese at leat of the 10% in 6 months)	
	\Temporal\Hba1c\	TimeToTarget, if reach a fixed threshold (7-7.5-8) in 6 months	

Appendix B


Pre-processing
(pseudo-code lines 13-17)
Event Log
E = {A,B,C} (pseudo-code line 3)
FE = {A,B}
th = Threshold
mat_data
(pseudo-code line 15)
mat_date_start
(pseudo-code line 16)
mat_date_end
(pseudo-code line 17)
CALL the
FUNCTION
FE = {A,B} (pseudo-code line 22)
Supp(A) = 5/10 > th (pseudo code line 27)
Function find_history (pseudo code lines
21-44) on histories starting with event A
mat_data
mat_data_new
(pseudo-code line 36)
mat_date_start_new (pseudo-code line 37)
mat_date_end_new
(pseudo-code line 38)
Event duration (A)
Care flow starting with A
Compute statistics about event A from Event duration array
Compute statistics on the sequence where A is the last event of the history (e.g. patient 5)
A
5 pts
Median 12 days
25perc 4 days
75perc 116 days
Total duration: 12 days
Out
1 paz
(pseudo-code lines 40-41)

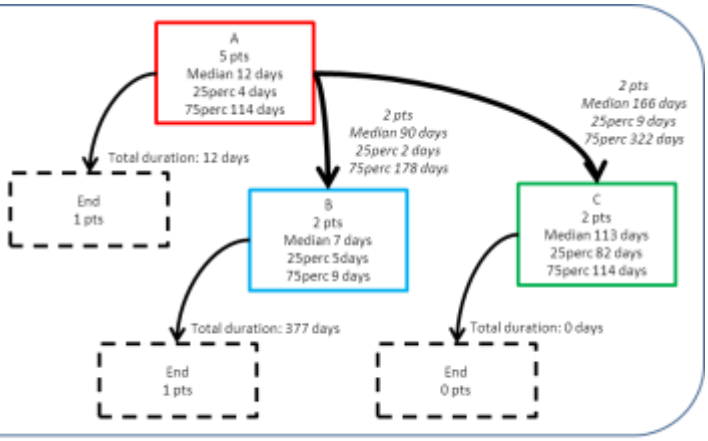

mat_data_new
(pseudo-code line 42)
mat_date_start_new
(pseudo-code line 42)
mat_date_end_new
(pseudo-code line 42)
CALL the FUNCTION


FE = {B, C, 0}
Supp(B) = 2/5 > th
Supp (C) = 2/5 > th
Function find_history (pseudo code lines 21-44)
on histories starting with events B and C
mat_data_new
mat_data_new'
(pseudo code line 36)
mat_data_new''
(pseudo code line 36)
mat_date_start_new (pseudo code line 37)
I
mat_date_end_new
(pseudo code line 38)
Event duration (B)
Event duration (C)
II
mat_date_start_new
mat_date_end_new
Branch (A → B)
Branch (A → C)
mat_date_end_new (A) - mat_date_start_new (B)
mat_date_end_new (A) - mat_date_start_new (C)

For both B and C

(pseudo code lines 40,41)

- I Compute statistics about the events B and C Event duration arrays
- Compute statistics on the sequence where B, C are the last event of the history (e.g. patient 1 for branch A, B) form the first event A
- II *Compute statistics for the branches A →B and A →C*



Bibliography

A.-J., P. et al., 2010. Stressful life events and the metabolic syndrome: The prevalence, prediction and prevention of diabetes (PPP)-botnia study. *Diabetes Care*, 33(2), pp.378–384. Available at: http://care.diabetesjournals.org/content/33/2/378.full.pdf+html%5Cnhttp://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed9&NEWS=N&AN=2010070033.

van der Aalst, W., 2011. Process mining: discovering and improving Spaghetti and Lasagna processes. *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, (c), pp.1–7. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6129461.

Van Der Aalst, W., Weijters, T. & Maruster, L., 2004. Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9), pp.1128–1142.

van der Aalst, W.M.P. et al., 2015. *Process Discovery Using Localized Events*,

van der Aalst, W.M.P., 2011. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, Available at: http://www.ncbi.nlm.nih.gov/pubmed/18487736.

van der Aalst, W.M.P. & Weijters, A.J.M.M., 2004. Process mining: a research agenda. *Computers in Industry*, 53(3), pp.231–244. Available at: http://www.sciencedirect.com/science/article/pii/S0166361503001945.

Abraham, J. & Reed, T., 2002. Progress, innovation and regulatory science in drug development: the politics of international standard-setting. *Social Studies of Science*, 32(3), pp.337–369.

Ackoff, R.L., 1989. From data to wisdom. *Journal of Applied Systems Analysis*, 16(1), pp.3–9.

Adlassnig, K.P. et al., 2006. Temporal representation and reasoning in medicine: Research directions and challenges. *Artificial Intelligence in Medicine*, 38(2), pp.101–113.

Agrawal, R. & Srikant, R., 1995. Mining sequential patterns. *Proceedings of the Eleventh International Conference on Data Engineering*.

Albers, D.J. et al., 2014. Dynamical phenotyping: Using temporal analysis of clinically collected physiologic data to stratify populations. *PLoS ONE*, 9(6).

Alberti, K.G. & Zimmet, P.Z., 1998. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. *Diabetic medicine : a journal of the British Diabetic Association*, 15(7), pp.539–553.

Allen, J.F., 1984. Towards a general theory of action and time. *Artificial Intelligence*, 23(2), pp.123–154.

American Diabetes Association, 2014. Standards of medical care in diabetes--2014. *Diabetes care*, 37 Suppl 1(October 2013), pp.S14-80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24357209.

Ampudia-Blasco, F.J. et al., 2015. A decision support tool for appropriate glucose-lowering therapy in patients with type 2 diabetes. *Diabetes Technol Ther*, 17(3), pp.194–202.

Anderson, A.E. et al., 2015. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. *Journal of biomedical informatics*, 60, pp.162–168. Available at: http://www.sciencedirect.com/science/article/pii/S1532046415002865.

Arfè, A. & Corrao, G., 2015. Tutorial: strategies addressing detection bias were reviewed and implemented for investigating the statins-diabetes association. *Journal of Clinical Epidemiology*, 68(5), pp.480–488.

Aronson, R. et al., 2014. OpT2mise: A Randomized Controlled Trial to Compare Insulin Pump Therapy

with Multiple Daily Injections in the Treatment of Type 2 Diabetes-Research Design and Methods. *Diabetes technology & therapeutics*, 16(7), pp.414–20. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24735134.

Ashley, E. a, 2015. The precision medicine initiative: a new national effort. *Jama*, Published(21), pp.E1-2.

Augstein, P. et al., 2010. Translation of personalized decision support into routine diabetes care. *Journal of diabetes science and technology*, 4(6), pp.1532–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3005067&tool=pmcentrez&rendertype=abstract.

Augusto, J.C., 2005. Temporal reasoning for decision support in medicine. *Artificial Intelligence in Medicine*, 33(1), pp.1–24.

Ayres, J. et al., 2002. Sequential Pattern mining using a bitmap representation. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, p.429. Available at: http://portal.acm.org/citation.cfm?doid=775047.775109.

Bahcall, O., 2015. Precision medicine. *Nature*, 526(7573), pp.335–335. Available at: http://www.nature.com/doifinder/10.1038/526335a.

Barbour, V. et al., 2013. Health care delivery. Open mHealth architecture: an engine for health care innovation. *PLoS Medicine*, 10(2), p.e10011395.

Barlow, J. & Krassas, G., 2013. Improving management of type 2 diabetes: Findings of the Type2Care clinical audit. *Australian Family Physician*, 42(1), pp.57–60.

Barrett, J.S. et al., 2008. Integration of modeling and simulation into hospital-based decision support systems guiding pediatric pharmacotherapy. *BMC medical informatics and decision making*, 8, p.6.

Batal, I. et al., 2009a. Multivariate Time Series Classification with Temporal Abstractions. *Florida Artificial Intelligence Research Society Conference*, pp.344–349. Available at: http://www.aaai.org/ocs/index.php/FLAIRS/2009/paper/view/48.

Batal, I. et al., 2009b. Multivariate Time Series Classification with Temporal Abstractions. *Florida Artificial Intelligence Research Society Conference*, pp.344–349.

Batley, N.J. et al., 2011. Implementation of an emergency department computer system: Design features that users value. *Journal of Emergency Medicine*, 41(6), pp.693–700.

Bekkar, M., Djemaa, H.K. & Alitouche, T.A., 2013. Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*, 3(10), pp.27–38. Available at: http://www.iiste.org/Journals/index.php/JIEA/article/view/7633.

Belard, A. et al., 2016. Precision diagnosis: a view of the clinical decision support systems (CDSS) landscape through the lens of critical care. *Journal of Clinical Monitoring and Computing*, pp.1–11.

Bellazzi, R., 2014. Big data and biomedical informatics: a challenging opportunity. *Yearbook of medical informatics*, 9, pp.8–13. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4287065&tool=pmcentrez&rendertype=abstract.

Bellazzi, R. et al., 2015. Big Data Technologies: New Opportunities for Diabetes Management. *Journal of Diabetes Science and Technology*. Available at: http://dst.sagepub.com/lookup/doi/10.1177/1932296815583505.

Bellazzi, R. et al., 2000. Intelligent analysis of clinical time series: An application in the diabetes mellitus domain. *Artificial Intelligence in Medicine*, 20(1), pp.37–57.

Bellazzi, R., Sacchi, L. & Concaro, S., 2009. Methods and tools for mining multivariate temporal data in

clinical and biomedical applications. In *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*. pp. 5629–5632.

Van Belle, V. & Van Calster, B., 2015. Visualizing risk prediction models. *PLoS ONE*, 10(7).

Benin, A.L. et al., 2011. How Good Are the Data? Feasible Approach to Validation of Metrics of Quality Derived From an Outpatient Electronic Health Record. *American Journal of Medical Quality*, 26(6), pp.441–451. Available at: http://ajm.sagepub.com/cgi/doi/10.1177/1062860611403136.

Bettini, C., Wang, X.S. & Jajodia, S., 1996. A General Framework for Time Granularity and Its Application to Temporal Reasoning. In *Third International Workshop on Temporal Representation and Reasoning (TIME-96)*. pp. 104--111.

Blake, P.M. et al., 2011. Toward an integrated knowledge environment to support modern oncology. *Cancer journal (Sudbury, Mass.)*, 17(4), pp.257–63. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21799334.

Bødker, K. & Granlien, M.F., 2008. Computer support for shared care of diabetes: findings from a Danish case. *Studies in health technology and informatics*, 136, pp.389–94. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18487762.

Bouarfa, L. & Dankelman, J., 2012. Workflow mining and outlier detection from clinical activity logs. *Journal of Biomedical Informatics*, 45(6), pp.1185–1190.

Bouaud, J., Falcoff, H. & Seroussi, B., 2013. Simultaneously authoring and modeling clinical practice guidelines: a case study in the therapeutic management of type 2 diabetes in France. *Studies in health technology and informatics*, 186, pp.108–112.

Bourne, P.E. et al., 2015. The NIH big data to knowledge (BD2K) initiative. *Journal of the American Medical Informatics Association*, 22(6), pp.1114–1114.

Bowman, S., 2013. Impact of electronic health record systems on information integrity: quality and safety implications. *Perspectives in health information management / AHIMA, American Health Information Management Association*, 10, p.1c. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3797550&tool=pmcentrez&rendertype=abstract.

Brambilla, M. & Fraternali, P., 2015. *Interaction Flow Modeling Language*, Available at: http://www.sciencedirect.com/science/article/pii/B9780128001080000102.

Brown, D.E., 2008. Introduction to Data Mining for Medical Informatics. *Clinics in Laboratory Medicine*, 28(1), pp.9–35.

Cadarette, S.M. & Burden, A.M., 2010. Measuring and improving adherence to osteoporosis pharmacotherapy. *Current Opinion in Rheumatology*, 22(4), pp.397–403.

Canavero, I. et al., 2016. Safely Addressing Patients with Atrial Fibrillation to Early Anticoagulation after Acute Stroke. *Journal of Stroke and Cerebrovascular Diseases*. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1052305716302944.

Canestaro, W.J., Austin, M. a & Thummel, K.E., 2014. Genetic factors affecting statin concentrations and subsequent myopathy: a HuGENet systematic review. *Genetics in medicine : official journal of the American College of Medical Genetics*, 16(11), pp.810–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24810685.

Caron, F. et al., 2012. Advanced care-flow mining and analysis. *Lecture Notes in Business Information Processing*, 99 LNBIP, pp.167–168. Available at: http://link.springer.com/chapter/10.1007/978-3-642-28108-2_18.

Caron, F. et al., 2014. Monitoring care processes in the gynecologic oncology department. *Computers in Biology and Medicine*, 44(1), pp.88–96.

Castaneda, C. et al., 2015. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of clinical bioinformatics*, 5, p.4. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4381462&tool=pmcentrez&rendertype=abstract.

Cederberg, H. et al., 2015. Increased risk of diabetes with statin treatment is associated with impaired insulin sensitivity and insulin secretion: a 6 year follow-up study of the METSIM cohort. *Diabetologia*, 58(5), pp.1109–1117.

Chan, J.C.N. et al., 2009. The joint asia diabetes evaluation (JADE) program: A web-based program to translate evidence to clinical practice in type 2 diabetes. *Diabetic Medicine*, 26(7), pp.693–699.

Chaudhry, A. & Feest, T., 2011. Chapter 14: Enhancing access to UK renal registry data through innovative online data visualisations. *Nephron - Clinical Practice*, 119(SUPPL. 2).

Cheong, C. et al., 2008. Patient adherence and reimbursement amount for antidiabetic fixed-dose combination products compared with dual therapy among texas medicaid recipients. *Clinical Therapeutics*, 30(10), pp.1893–1907.

Chesani, F. et al., 2008. Compliance checking of cancer-screening careflows: An approach based on Computational Logic. In *Studies in Health Technology and Informatics*. pp. 183–192.

Cho, B.H. et al., 2008. Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. *Artificial Intelligence in Medicine*, 42(1), pp.37–53.

Chuang, K. et al., 2011. Long-term air pollution exposure and risk factors for cardiovascular diseases among the elderly in Taiwan. *Occup Environ Med*, 68(2), pp.64–68.

Chui, K.K.H. et al., 2011. Visual analytics for epidemiologists: Understanding the interactions between age, time, and disease with multi-panel graphs. *PLoS ONE*, 6(2).

Cichosz, S.L., Johansen, M.D. & Hejlesen, O., 2015. Toward Big Data Analytics: Review of Predictive Models in Management of Diabetes and Its Complications. *Journal of diabetes science and technology*, 10(1), pp.27–34. Available at: http://dst.sagepub.com/content/10/1/27.full.

Cimino, J.J., 1997. Intranet technology in hospital information systems. In *Studies in Health Technology and Informatics*. pp. 102–109.

Cimino, J.J. et al., 2013. Practical choices for infobutton customization: experience from four sites. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2013, pp.236–45. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3900175&tool=pmcentrez&rendertype=abstract.

Cimino, J.J. et al., 2007. Redesign of the Columbia University Infobutton Manager. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pp.135–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2655841&tool=pmcentrez&rendertype=abstract.

Cleveringa, F.G. et al., 2013. Computerized decision support systems in primary care for type 2 diabetes patients only improve patients' outcomes when combined with feedback on performance and case management: a systematic review. *Diabetes Technol Ther*, 15(2), pp.180–192. Available at: http://online.liebertpub.com/doi/pdfplus/10.1089/dia.2012.0201.

Cleveringa, F.G.W. et al., 2008. Combined task delegation, computerized decision support, and feedback improve cardiovascular risk for type 2 diabetic patients: a cluster randomized trial in primary care. *Diabetes care*, 31(12), pp.2273–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2584178&tool=pmcentrez&rendertype

=abstract.

Collins, F.S. & Varmus, H., 2015. A new initiative on precision medicine. *N Engl J Med*, 372(9), pp.793–795.

Colombo, G.L. et al., 2012. Antidiabetic therapy in real practice: Indicators for adherence and treatment cost. *Patient Preference and Adherence*, 6, pp.653–661.

Combi, C. et al., 2014. Mining approximate temporal functional dependencies with pure temporal grouping in clinical databases. *Computers in Biology and Medicine*, 62, pp.306–324.

Combi, C., 2004. Representing and Reasoning about Temporal Granularities. *Journal of Logic and Computation*, 14(1), pp.51–77. Available at: http://logcom.oxfordjournals.org/content/14/1/51.abstract.

Combi, C. et al., 2009. Temporal similarity measures for querying clinical workflows. *Artificial Intelligence in Medicine*, 46(1), pp.37–54.

Combi, C. & Chittaro, L., 1999. Abstraction on clinical data sequences: An object-oriented data model and a query language based on the event calculus. *Artificial Intelligence in Medicine*, 17(3), pp.271–301.

Combi, C. & Shahar, Y., 1997. Temporal reasoning and temporal data maintenance in medicine: Issues and challenges. *Computers in Biology and Medicine*, 27(5), pp.353–368.

Concaro, S. et al., 2011. Mining health care administrative data with temporal association rules on hybrid events. *Methods of Information in Medicine*, 50(2), pp.166–179.

Conway, M. et al., 2011. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. *AMIA Annu Symp Proc*, 2011, pp.274–283.

Cook, J.E. & Wolf, A.L., 1998. Discovering models of software processes from event-based data. *ACM Transactions on Software Engineering and Methodology*, 7(3), pp.215–249.

Cryer, P.E., 2008. The barrier of hypoglycemia in diabetes. *Diabetes*, 57(12), pp.3169–3176.

Dagliati, A., Sacchi, L., Bucalo, M., et al., 2014. A data gathering framework to collect Type 2 diabetes patients data. *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, (June), pp.244–247. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6864349.

Dagliati, A. et al., 2015. Integration of Administrative, Clinical, and Environmental Data to Support the Management of Type 2 Diabetes Mellitus: From Satellites to Clinical Care. *Journal of diabetes science and technology*, 10(1), pp.19–26.

Dagliati, A., Sacchi, L., Cerra, C., et al., 2014. Temporal data mining and process mining techniques to identify cardiovascular risk-associated clinical pathways in Type 2 diabetes patients. *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp.240–243. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6864348.

Das, a K. et al., 1992. An extended SQL for temporal data management in clinical decision-support systems. *Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care*, pp.128–32. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2248091&tool=pmcentrez&rendertype=abstract.

Disease, Committee on A Framework for Developing a New Taxonomy of Disease, Board of Life Sciences, Division of Earth and Life Sciences, N.R.C. of e N.A., 2011. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*, Available at: http://www.ncbi.nlm.nih.gov/pubmed/22536618.

Dixon, B.E. et al., 2013. Improving Medication Adherence for Chronic Disease Using Integrated e-Technologies. *Medinfo 2013*, 18(7), p.2013.

Dombrowsky, E. et al., 2011. Evaluating performance of a decision support system to improve methotrexate pharmacotherapy in children and young adults with cancer. *Therapeutic drug monitoring*, 33(1), pp.99–107. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3074357&tool=pmcentrez&rendertype=abstract.

Donsa, K. et al., 2016. Impact of errors in paper-based and computerized diabetes management with decision support for hospitalized patients with type 2 diabetes. A post-hoc analysis of a before and after study. *International Journal of Medical Informatics*, 90, pp.58–67.

Ebrahiminia, V. et al., 2006. Design of a decision support system for chronic diseases coupling generic therapeutic algorithms with guideline-based specific rules. *Studies in health technology and informatics*, 124, pp.483–488. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17108565&retmode=ref&cmd=prlinks%5Cnpapers2://publication/uuid/6F062E36-56F8-4557-8561-756D917D1B67.

Elixhauser, a, Steiner, C. & Palmer, L., 2014. Clinical Classifications Software (CCS), 2014. *U.S. Agency for Healthcare Research and Quality*, (November 2013), pp.1–54. Available at: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

Etheredge, L.M., 2014. Rapid learning: A breakthrough agenda. *Health Affairs*, 33(7), pp.1155–1162.

Fauci, A. et al., 2008. *Harrison's principles of internal medicine*,

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), p.37.

Fei Wang, N.L., 2013. A Framework for Mining Signatures from Event Sequences and Its Applications in Healthcare Data. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 35(2), p.272. Available at: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6200289.

Fernández-Llatas, C. et al., 2010. Activity-based Process Mining for Clinical Pathways computer aided design. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*. pp. 6178–6181.

Fetter, R.B. & Freeman, J.L., 1986. Diagnosis related groups: product line management within hospitals. *Academy of management review. Academy of Management*, 11(1), pp.41–54.

Fleurence, R.L. et al., 2014. Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association : JAMIA*, 21(4), pp.578–582. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=24821743&retmode=ref&cmd=prlinks%5Cnpapers2://publication/doi/10.1136/amiajnl-2014-002747.

Frey, L.J., Lenert, L. & Lopez-Campos, G., 2014. EHR Big Data Deep Phenotyping. Contribution of the IMIA Genomic Medicine Working Group. *Yearbook of medical informatics*, 9, pp.206–11. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4287080&tool=pmcentrez&rendertype=abstract.

Frith, K.H., Anderson, F. & Sewell, J.P., 2010. Assessing and selecting data for a nursing services dashboard. *The Journal of nursing administration*, 40(1), pp.10–16.

Gabriel-Sánchez, R. et al., 2009. Metabolic syndrome in bipolar disorders in Spain: Findings from the population-based case-control BIMET-VIVA study. *European Neuropsychopharmacology*, 19((Gabriel-Sánchez R.; Lorenzo-Carrascosa L.; Alonso-Arroyo M.) Clinical Epidemiology Division and RECAVA Network, Hospital Universitario La Paz, Madrid, Spain), p.S466. Available at: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L70044753%5Cnhttp://dx.doi.org/10.1016/S0924-977X(09)70730-

8%5Cnhttp://sfx.unimi.it:9003/unimi?sid=EMBASE&issn=0924977X&id=doi:10.1016%2FS0924-977X%2809%2970730-8&atitle=Metabolic+sy.

Gálvez, J. a et al., 2014. Visual analytical tool for evaluation of 10-year perioperative transfusion practice at a children's hospital. *Journal of the American Medical Informatics Association*, 21(3), pp.529–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24363319.

Gandomi, A. & Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), pp.137–144.

Garofalakis, M. & Rastogi, R., 1999. SPIRIT: Sequential pattern mining with regular expression constraints. … *of the International Conference on Very* …, pp.223–234. Available at: http://citeseer.ist.psu.edu/247008.html%5Cnpapers2://publication/uuid/04E1DB4D-FEB0-41DC-9E63-522B5BE8B147.

van Gemert-Pijnen, J.E.W.C. et al., 2011. A holistic framework to improve the uptake and impact of eHealth technologies. *Journal of medical Internet research*, 13(4).

Georga, E. et al., 2009. Data mining for blood glucose prediction and knowledge discovery in diabetic patients: The METABO diabetes modeling and management system. In *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*. pp. 5633–5636.

Goldstein, A. & Shahar, Y., 2016. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data. *Journal of Biomedical Informatics*, 61, pp.159–175.

Gorina, Y. & Kramarow, E.A., 2011. Identifying chronic conditions in medicare claims data: Evaluating the chronic condition data warehouse algorithm. *Health Services Research*, 46(5), pp.1610–1627.

Gotz, D. et al., 2012. ICDA: a platform for Intelligent Care Delivery Analytics. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012(1), pp.264–73. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3540495&tool=pmcentrez&rendertype=abstract.

Gotz, D., Wang, F. & Perer, A., 2014. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of Biomedical Informatics*, 48, pp.148–159.

Guillén, A. et al., 2011. METABO: A new paradigm towards diabetes disease management. An innovative business model. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. pp. 3554–3557.

Halamka, J.D., 2014. Early Experiences With Big Data At An Academic Medical Center. *Health Affairs*, 33(7), pp.1132–1138. Available at: http://content.healthaffairs.org/cgi/doi/10.1377/hlthaff.2014.0031%5Cnhttp://content.healthaffairs.org/content/33/7/1132.abstract.

Hall, M.A. et al., 2014. Environment-wide association study (ewas) for type 2 diabetes in the marshfield personalized medicine research project biobank. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 19, pp.200–11.

Halpern, Y. et al., 2016. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association : JAMIA*, p.ocw011. Available at: http://jamia.oxfordjournals.org/content/early/2016/04/26/jamia.ocw011.abstract.

Hamed, K.H. & Ramachandra Rao, A., 1998. A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, 204(1–4), pp.182–196.

Harper, E., 2014. Can big data transform electronic health records into learning health systems? In *Studies in Health Technology and Informatics*. pp. 470–475.

Hauskrecht, M. et al., 2013. Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics*, 46(1), pp.47–55.

Häussler, B. et al., 2007. Risk assessment in diabetes management: how do general practitioners estimate risks due to diabetes? *Quality & safety in health care*, 16(3), pp.208–12. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-34250867439&partnerID=tZOtx3y1.

Heselmans, A. et al., 2013. Feasibility and impact of an evidence-based electronic decision support system for diabetes care in family medicine: protocol for a cluster randomized controlled trial. *Implementation Science: IS*, 8(1), p.83.

Himes, B.E. et al., 2008. Characterization of patients who suffer asthma exacerbations using data extracted from electronic medical records. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pp.308–312.

Hippisley-Cox, J. et al., 2010. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *Bmj*, 341, p.c6624. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21148212%5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC2999889/pdf/bmj.c6624.pdf.

Holbrook, A. et al., 2009. Individualized electronic decision support and reminders to improve diabetes care in the community: COMPETE II randomized trial. *CMAJ*, 181(1–2), pp.37–44.

Hope, W.W. et al., 2013. Software for dosage individualization of voriconazole for immunocompromised patients. *Antimicrobial Agents and Chemotherapy*, 57(4), pp.1888–1894.

Höppner, F. & Klawonn, F., 2001. Finding informative rules in interval sequences. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 125–134.

Hovenga, E.J.S. & Grain, H., 2013. Health Data and Data Governance. *Studies In Health Technology And Informatics*, 193, pp.67–92. Available at: http://ezproxy.library.uvic.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=mnh&AN=24018511&site=ehost-live&scope=site.

Hripcsak, G. & Albers, D.J., 2012. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, pp.117–121.

Hripcsak, G. & Albers, D.J., 2013. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 20(1), pp.117–21. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3555337&tool=pmcentrez&rendertype=abstract.

Hripcsak, G., Albers, D.J. & Perotte, A., 2015. Parameterizing time in electronic health record studies. *Journal of the American Medical Informatics Association*, 18(Suppl 1), pp.i109–i115.

Huang, B., Zhu, P. & Wu, C., 2012. Customer-centered careflow modeling based on guidelines. *Journal of Medical Systems*, 36(5), pp.3307–3319.

Huang, Y. et al., 2007. Feature selection and classification model construction on type 2 diabetic patients' data. *Artificial Intelligence in Medicine*, 41(3), pp.251–262.

Huang, Z. et al., 2014. Discovery of clinical pathway patterns from event logs using probabilistic topic models. *Journal of Biomedical Informatics*, 47, pp.39–57.

Huang, Z. et al., 2013. Summarizing clinical pathways from event logs. *Journal of Biomedical Informatics*, 46(1), pp.111–127.

Huang, Z., Lu, X. & Duan, H., 2012. On mining clinical pathway patterns from medical behaviors.

Artificial Intelligence in Medicine, 56(1), pp.35–50.

IBM, 2013. Wrangling big data: Fundamentals of data lifecycle management. *IBM Managing data lifecycle*.

Israel, O., Sconfienza, L.M. & Lipsky, B.A., 2014. Diagnosing diabetic foot infection: the role of imaging and a proposed flow chart for assessment. *The quarterly journal of nuclear medicine and molecular imaging : official publication of the Italian Association of Nuclear Medicine (AIMN) [and] the International Association of Radiopharmacology (IAR), [and] Section of the Society of* Radiopharmaceutica, 58(1), pp.33–45.

Jaccard, P., 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37(JANUARY 1901), pp.547–579.

Jagadeesh Chandra Bose, R.P. & Van Der Aalst, W.M.P., 2012. Process diagnostics using trace alignment: Opportunities, issues, and challenges. *Information Systems*, 37(2), pp.117–141.

Janghorbani, M. & Momeni, F., 2014. Systematic review and metaanalysis of air pollution exposure and risk of diabetes. *Eur J Epidemiol*.

Jensen, P.B., Jensen, L.J. & Brunak, S., 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6), pp.395–405. Available at: http://dx.doi.org/10.1038/nrg3208.

Juarez, J.M. et al., 2015. Spatiotemporal data visualisation for homecare monitoring of elderly people. *Artificial Intelligence in Medicine*, 65(2), pp.97–111.

Kahn, M.G. & Ranade, D., 2010. The impact of electronic medical records data sources on an adverse drug event quality measure. *Journal of the American Medical Informatics Association : JAMIA*, 17(2), pp.185–91. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-77953161770&partnerID=tZOtx3y1.

Kaltoft, M.K. et al., 2014. Enhancing informatics competency under uncertainty at the point of decision: A knowing about knowing vision. In *Studies in Health Technology and Informatics*. pp. 975–979.

Kam, P. & Fu, A.W., 2000. Discovering Temporal Patterns for Interval-based Events. *Lecture Notes in Computer Science*, 1874(5), pp.317–326. Available at: http://www.springerlink.com/index/b5eg19hu445vx0q7.pdf.

Kammoun, F. & Ayed, M. Ben, 2014. Clinical Dynamic Decision Support System based on temporal association rules. *2nd Middle East Conference on Biomedical Engineering*, pp.289–292.

Kaneko, Y. et al., 2013. The search for common pathways underlying asthma and COPD. *International Journal of COPD*, 8, pp.65–78.

Kannel, W.B., Hjortland, M. & Castelli, W.P., 1974. Role of diabetes in congestive heart failure: The Framingham study. *The American Journal of Cardiology*, 34(1), pp.29–34.

Kannel, W.B. & McGee, D.L., 1979. Diabetes and glucose tolerance as risk factors for cardiovascular disease: The Framingham study. *Diabetes Care*, 2(2), pp.120–126.

Katal, A., Wazid, M. & Goudar, R.H., 2013. Big data: Issues, challenges, tools and Good practices. In *2013 6th International Conference on Contemporary Computing, IC3 2013*. pp. 404–409.

Keim, D.A. et al., 2008. Visual analytics: Scope and challenges. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 76–90.

Kho, A.N. et al., 2011. Electronic medical records for genetic research: results of the eMERGE consortium. *Science translational medicine*, 3(79), p.79re1. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3690272&tool=pmcentrez&rendertype=abstract.

Kohane, I.S., Churchill, S.E. & Murphy, S.N., 2012. A translational engine at the national scale: informatics for integrating biology and the bedside. *Journal of the American Medical Informatics Association*, 19(2), pp.181–185.

Kohn, M.S. et al., 2014. IBM's Health Analytics and Clinical Decision Support. *Yearbook of medical informatics*, 9, pp.154–62. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4287097&tool=pmcentrez&rendertype=abstract.

Krumholz, H.M., 2014. Big data and new knowledge in medicine: The thinking , training , and tools needed for a learning health system. *Health Affairs*, 33(7), pp.1163–1170.

Kumar, S.J.J. & Madheswaran, M., 2012. An improved medical decision support system to identify the diabetic retinopathy using fundus images. In *Journal of Medical Systems*. pp. 3573–3581.

Last Mark, Tosas Olga, Cassarino Tiziano Gallo, Kozlakidis Zisis, E.J., 2016. Evolving classification of intensive care patients from event data. *Artificial Intelligence in Medicine*, 69, pp.22–32. Available at: http://dx.doi.org/10.1016/j.artmed.2016.04.001.

Leemans, M. & van der Aalst, W.M.P., 2015. Discovery of frequent episodes in event logs. In *Lecture Notes in Business Information Processing*. pp. 1–31.

Lewis, S.N. et al., 2011. Prediction of disease and phenotype associations from Genome-Wide association studies. *PLoS ONE*, 6(11).

Li, X. et al., 2015. Analysis of Care Pathway Variation Patterns in Patient Records. In *Studies in Health Technology and Informatics*. pp. 692–696.

Lim, S. et al., 2011. Improved glycemic control without hypoglycemia in elderly diabetic patients using the ubiquitous healthcare service, a new medical information system. *Diabetes Care*, 34(2), pp.308–313.

Lin, T.T.-L. et al., 2015. The Effect of Diabetes, Hyperlipidemia, and Statins on the Development of Rotator Cuff Disease: A Nationwide, 11-Year, Longitudinal, Population-Based Follow-up Study. *The American Journal of Sports Medicine*.

Lin, Y.K., Chen, H. & Brown, R.A., 2013. MedTime: A temporal information extraction system for clinical narratives. *Journal of Biomedical Informatics*, 46(SUPPL.).

Linder, J. a et al., 2010. Electronic health record feedback to improve antibiotic prescribing for acute respiratory infections. *The American journal of managed care*, 16(12 Suppl HIT), pp.e311-9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21322301.

Lipton, J.A. et al., 2010. Evaluation of a clinical decision support system for glucose control: impact of protocol modifications on compliance and achievement of glycemic targets. *Critical Pathways in Cardiology: A Journal of Evidence-Based Medicine*, 9(3), pp.140–147. Available at: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=medl&AN=20802267.

Liu, C. et al., 2015. Temporal Phenotyping from Longitudinal Electronic Health Records. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, pp.705–714. Available at: http://dl.acm.org/citation.cfm?id=2783258.2783352.

Liu, H. et al., 2013. An Efficacy Driven Approach for Medication Recommendation in Type 2 Diabetes Treatment Using Data Mining Techniques. *Medinfo 2013: Proceedings of the 14th World Congress on Medical and Health Informatics, Pts 1 and 2*, 192, p.1071.

Liu, H. et al., 2015. Synthesizing Analytic Evidence to Refine Care Pathways. In *Studies in Health Technology and Informatics*. pp. 70–74.

Liu, H., Mei, J. & Xie, G., 2012. Towards collaborative chronic care using a clinical guideline-based

decision support system. In *Studies in Health Technology and Informatics*. pp. 492–496.

Liu, Z. & Hauskrecht, M., 2015. Clinical time series prediction: Toward a hierarchical dynamical system framework. *Artificial Intelligence in Medicine*, 65(1), pp.5–18.

Lo-Ciganic, W.-H. et al., 2015. Using machine learning to examine medication adherence thresholds and risk of hospitalization. *Medical care*, 53(8), pp.720–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26147866.

Lupse, O.S. et al., 2014. Supporting diagnosis and treatment in medical care based on Big Data processing. *Studies in health technology and informatics; Stud.Health Technol.Inform.*, pp.65–69.

Mandel, J.C. et al., 2016. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, pp.1–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26911829.

Mandl, K.D. et al., 2012. The SMART Platform: early experience enabling substitutable applications for electronic health records. *Journal of the American Medical Informatics Association*, 19(4), pp.597–603.

Mane, K.K. et al., 2012. VisualDecisionLinc: A visual analytics approach for comparative effectiveness-based clinical decision support in psychiatry. *Journal of Biomedical Informatics*, 45(1), pp.101–106.

Mannila, H., Toivonen, H. & Verkamo, a. I., 1995. Discovering Frequent Episodes in Sequences. *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pp.210–215. Available at: http://ukpmc.ac.uk/abstract/CIT/637423.

Mans, R.S. et al., 2015. Process mining in healthcare: Data challenges when answering frequently posed questions. *Methods in Molecular Biology*, 1246(Data Mining in Clinical Medicine).

Mathers, C.D. & Loncar, D., 2006. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Medicine*, 3(11), pp.2011–2030.

McCarty, C.A. et al., 2011. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC medical genomics*, 4(1), p.13. Available at: http://www.biomedcentral.com/1755-8794/4/13.

McGlynn, E.A. et al., 2014. Developing a data infrastructure for a learning health system: the PORTAL network. *Journal of the American Medical Informatics Association : JAMIA*, 21, pp.596–601. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24821738.

Meystre, S.M., Deshmukh, V.G. & Mitchell, J., 2009. A clinical use case to evaluate the i2b2 Hive: predicting asthma exacerbations. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2009, pp.442–446.

Mick, J., 2011. Data-Driven Decision Making. *JONA: The Journal of Nursing Administration*, 41(10), pp.391–393.

Mitsa, T., 2010. *Temporal Data Mining*, Available at: http://www.ncbi.nlm.nih.gov/pubmed/18194720.

Mitsch, C. et al., 2016. Clinical Decision Support for the Classification of Diabetic Retinopathy : A Comparison.

Moghimi, F.H., Cheung, M. & Wickramasinghe, N., 2013. Applying predictive analytics to develop an intelligent risk detection application for healthcare contexts. In *Studies in Health Technology and Informatics*. p. 926.

Möller, K., 2013. Lifecycle models of data-centric systems and domains. *Semantic Web*, 4(1), pp.67–88.

Moskovitch, R. & Shahar, Y., 2015. Classification-driven temporal discretization of multivariate time series. *Data Mining and Knowledge Discovery*, 29(4), pp.871–913.

Moskovitch, R. & Shahar, Y., 2009. Medical Temporal-Knowledge Discovery via Temporal Abstraction. *AMIA Annual Symposium proceedings AMIA Symposium AMIA Symposium*, 2009, pp.452–456. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2815492&tool=pmcentrez&rendertype=abstract.

Mould, D.R. & Dubinsky, M.C., 2015. Dashboard systems: Pharmacokinetic/pharmacodynamic mediated dose optimization for monoclonal antibodies. *Journal of Clinical Pharmacology*, 55(S3), pp.S51–S59.

Mouttham, A., Peyton, L. & Kuziemsky, C., 2011. Leveraging performance analytics to improve integration of care. In *Proceeding of the 3rd workshop on Software engineering in health care - SEHC '11*. p. 56. Available at: http://dl.acm.org/citation.cfm?id=1987993.1988005.

Murdoch, T.B. & Detsky, A.S., 2013. The inevitable application of big data to health care. *Jama*, 309(13), pp.1351–1352. Available at: http://dx.doi.org/10.1001/jama.2013.393%5Cnhttp://jama.jamanetwork.com/data/Journals/JAMA/926712/jvp130007_1351_1352.pdf.

Murphy, S.N. et al., 2010. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association : JAMIA*, 17(2), pp.124–130.

Nadkarni, G.N. et al., 2014. Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2014, pp.907–16. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4419875&tool=pmcentrez&rendertype=abstract.

Nelson, D.L. & Cox, M.M., 2013. *Lehninger Principles of Biochemistry 6th ed.*,

Neubauer, K.M. et al., 2015. Standardized Glycemic Management with a Computerized Workflow and Decision Support System for Hospitalized Patients with Type 2 Diabetes on Different Wards. *Diabetes Technology & Therapeutics*, 17(10), pp.685–692. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26355756.

Newton, K.M. et al., 2013. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc*, 20(e1), pp.e147-54. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23531748.

Ng, K. et al., 2014. PARAMO: A PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. *Journal of Biomedical Informatics*, 48, pp.160–170.

Nichols, G.A. et al., 2015. Impact on glycated haemoglobin of a biological response-based measure of medication adherence. *Diabetes, Obesity & Metabolism*.

Nikfarjam, A., Emadzadeh, E. & Gonzalez, G., 2013. Towards generating a patient's timeline: Extracting temporal relationships from clinical notes. *Journal of Biomedical Informatics*, 46(SUPPL.).

Nodelman, U., Shelton, C.R. & Koller, D., 2002. Continuous time Bayesian networks. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, (June), pp.378–387. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.1813&rep=rep1&type=pdf.

O'Connor, P.J. et al., 2011. Diabetes performance measures: Current status and future directions. In *Diabetes Care*. pp. 1651–1659.

O'Connor, P.J. et al., 2011. Impact of electronic health record clinical decision support on diabetes care: a randomized trial. *Annals of family medicine*, 9(1), pp.12–21. Available at: http://www.sciencedirect.com/science/article/pii/S1544170911600036.

O'Reilly, D. et al., 2012. Cost-effectiveness of a shared computerized decision support system for diabetes

linked to electronic medical records. *Journal of the American Medical Informatics Association*, 19(3), pp.341–345.

Ola, O. & Sedig, K., 2014. The challenge of big data in public health: an opportunity for visual analytics. *Online Journal of Public Health Informatics*, 5(3), pp.e223, 1–21. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3959916&tool=pmcentrez&rendertype=abstract.

den Ouden, H. et al., 2015. Shared decision making in type 2 diabetes with a support decision tool that takes into account clinical factors, the intensity of treatment and patient preferences: design of a cluster randomised (OPTIMAL) trial. *BMC family practice*, 16(1), p.27. Available at: http://www.biomedcentral.com/1471-2296/16/27.

Pacheco, J.A., Thompson, W. & Kho, A., 2011. Automatically detecting problem list omissions of type 2 diabetes cases using electronic medical records. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011, pp.1062–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243294&tool=pmcentrez&rendertype=abstract.

Palacios, B.C.-S.C.M.M.J., 2016. Development of a clinical decision support system for antibiotic management in a hospital environment. *Progress in Artificial Intelligence*, 5(3), pp.181–197.

Palmer, A.J. et al., 2004. The CORE Diabetes Model: Projecting long-term clinical outcomes, costs and cost-effectiveness of interventions in diabetes mellitus (types 1 and 2) to support clinical and reimbursement decision-making. *Current medical research and opinion*, 20 Suppl 1, pp.S5–S26.

Palmieri, L. et al., 2004. Evaluation of the global cardiovascular absolute risk: the Progetto CUORE individual score. *Annali dell'Istituto superiore di sanita*, 40(4), pp.393–399.

Panzarasa, S. et al., 2004. A careflow management system for chronic patients. *Studies in Health Technology and Informatics*, 107, pp.673–677.

Panzarasa, S. et al., 2002. Evidence-based careflow management systems: The case of post-stroke rehabilitation. *Journal of Biomedical Informatics*, 35(2), pp.123–139.

Park, S.K. et al., 2015. Long-Term Exposure to Air Pollution and Type 2 Diabetes Mellitus in a Multiethnic Cohort. *American Journal of Epidemiology*, 181(5), pp.327–336. Available at: http://aje.oxfordjournals.org/cgi/doi/10.1093/aje/kwu280.

Parker, R.F. et al., 2014. The effect of using a shared electronic health record on quality of care in people with type 2 diabetes. *J Diabetes Sci Technol*, 8(5), pp.1064–1065. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24876452.

Parsons, a. et al., 2012. Validity of electronic health record-derived quality measurement for performance monitoring. *Journal of the American Medical Informatics Association*, 19(4), pp.604–609.

Patel, D., Hsu, W. & Lee, M.L., 2008. Mining relationships among interval-based events for classification. *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp.393–404. Available at: http://doi.acm.org/10.1145/1376616.1376658.

Pathak, J., Kho, A.N. & Denny, J.C., 2013. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association : JAMIA*, 20(December), pp.e206-11. Available at: http://jamia.oxfordjournals.org/content/20/e2/e206.abstract.

Peek, N., Holmes, J.H. & Sun, J., 2014. Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics. *Yearbook of medical informatics*, 9(1), pp.42–47.

Peissig, P.L. et al., 2014. Relational machine learning for electronic health record-driven phenotyping. *Journal of Biomedical Informatics*, 52, pp.260–270.

Peleg, M. et al., 2008. Lessons learned from adapting a generic narrative diabetic-foot guideline to an institutional decision-support system. In *Studies in Health Technology and Informatics*. pp. 243–252.

Penno, G. et al., 2013. HbA1c variability as an independent correlate of nephropathy, but not retinopathy, in patients with type 2 diabetes: The renal insufficiency and cardiovascular events (RIACE) Italian Multicenter Study. *Diabetes Care*, 36(8), pp.2301–2310.

Perer, A., Wang, F. & Hu, J., 2015. Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of Biomedical Informatics*, 56, pp.369–378.

Peters, S.G. & Buntrock, J.D., 2014. Big data and the electronic health record. *The Journal of ambulatory care management*, 37(3), pp.206–210.

Peters, S.G. & Khan, M.A., 2014. Electronic health records: current and future use. *Journal of comparative effectiveness research*, 3(5), pp.515–522.

Peterson, A.M. et al., 2007. A checklist for medication compliance and persistence studies using retrospective databases. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, 10(1), pp.3–12.

Pivovarov, R. et al., 2014. Temporal trends of hemoglobin A1c testing. *Journal of the American Medical Informatics Association : JAMIA*, pp.1–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24928176.

Post, A.R., Kurc, T., Willard, R., et al., 2013. Temporal abstraction-based clinical phenotyping with Eureka! *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2013, pp.1160–1169. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84901280296&partnerID=tZOtx3y1.

Post, A.R., Kurc, T., Cholleti, S., et al., 2013. The Analytic Information Warehouse (AIW): A platform for analytics using electronic health record data. *Journal of Biomedical Informatics*, 46(3), pp.410–424.

Post, A.R. & Harrison, J.H., 2007a. PROTEMPA: A Method for Specifying and Identifying Temporal Sequences in Retrospective Data for Patient Selection. *Journal of the American Medical Informatics Association*, 14, pp.674–683.

Post, A.R. & Harrison, J.H., 2007b. PROTEMPA: A Method for Specifying and Identifying Temporal Sequences in Retrospective Data for Patient Selection. *Journal of the American Medical Informatics Association*, 14(5), pp.674–683.

Post, A.R. & Harrison, J.H., 2008. Temporal Data Mining. *Clinics in Laboratory Medicine*, 28(1), pp.83–100.

Quaglini, S. et al., 2001. Flexible guideline-based patient careflow systems. *Artificial Intelligence in Medicine*, 22(1), pp.65–80.

Quaglini, S. et al., 2000. Guideline-based careflow systems. *Artificial Intelligence in Medicine*, 20(1), pp.5–22.

R.a, A., D.a, G. & F.b, L., 1998. Mining process models from workflow logs. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1377 LNCS, pp.469–483. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84890470685&partnerID=40&md5=6b1646f986d196c756dad381edee33a4.

Rachel, R. & Michelle, S., 2014. Electronic Health Records-Based Phenotyping. *Rethinking Clinical Trials*. Available at: http://sites.duke.edu/rethinkingclinicaltrials/ehr-phenotyping/.

Raghupathi, W. & Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2, p.3. Available at: http://www.hissjournal.com/content/2/1/3.

Rajagopalan, S. & Brook, R.D., 2012. Air Pollution and Type 2 Diabetes Mechanistic Insights. *Diabetes*, 61(December), pp.3037–3045.

Ramyachitra, D. & Manikandan, P., 2014. Imbalanced Dataset Classification and Solutions: a Review. *International Journal of Computing and Business Research (IJCBR)*, 5(4). Available at: http://www.researchmanuscripts.com/July2014/2.pdf.

Rasmussen, L. V et al., 2015. A Modular Architecture for Electronic Health {Record-Driven} Phenotyping. *AMIA Jt Summits Transl Sci Proc*, 2015, pp.147–151.

Rebuge, Á. & Ferreira, D.R., 2012. Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, 37(2), pp.99–116.

Resetar, E. et al., 2005. Customizing a commercial rule base for detecting drug-drug interactions. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, p.1094.

Retnakaran, R. et al., 2006. Risk factors for renal dysfunction in type 2 diabetes: U.K. Prospective Diabetes Study 74. *Diabetes*, 55(6), pp.1832–1839.

Reza, A.W. & Eswaran, C., 2011. A decision support system for automatic screening of non-proliferative diabetic retinopathy. *Journal of Medical Systems*, 35(1), pp.17–24.

Riazi, H. et al., 2016. Conceptual Framework for Developing a Diabetes Information Network. *Acta Informatica Medica*, 24(3), p.186. Available at: http://www.scopemed.org/?mno=231293.

Richesson, R.L., Rusincovitch, S.A., et al., 2013. A comparison of phenotype definitions for diabetes mellitus. *Journal of the American Medical Informatics Association : JAMIA*, 20(e2), pp.e319-26. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3861928&tool=pmcentrez&rendertype=abstract.

Richesson, R.L., Hammond, W.E., et al., 2013. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *Journal of the American Medical Informatics Association : JAMIA*, 20(e2), pp.e226-31. Available at: http://jamia.bmj.com/content/20/e2/e226.full.

Robinson, P.N., 2012. Deep phenotyping for precision medicine. *Human Mutation*, 33(5), pp.777–780.

Rodbard, D. & Vigersky, R. a, 2011. Design of a decision support system to help clinicians manage glycemia in patients with type 2 diabetes mellitus. *Journal of diabetes science and technology*, 5(2), pp.402–11. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125935&tool=pmcentrez&rendertype=abstract.

Roisin RR, 2016. Chronic Obstructive Pulmonary Disease Updated 2010 Global Initiative for Chronic Obstructive Lung Disease. *Global Initiative for Chronic Obstructive Lung Disease. Inc*, pp.1–94.

Rojas, E. et al., 2016. Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*, 61, pp.224–236.

Rothman, K.J., Greenland, S. & Associate, T.L.L., 2014. Modern Epidemiology, 3rd Edition. *The Hastings Center report*, 44 Suppl 2, p.insidebackcover.

Rumsfeld, J.S., Joynt, K.E. & Maddox, T.M., 2016. Big data analytics to improve cardiovascular care: promise and challenges. *Nature reviews. Cardiology*.

Rusincovitch, S.A. et al., 2013. Framework for Curating and Applying Data Elements within Continuing Use Data: A Case Study from the Durham Diabetes Coalition. *AMIA Jt Summits Transl Sci Proc*, 2013(April), p.228. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24303271.

Russell, K.G. & Rosenzweig, J., 2007. Improving outcomes for patients with diabetes using Joslin Diabetes Center's Registry and Risk Stratification system. *Journal of healthcare information management : JHIM*, 21(2), pp.26–33.

Sacchi, L., Larizza, C., Combi, C., et al., 2007. Data mining with Temporal Abstractions: Learning rules from time series. *Data Mining and Knowledge Discovery*, 15(2), pp.217–247.

Sacchi, L. et al., 2015a. JTSA: An open source framework for time series abstractions. *Computer Methods and Programs in Biomedicine*, 121(3), pp.175–188. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0169260715001492.

Sacchi, L. et al., 2015b. JTSA: An open source framework for time series abstractions. *Computer Methods and Programs in Biomedicine*, 121(3), pp.175–188.

Sacchi, L. et al., 2015c. JTSA: An open source framework for time series abstractions. *Computer Methods and Programs in Biomedicine*, pp.1–14. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0169260715001492.

Sacchi, L., Larizza, C., Magni, P., et al., 2007. Precedence Temporal Networks to represent temporal relationships in gene expression data. *Journal of Biomedical Informatics*, 40(6), pp.761–774.

Sacchi, L., Dagliati, A. & Bellazzi, R., 2015. Analyzing Complex Patients Temporal Histories: New Frontiers in Temporal Data Mining. In *Data Mining in Clinical Medicine, Methods in Molecular Biology*.

Sáenz, A. et al., 2012. Development and validation of a computer application to aid the physician's decision-making process at the start of and during treatment with insulin in type 2 diabetes: a randomized and controlled trial. *Journal of diabetes science and technology*, 6(3), pp.581–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3440036&tool=pmcentrez&rendertype=abstract.

Sanchez, F.M. et al., 2014. Exposome informatics : considerations for the design of future biomedical research information systems. *J Am Med Inform Assoc*, 21, pp.386–390.

Sarkar, I.N. et al., 2011. Translational bioinformatics: linking knowledge across biological and clinical realms. *Journal of the American Medical Informatics Association : JAMIA*, 18(4), pp.354–357.

Savova, G. et al., 2009. Towards temporal relation discovery from the clinical narrative. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2009(February 2016), pp.568–572.

Schoen, D.E., Glance, D.G. & Thompson, S.C., 2015. Clinical decision support software for diabetic foot risk stratification: development and formative evaluation. *Journal of Foot and Ankle Research*, 8, p.73. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4676878/.

Segagni, D. et al., 2015. Improving Clinical Decisions on T2DM Patients Integrating Clinical, Administrative and Environmental Data. In *Studies in Health Technology and Informatics*. pp. 682–686.

Segagni, D. et al., 2011. The ONCO-I2b2 project: Integrating biobank information and clinical data to support translational research in oncology. In *Studies in Health Technology and Informatics*. pp. 887–891.

Shahar, Y., 1997. A framework for knowledge-based temporal abstraction. *Artificial Intelligence*, 90(1–2), pp.79–133.

Shahar, Y., 1999. Timing is everything: Temporal reasoning and temporal data maintenance in medicine. *Artificial Intelligence in Medicine*, pp.30–46. Available at: http://dx.doi.org/10.1007/3-540-48720-4_3.

Shalom, E., Shahar, Y. & Lunenfeld, E., 2016. An architecture for a continuous, user-driven, and data-driven application of clinical guidelines and its evaluation. *Journal of Biomedical Informatics*, 59, pp.130–148.

Shaw, J.E., Sicree, R.A. & Zimmet, P.Z., 2010. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Research and Clinical Practice*, 87(1), pp.4–14.

Shivade, C. et al., 2014. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2), pp.221–230. Available at:

http://jamia.oxfordjournals.org/lookup/doi/10.1136/amiajnl-2013-001935%5Cnhttp://jamia.bmj.com/content/early/2013/11/07/amiajnl-2013-001935.long%5Cn%3CGo to ISI%3E://WOS:000331263600007.

Sim, I. et al., 2001. Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association : JAMIA*, 8(6), pp.527–534. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=130063&tool=pmcentrez&rendertype=abstract.

Simms, R.A. et al., 2013. Development of maternity dashboards across a UK health region; Current practice, continuing problems. *European Journal of Obstetrics Gynecology and Reproductive Biology*, 170(1), pp.119–124.

Simon-Tuval, T. et al., 2015. The association between adherence to cardiovascular medications and healthcare utilization. *The European journal of health economics: HEPAC: health economics in prevention and care*.

Simpao, A.F. et al., 2014. Optimization of drug-drug interaction alert rules in a pediatric hospital's electronic health record system using a visual analytics dashboard. *Journal of the American Medical Informatics Association*, 22(2), pp.361–9. Available at: http://jamia.oxfordjournals.org/cgi/doi/10.1136/amiajnl-2013-002538%5Cnhttp://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=25318641&retmode=ref&cmd=prlinks.

Simpao, A.F., Ahumada, L.M. & Rehman, M.A., 2015. Big data and visual analytics in anaesthesia and health care. *British Journal of Anaesthesia*, 115(3), pp.350–356.

Singh, A. et al., 2015. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *Journal of Biomedical Informatics*, 53, pp.220–228.

Slonim, N. et al., 2012. Knowledge-analytics synergy in clinical decision support. In *Studies in Health Technology and Informatics*. pp. 703–707.

Smith, M.C. & Wrobel, J.P., 2014. Epidemiology and clinical impact of major comorbidities in patients with COPD. *Int J Chron Obstruct Pulmon Dis*, 9, pp.871–888.

Solti, I. et al., 2008. Natural language processing of clinical trial announcements: exploratory-study of building an automated screening application. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, p.1142.

Sprague, A.E. et al., 2013. Measuring quality in maternal-newborn care: developing a clinical dashboard. *J Obstet Gynaecol Can*, 35(1), pp.29–38. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23343794.

Spratt, S.E. et al., 2015. Methods and initial findings from the Durham Diabetes Coalition: Integrating geospatial health technology and community interventions to reduce death and disability. *Journal of Clinical and Translational Endocrinology*, 2(1), pp.26–36.

Srinivasan Suresh, 2014. Big Data and Predictive Analytics Applications in the Care of Children. *IT Professional*, 16(1), pp.13–15. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6756866.

Stacey, M. & McGregor, C., 2007. Temporal abstraction in intelligent clinical data analysis: A survey. *Artificial Intelligence in Medicine*, 39(1), pp.1–24. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17011175.

Steiner, J.F. & Prochazka, A. V., 1997. The assessment of refill compliance using pharmacy records: Methods, validity, and applications. *Journal of Clinical Epidemiology*, 50(1), pp.105–116.

Stekhoven, D.J. & B??hlmann, P., 2012. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), pp.112–118.

Stella, F. & Amer, Y., 2012. Continuous time Bayesian network classifiers. *Journal of Biomedical Informatics*, 45(6), pp.1108–1119.

Stratton, I.M. et al., 2000. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. *BMJ (Clinical research ed.)*, 321(7258), pp.405–412.

Stratton, I.M. et al., 2001. UKPDS 50: risk factors for incidence and progression of retinopathy in Type II diabetes over 6 years from diagnosis. *Diabetologia*, 44(2), pp.156–163.

Sun, W., Rumshisky, A. & Uzuner, O., 2013. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46(SUPPL.).

Tamayo, T. et al., 2014. Is particle pollution in outdoor air associated with metabolic control in type 2 diabetes? *PLoS ONE*, 9(3).

Tan, X. et al., 2010. [Computer-assisted screening system for individualized treatment of type 2 diabetes mellitus]. *Nan fang yi ke da xue xue bao = Journal of Southern Medical University*, 30(9), pp.2134–2138.

Tenenbaum, J.D. et al., 2016. An informatics research agenda to support precision medicine: seven key areas. *Journal of the American Medical Informatics Association*, 1(919), p.ocv213. Available at: http://jamia.oxfordjournals.org/lookup/doi/10.1093/jamia/ocv213.

Thiering, E. & Heinrich, J., 2015. Epidemiology of air pollution and diabetes. *Trends in Endocrinology & Metabolism*, pp.1–11. Available at: http://dx.doi.org/10.1016/j.tem.2015.05.002.

Thomas, J. & Kielman, J., 2009. Challenges for visual analytics. *Information Visualization*, 8(4), pp.309–314. Available at: http://ivi.sagepub.com/content/8/4/309.

Tomasallo, C.D. et al., 2014. Estimating Wisconsin asthma prevalence using clinical electronic health records and public health data. *American Journal of Public Health*, 104(1).

Toussi, M. et al., 2008. An automated method for analyzing adherence to therapeutic guidelines: application in diabetes. *Studies in health technology and informatics*, 136, pp.339–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18487754.

Triplitt, C.L., 2012. Examining the mechanisms of glucose regulation. *The American journal of managed care*, 18(1 Suppl), pp.S4-10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22559855.

Tsumoto, S. et al., 2014. Similarity-based behavior and process mining of medical practices. *Future Generation Computer Systems*, 33, pp.21–31.

Tu, S.W. et al., 2011. A practical method for transforming free-text eligibility criteria into computable criteria. *Journal of Biomedical Informatics*, 44(2), pp.239–250.

Tuomilehto, J. et al., 2001. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *The New England journal of medicine*, 344(18), pp.1343–50. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11333990.

Vaitsis, C., Nilsson, G. & Zary, N., 2014. Big Data in Medical Informatics: Improving Education Through Visual Analytics. In *Studies in Health Technology and Informatics*. pp. 1163–1167.

Van Velsen, L., Wentzel, J. & Van Gemert-Pijnen, J.E.W.C., 2013. Designing ehealth that matters via a multidisciplinary requirements development approach. *Journal of Medical Internet Research*, 15(6).

Verbeek, H.M.W. et al., 2006. Interoperability in the ProM framework. In *CEUR Workshop Proceedings*.

Verduijn, M. et al., 2007. Temporal abstraction for feature extraction: A comparative case study in prediction from intensive care monitoring data. *Artificial Intelligence in Medicine*, 41(1), pp.1–12.

de Vries, F.M. et al., 2015. Does a cardiovascular event change adherence to statin treatment in patients

with type 2 diabetes? A matched cohort design. *Current Medical Research and Opinion*, 31(4), pp.595–602.

Vrijens, B. et al., 2012. A new taxonomy for describing and defining adherence to medications. *British Journal of Clinical Pharmacology*, 73(5), pp.691–705.

Wagholikar, K.B. et al., 2016. SMART-on-FHIR implemented over i2b2. *Journal of the American Medical Informatics Association : JAMIA*, pp.1–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27274012.

Waitman, L.R. et al., 2011. Adopting real-time surveillance dashboards as a component of an enterprisewide medication safety strategy. *Joint Commission Journal on Quality and Patient Safety*, 37(7), pp.326–332.

Wang, C.-J. et al., 2013. Bioinformatics method to analyze the mechanism of pancreatic cancer disorder. *Journal of computational biology : a journal of computational molecular cell biology*, 20(6), pp.444–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23614574.

Wang, F. et al., 2012. Towards heterogeneous temporal clinical event pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*. p. 453. Available at: http://dl.acm.org/citation.cfm?doid=2339530.2339605.

Wang, L.-Y. et al., 2015. Time-dependent variation of pathways and networks in a 24-hour window after cerebral ischemia-reperfusion injury. *BMC Systems Biology*, 9(1), pp.1–11. Available at: http://www.biomedcentral.com/1752-0509/9/11.

Warner, J.L. et al., 2016. Classification of hospital acquired complications using temporal clinical information from a large electronic health record. *Journal of Biomedical Informatics*, 59, pp.209–217.

Webber, B. et al., 1998. Exploiting multiple goals and intentions in decision support for the management of multiple trauma: a review of the TraumAID project. *Artificial Intelligence*, 105(1–2), pp.263–293.

Wei, W.Q. et al., 2012. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc*, 19(2), pp.219–224.

Wei, W.Q. et al., 2013. The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *International Journal of Medical Informatics*, 82(4), pp.239–247.

Weijters, A.J.M.M. & Ribeiro, J.T.S., 2011. Flexible heuristics miner (FHM). In *IEEE SSCI 2011: Symposium Series on Computational Intelligence - CIDM 2011: 2011 IEEE Symposium on Computational Intelligence and Data Mining*. pp. 310–317.

Weiskopf, N.G. & Weng, C., 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association : JAMIA*, 20(1), pp.144–51. Available at: http://jamia.oxfordjournals.org/content/20/1/144.abstract.

Welch, G. et al., 2015. An internet-based diabetes management platform improves team care and outcomes in an urban latino population. *Diabetes care*, 38(4), pp.561–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25633661.

West, V.L., Borland, D. & Hammond, W.E., 2014. Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association*, pp.330–339. Available at: http://jamia.oxfordjournals.org/cgi/doi/10.1136/amiajnl-2014-002955.

Whiting, D.R. et al., 2011. IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes Research and Clinical Practice*, 94(3), pp.311–321.

WHO, 2016. Chronic obstructive pulmonary disease (COPD). Available at:

http://www.who.int/respiratory/copd/en/.

WHO, 2006. *Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation*, Available at: http://whqlibdoc.who.int/publications/2006/9241594934_eng.pdf%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Definition+and+diagnosis+of+diabetes+mellitus+and+intermediate+hyperglycemia:+report+of+a+WHO/IDF+consultation#0.

WILBANKS, B.A. & LANGFORD, P.A., 2014. A Review of Dashboards for Data Analytics in Nursing. *CIN: Computers, Informatics, Nursing*, 32(11), pp.545–549. Available at: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00024665-201411000-00006.

Winarko, E. & Roddick, J.F., 2007. ARMADA - An algorithm for discovering richer relative temporal association rules from interval-based data. *Data and Knowledge Engineering*, 63(1), pp.76–90.

World Health Organization, 2016. *Global Report on Diabetes*, Available at: http://www.who.int/about/licensing/%5Cnhttp://apps.who.int/iris/bitstream/10665/204871/1/9789241565257_eng.pdf.

Wright, D.F.B. & Duffull, S.B., 2013. A bayesian dose-individualization method for warfarin. *Clinical Pharmacokinetics*, 52(1), pp.59–68.

Xu, J. et al., 2015. Review and evaluation of electronic health records-driven phenotype algorithm authoring tools for clinical and translational research. *Journal of the American Medical Informatics Association*, 22(6), pp.1251–1260.

Yahi, A. & Tatonetti, N.P., 2015. A knowledge-based, automated method for phenotyping in the EHR using only clinical pathology reports. *AMIA Jt Summits Transl Sci Proc*, 2015, pp.64–68.

Yang, W.S. & Hwang, S.Y., 2006. A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications*, 31(1), pp.56–58.

Yazdanpanah, M., Chen, C. & Graham, J., 2013. Secondary analysis of publicly available data reveals superoxide and oxygen radical pathways are enriched for associations between type 2 diabetes and low-frequency variants. *Annals of Human Genetics*, 77(6), pp.472–481.

Yu, H., Zhang, L. & Liu, W., 2008. [Medical knowledge discovery system research based on computer--epidemiological data mining of complications in diabetes mellitus]. *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi*, 25(2), pp.295–299. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18610609.

Yu, Y. et al., 2014. Care Pathway Workbench: Evidence Harmonization from Guideline and Data. In *Studies in Health Technology and Informatics*. pp. 23–27.

Yun Chen & Hui Yang, 2014. Heterogeneous postsurgical data analytics for predictive modeling of mortality risks in intensive care units. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2014, pp.4310–4314.

Zaki, M.J., 2001. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1–2), pp.31–60.

Zanolin, M.E. et al., 2004. The role of climate on the geographic variability of asthma, allergic rhinitis and respiratory symptoms: results from the Italian study of asthma in young adults. *Allergy*, 59(3), pp.306–314. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14982513%5Cnhttp://onlinelibrary.wiley.com/store/10.1046/j.1398-9995.2003.00391.x/asset/j.1398-9995.2003.00391.x.pdf?v=1&t=i7qx4xku&s=e5c300398f5b5c7300c55edaace63f4096def83f.

Zhang, L. et al., 2008. Discovering during-temporal patterns (DTPs) in large temporal databases. *Expert*

Systems with Applications, 34(2), pp.1178–1189.

Zhang, Y. et al., 2016. Application and exploration of big data mining in clinical medicine. *Chinese Medical Journal*, 129(6), pp.731–738.

Zhou, S. et al., 2016. Diversity of Gut Microbiota Metabolic Pathways in 10 Pairs of Chinese Infant Twins. *Plos One*, 11(9), p.e0161627. Available at: http://dx.plos.org/10.1371/journal.pone.0161627.

Ziemer, D.C. et al., 2006. An informatics-supported intervention improves diabetes control in a primary care setting. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, p.1160. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1839402&tool=pmcentrez&rendertype=abstract.

Zillner, S. et al., 2014. User Needs and Requirements Analysis for Big Data Healthcare Applications. In *Studies in Health Technology and Informatics*. pp. 657–661.