UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Programa de Doctorado en Tecnologías para la Salud y el Bienestar

Doctoral Thesis

Dynamic risk models for characterising chronic diseases' behaviour using Process Mining techniques

Author:
Zoe Valero-Ramon

Supervisors:
Dr. Carlos Fernandez-Llatas
Dr. Vicente Traver

7 de enero de 2022

Abstract

Risk models in the healthcare domain are statistical methods that provide early warnings about a person's risk for an adverse episode in the future. They usually use the information routinely stored in Hospital Information Systems to offer an individual probability for developing a future negative outcome in a given period.

Concretely, in the field of chronic diseases that share common risk factors, risk models are based on the analysis of those risk factors -raised blood pressure, raised glucose levels, abnormal blood lipids, and overweight and obesity- and their associated biometric measures. These measures are collected during clinical practice frequently in a periodic manner, and accordingly, they are incorporated into the risk models to support clinicians' decision-making.

Data-Driven techniques could be used to create these temporal-aware risk models, considering the patients' history included in Electronic Health Records, and extracting knowledge from raw data. However, in the healthcare domain, Data Mining results are usually perceived by the health experts as black-boxes, and in consequence, they do not trust in the algorithms' decisions. The Interactive paradigm allows experts to understand the results, in that sense, professionals can correct those models according to their knowledge and experience, providing perceptual and cognitive models. In this context, Process Mining is a Data Mining technique that enables the implementation of the Interactive paradigm, offering a clear care process understanding and providing human-understandable models.

Chronic conditions are usually described by static pictures of variables, such as genetic, physiological, environmental, and behavioural factors. Nevertheless, the dynamic, temporal, and behavioural perspectives are not commonly considered in the risk models. That means the last status of the risk becomes the actual status of the patient. However, the patients' condition could be influenced by their past dynamic circumstances.

The objective of this thesis is to provide a novel risk vision based on Data-Driven technologies offering a dynamic view of the patients' evolution regarding their chronic condition. Technically, it supposes to approach risk models incorporating the dynamic and behavioural perspective of patients to the risk models thanks to the information included in the Electronic Health Records. The results obtained throughout this thesis show how Process Mining technologies can bring a dynamic and interactive view of chronic disease risk models. These results can support health professionals in daily practice for a better understanding of the patients' health condition and a better classification of their risk status.

Resumen

Los modelos de riesgo en el ámbito de la salud son métodos estadísticos que brindan advertencias tempranas sobre el riesgo de una persona de sufrir un episodio adverso en el futuro. Por lo general, utilizan la información almacenada de forma rutinaria en los sistemas de información hospitalaria para ofrecer una probabilidad individual de desarrollar un resultado negativo futuro en un período determinado. Concretamente, en el campo de las enfermedades crónicas que comparten factores de riesgo comunes, los modelos de riesgo se basan en el análisis de esos factores de riesgo -tensión arterial elevada, glucemia elevada, lípidos sanguíneos anormales, sobrepeso y obesidad- y sus medidas biométricas asociadas. Estas medidas se recopilan durante la práctica clínica de manera periódica y, se incorporan a los modelos de riesgo para apoyar a los médicos en la toma de decisiones.

Para crear modelos de riesgo que incluyan la variable temporal, se podrían utilizar técnicas basadas en datos (Data-Driven), de forma que se tuviera en cuenta el historial de los pacientes almacenado en los registros médicos electrónicos, extrayendo conocimiento de los datos en bruto. Sin embargo, en el ámbito de la salud, los resultados de la minería de datos suelen ser percibidos por los expertos en salud como cajas negras y, en consecuencia, no confían en sus decisiones. El paradigma Interactivo permite a los expertos comprender los resultados, para que los profesionales puedan corregir esos modelos de acuerdo con su conocimiento y experiencia, proporcionando modelos perceptivos y cognitivos. En este contexto, la minería de procesos es una técnica de minería de datos que permite la implementación del paradigma Interactivo, ofreciendo una comprensión clara del proceso de atención y proporcionando modelos comprensibles para el ser humano.

Las condiciones crónicas generalmente se describen mediante imágenes estáticas de variables, como factores genéticos, fisiológicos, ambientales y de comportamiento. Sin embargo, la perspectiva dinámica, temporal y de comportamiento no se consideran comúnmente en los modelos de riesgo. Eso significa que el último estado de riesgo se convierte en el estado real del paciente. No obstante, la condición de los pacientes podría verse influenciada por sus condiciones dinámicas pasadas.

El objetivo de esta tesis es proporcionar una visión novedosa del riesgo asociado a un paciente, basada en tecnologías Data-Driven que ofrezcan una visión dinámica de su evolución con respecto a su condición crónica. Técnicamente, supone abordar los modelos de riesgo incorporando la perspectiva dinámica y comportamental de los pacientes gracias a la información incluida en la Historia Clínica Electrónica. Los resultados obtenidos a lo largo de esta tesis muestran cómo las tecnologías de minería de procesos pueden aportar una visión dinámica e interactiva de los modelos de riesgo de enfermedades crónicas. Estos resultados pueden ayudar a los profesionales de la salud en la práctica diaria para una mejor comprensión del estado de salud de los pacientes y una mejor clasificación de su estado de riesgo.

Resum

Els models de risc en l'àmbit de la salut són mètodes estadístics que brinden advertències primerenques sobre el risc d'una persona de patir un episodi advers en el futur. Generalment, utilitzen la informació emmagatzemada de forma rutinària en els sistemes d'informació hospitalària per a oferir una probabilitat individual de desenrotllar un resultat negatiu futur en un període determinat. Concretament, en el camp de les malalties cròniques que compartixen factors de risc comú, els models de risc es basen en l'anàlisi d'eixos factors de risc -tensió arterial elevada, glucèmia elevada, lípids sanguinis anormals, sobrecàrrega i obesitat- i les seues mesures biomètriques associades. Estes mesures es recopilen durant la pràctica clínica ben sovint de manera periòdica i, en conseqüència, s'incorporen als models de risc i recolzen la presa de decisions dels metges.

Per a crear estos models de risc que incloguen la variable temporal es podrien utilitzar tècniques basades en dades (Data-Driven) , de manera que es tinguera en compte l'historial dels pacients disponible en els registres mèdics electrònics, extraient coneixement de les dades en brut. No obstant això, en l'àmbit de la salut, els resultats de la mineria de dades solen ser percebuts pels experts en salut com a caixes negres i, en conseqüència, no confien en les decisions dels algoritmes. El paradigma Interactiu permet als experts comprendre els resultats, perquè els professionals puguen corregir eixos models d'acord amb el seu coneixement i experiència, proporcionant models perceptius i cognitius. En este context, la mineria de processos és una tècnica de mineria de dades que permet la implementació del paradigma Interactiu, oferint una comprensió clara del procés d'atenció i proporcionant models comprensibles per al ser humà.

Les condicions cròniques generalment es descriuen per mitjà d'imatges estàtiques de variables, com a factors genètics, fisiològics, ambientals i de comportament. No obstant això, la perspectiva dinàmica, temporal i de comportament no es consideren comunament en els models de risc. Això significa que l'últim estat de risc es convertix en l'estat real del pacient. No obstant això, la condició dels pacients podria veure's influenciada per les seues condicions dinàmiques passades.

L'objectiu d'esta tesi és proporcionar una visió nova del risc, associat a un pacient, basada en tecnologies Data-Driven que oferisquen una visió dinàmica de l'evolució dels pacients respecte a la seua condició crònica. Tècnicament, suposa abordar els models de risc incorporant la perspectiva dinàmica i el comportament dels pacients als models de risc gràcies a la informació inclosa en la Història Clínica Electrònica. Els resultats obtinguts al llarg d'esta tesi mostren com les tecnologies de mineria de processos poden aportar una visió dinàmica i interactiva dels models de risc de malalties cròniques. Estos resultats poden ajudar els professionals de la salut en la pràctica diària per a una millor comprensió de l'estat de salut dels pacients i una millor classificació del seu estat de risc.

Contents

List of Figures

List of Tables

List of Abbreviations

BF	**B**ody **F**at
BMI	**B**ody **M**ass **I**ndex
BP	**B**lood **P**ressure
BPV	**B**lood **P**ressure **V**ariability
CC	**C**alf **C**ircunference
CG	**C**linical **G**uidelines
CKD	**C**hronic **K**idney **D**isease
CSV	**C**omma-**s**epareted **V**alues
DBP	**D**iastolic **B**lood **P**ressure
DM	**D**ata **M**ining
DRG	**D**iagnosis-**R**elated **G**roup
EHR	**E**lectronic **H**ealth **R**ecords
FPG	**F**asting **P**lasma **G**lucose
ICD	**I**nternational Statistical **C**lassification of **D**iseases and Related Health Problems
IoT	**I**nternet **o**f **T**hings
IPI	**I**nteractive **P**rocess **I**ndicator
IPM	**I**nteractive **P**rocess **M**ining
IT	**I**nformation of **T**echnology
KBTA	**K**nowledge-**B**ased **T**emporal **A**bstractions
KPI	**K**ey **P**erformance **I**ndicator
MNA	**M**ini **N**utritional **A**ssessment
ML	**M**achine **L**earning
O	**O**bjective
PALIA	**P**arallel **A**ctivity **L**og **I**nference **A**lgorithm
PM	**P**rocess **M**ining
QTC	**Q**uality **T**hreshold **C**luster
RQ	**R**esearch **Q**uestion
SBP	**S**ystolic **B**lood **P**ressure
TA	**T**emporal **A**bstractions
TPA	**T**Time **P**arallel **A**utomaton
TTD	**T**Time **T**opological **D**istance
WHO	**W**orld **H**ealth **O**rganization
WTD	**W**eighted **T**opological **D**istance

Chapter 1

Introduction

Chronic conditions, defined by the World Health Organization as diseases of long duration and generally slow progression are by far the leading cause of mortality in Europe, representing 77% of the total disease burden and 86% of all deaths [1]. Some of the chronic diseases with higher impact are coronary heart disease, stroke, many varieties of cancer, depression, diabetes, asthma, chronic obstructive pulmonary disease, or hypertension among others. In fact, over 50 million people in Europe have more than one chronic disease, due to either random co-occurrence, possible shared underlying risk profile, or synergies in disease development [2]. The financial costs associated with treating chronic diseases are extremely high, and given that the average age of European populations is increasing, chronic diseases will continue to place substantial pressure on national budgets [3].

Chronic diseases share common risk factors and conditions, even more, they are well-established and well-known. While some risk factors, such as age, sex, or genetic make-up, cannot be modified, known as non-modifiable risk factors, many behavioural risk factors can be modified and are the same in women and men. Common modifiable risk factors are the unhealthy diet, the physical inactivity, and the tobacco use. The causes of these risk factors are express through the intermediate risk factors of raised blood pressure, raised glucose levels, abnormal blood lipids, and overweight and obesity [4]. The recognition of these intermediate risk factors and conditions is the conceptual basis for an integrated approach to a chronic disease. In fact, these intermediate risk factors are used in the preventive medicine approach to offer an individual probability for developing a future adverse outcome in a given period thanks to the risk models. Risk models use routinely gathered data (such as demographic, measures about metabolic factors, or utilisation history) and stratify the entire population by level of risk. Nevertheless, chronic conditions are not only due to these risk factors, they are also the result of a combination of other factors, such as genetic, physiological, environmental, and behavioural factors, some modifiable and some not, that occur during the process of the diseases. Therefore, one of the main risk models restraints relies on their own definition, committing static pictures of variables, without considering any dynamic and temporal perspective. For example, when talking about obesity chronic disease, when examine people following *miracle diets*, the concept of dynamic analysis acquires a critical importance. Having the complete picture of the patient actual status is paramount to incorporate the information of the evolution in the risk models. Risk and predictive models do not usually respond well to unexpected changes in patient's conditions, as they suit standard conditions rather than unusual or unpredictable ones [5]. Individual differences cause considerable variances in the execution of models, consequently, a new approach incorporating the diseases' behaviours to risk models could offer a better understanding. In this approach, the risk models will be process oriented and will be obtained thanks to the insights and information basically included in the clinical databases.

Actually, these data routinely collected electronically in the health sector supposes one of the best sources of information, that, in combination with increasing computing power, opens a new world of analysis and application for improving patient care in the field of risk modelling [6]. Nowadays, the massive introduction of Electronic Health Records in the medical systems and all facets of the healthcare processes, has generated an enormous amount of information, the testimony of the patient's journey along with the received care. In this scenario, healthcare professionals could have not only the information collected within the healthcare settings but also data coming from other multiple sources, such as personal and environmental data thanks to wearable, sensors, Internet of Things, mobile applications, or even social media.

All this information could play an important role in the management of chronic diseases, however, these data do not suppose any significant progress by themselves. In fact, this is not a matter of quantity of data, is what to do with these data, extracting knowledge from them and presenting the results to health professionals in a manner they trust and understand. An in-depth analysis of these data is necessary to obtain the knowledge that allows not only the improvement of the quality of the care provided to patients with chronic conditions but also to move towards a patient-centred healthcare model, within the personalised medicine paradigm. Obtaining insights and knowledge from the huge amount of data stored in Electronic Health Records and other data sources might transform the way we understand risk models and chronic diseases. Notwithstanding, this approach should be proposed and evaluated in concrete medical conditions, as chronic conditions are.

Therefore, the point of departure for this work is to understand how chronic diseases work, their common characteristics, and their consequences in patients' health but also in their analysis and management. It is needed to describe chronic diseases in the framework of this work, together with their common risk factors, and how risk factors are used through the risk models for health diseases to provide early warnings about a patient's risk for an adverse episode in the future. Risk models definition and use introduces benefits, but also limitations as they are currently defined, it is without considering the evolution perspective and patients' behaviour. These two limitations suppose the main challenges this work intends to solve.

Departing from the approach of risk models in the current preventive medicine, understood as statistical tools that offer *an individual probability of developing a future adverse outcome in a given period*, we consider their two main limitations, the none consideration of the temporal perspective neither the patient particularities, that are precisely the fundamentals for a dynamic approach. On the one hand, considering diseases dynamically; they evolve towards different destinations, especially when talking about chronic health problems. Similarly, the human being is not invariant; a person is changing throughout her/his biography in age, lifestyle, socioeconomic status, or inter-current diseases. On the other hand, patient characteristics should be contemplated for discovering more accurate stratification groups. Hence a risk model dynamically approached should examine the temporal perspective and maximise the process value based on each group condition and characteristics. Then, it is desirable to incorporate this knowledge in the definition of the risk models. As a consequence of the huge amount of available data, Data-Driven techniques could be used to create this *temporal-aware* risk models, considering the patients' history.

In this context, Data Mining techniques aim to extract useful knowledge from raw data. Interest in this field arose due to the advances in Information Technology and the rapid growth of business and scientific databases. These data hold valuable information such as trends and patterns, which can be used to improve decision making. However, Data Mining techniques are seen as *black-box* systems by health professionals. There is a wide range of *black-box* supervised Data Mining methods, which are capable of accurate predictions, but where obtained models are too complex to be easily understood by humans [7], this means that humans do not usually trust in their results as they do not have the rationale behind the decisions of the algorithm. Interactive frameworks are thought to implicate the experts in the process of automatic learning, supporting them in the process of cognition, and in the better understanding of the interventions, allowing the correction, and the selection of the best personalised solution based on their own experiences. In this approach, Process Mining techniques offer human understandable pattern recognition algorithms needed for the application of the Interactive paradigm [8]. Accordingly, the application of Process Mining techniques can reinforce health experts in the understanding of chronic diseases underlying processes by providing more understandable risk models.

For the research of this work, we have selected PMApp as a Process Mining tool, although the work described in this doctoral thesis is tool independent, and could be performed with other Process Mining tools, PMApp has been pioneer in the application of Interactive methodologies in several real healthcare scenarios [9]. Moreover, it has been developed within our research team, this fact allowed us a complete access to the tool for creating specific algorithms and plug-ins for the development of the research and the methodology.

This novel approach to risk models should shape a chronic condition and create patients' behaviour based stratification using Process Mining techniques. Furthermore, a formal methodology supporting this modelling would allow health professionals creating these new risk models. Consequently, this document presents the novel definition for the *temporal-aware* risk models, that we called *Dynamic Risk Models*, and how Process Mining in combination with trace clustering techniques could tackle this new definition, supporting the definition with an experimentation.

To do that, the work presented examines the behavioural modelling of chronic diseases, again using Process Mining techniques and data collected in clinical daily practice and stored in Electronic Health Records. Concretely, three Dynamic Risk Models are presented for three different chronic conditions, hypertension, hyperglycemia, and obesity. These three chronic conditions were selected because of their wider prevalence among the population, and specifically in obesity, because of its difficulty in the screening. The models have been obtained following concrete procedures through the experiments and trying to answer specific questions. This fact conducted to the formal introduction of an interactive and question based methodology for deploying Dynamic Risk Models based on the well-know PM^2 methodology [10], and its put into practise through a use case for obesity chronic disease. The methodology supports clinicians thanks to a process oriented view on what and how they want to inquire about underlying chronic conditions.

1.1. Hypothesis, research questions and objectives

1.1.1. Hypothesis

Given the necessity already presented to incorporate, both the temporal perspective and the dimension of the behaviour of the patients themselves, to the current definition of risk models for the specific case of chronic diseases, a new approach will be proposed for the risk models that includes these needs based on the use of Process Mining techniques. Our hypothesis to be validated is that:

> *The dynamic perspective of chronic conditions could be incorporated into risk models using historical data coming from Electronic Health Records and Process Mining techniques to analyse them, offering a novel way to better understand the diseases' behaviour with a process oriented view.*

Based on this hypothesis, and to discover novel risk models that consider the dynamic and behavioural perspective for chronic conditions, it is needed to evaluate the potentiality of the Process Mining techniques to approach risk models dynamically, and if the data included in the Electronic Health Records (EHR) provides the needed information. This fact was conducted to the formulation of the following specific research questions and objectives.

1.1.2. Specific research questions and objectives

In order to confirm the hypothesis, the following research questions were identified:

- **RQ1 -** Can the evolution of a chronic condition be modelled using Process Mining techniques in an understandable manner for healthcare professionals?
- **RQ2 -** Can we create a patients' behaviour based stratification (Dynamic Risk Models) using Process Mining techniques that allow us to build a perspective for better understanding the chronic conditions' evolution?
- **RQ3 -** Can we define a formal methodology for approaching the dynamic perspective of chronic conditions using Interactive Process Mining techniques to obtain Dynamic Risk Models for chronic diseases?

To achieve the research questions of the doctoral thesis, a set of secondary objectives were established:

- **O1 - To evaluate the viability of approaching a medical condition using Process Mining techniques to find behavioural dynamics.** Risk and predictive models in the preventive medicine approach are static snapshots of variables that usually do not consider the dynamic perspective. However, medical conditions vary and evolve. They are the result of a combination of several factors, behavioural, environmental, and physiological among others. Therefore, it is necessary to incorporate the dynamic perspective and those factors, and approach a medical condition as a process. In that sense, Process Mining techniques could find patterns and behavioural models in sequential data.
- **O2 - To approach chronic health conditions using Process Mining techniques and information from EHR to obtain behavioural risk models.** Chronic conditions comprise several characteristics that made them a perfect candidate to study the possibilities of discovering dynamic risk models associated with them. They use to be of long duration and patients are periodically monitor, consequently EHR store important information about the underlying process. The main idea is to model a chronic disease through a human-understandable graphical representation that could support healthcare professionals in comprehending their current awareness of the chronic disease processes, as it takes into consideration the disease's variability over time and patient nature, using available data coming from EHR.
- **O3 - To propose a formal methodology to translate chronic conditions evolution into dynamic flows using Interactive Process Mining techniques.** In this context, it is not just about representing time-stamped data in a qualitative way, but also a question of what data represent and how to show them. It is crucial allowing health professionals to be committed to the method, incorporating them into the procedure. For that, the Interactive paradigm working together with Process Mining offers a unique framework to develop the methodology that could allow translating clinical questions from healthcare professionals into visual patients' behaviour based stratification models that bring open questions out enabling them to manage actual processes. These models allow navigating them into data gaining insights, highlighting the real behaviour of the patients through understandable views that support the clinical questions.

These secondary objectives have been used to structure the work carried out during the document. This doctoral thesis is intended to create a new formal framework to discover visual dynamic risk models associated with chronic diseases. This means, on the one hand, to answer if the Process Mining approach can dynamically model medical conditions with the available data. And on the other hand, if the obtained models offer useful and novel knowledge to health professionals compared with the current risk models. For this, the proposed methodology should not only work for different chronic conditions but also contemplate a set of formal steps or stages that result in the desired model in the manner health professionals need it. And latest to define a set of concrete methods to implement the methodology in the appropriate manner and form.

1.2. Structure

Image 1.1 presents the overall structure of the document. Chapter 1 introduces the main hypothesis of the work, together with the motivation, the research questions, and research objectives. Chapter 2 describes the background of the risk models for health diseases, as they are currently defined, and the gap/problem the work tries to solve, how it has been approached in the literature and the new approach proposed. Chapter 3 describes what and how the research was approached in the work, identifying and introducing the main materials and methods used. Chapter 4 conducts a research to model a medical condition with Process Mining techniques using real data from a well-characterised and bounded population and disease. Chapter 5 examines the dynamic modelling of chronic conditions using Process Mining techniques with the information stored in the EHR, concretely develops two different dynamic risk models for two chronic conditions using the associated risk factors, high blood pressure, and hyperglycemia. Chapter 6 develops and describes a formal methodology for obtaining risk models for medical conditions. Besides, it provides an example of application to obesity chronic disease. Chapter 7 states how the work has committed the research question and the objective together with its main conclusion and introduces possible future work. Finally, Chapter 8 presents the main original contributions of the present work.

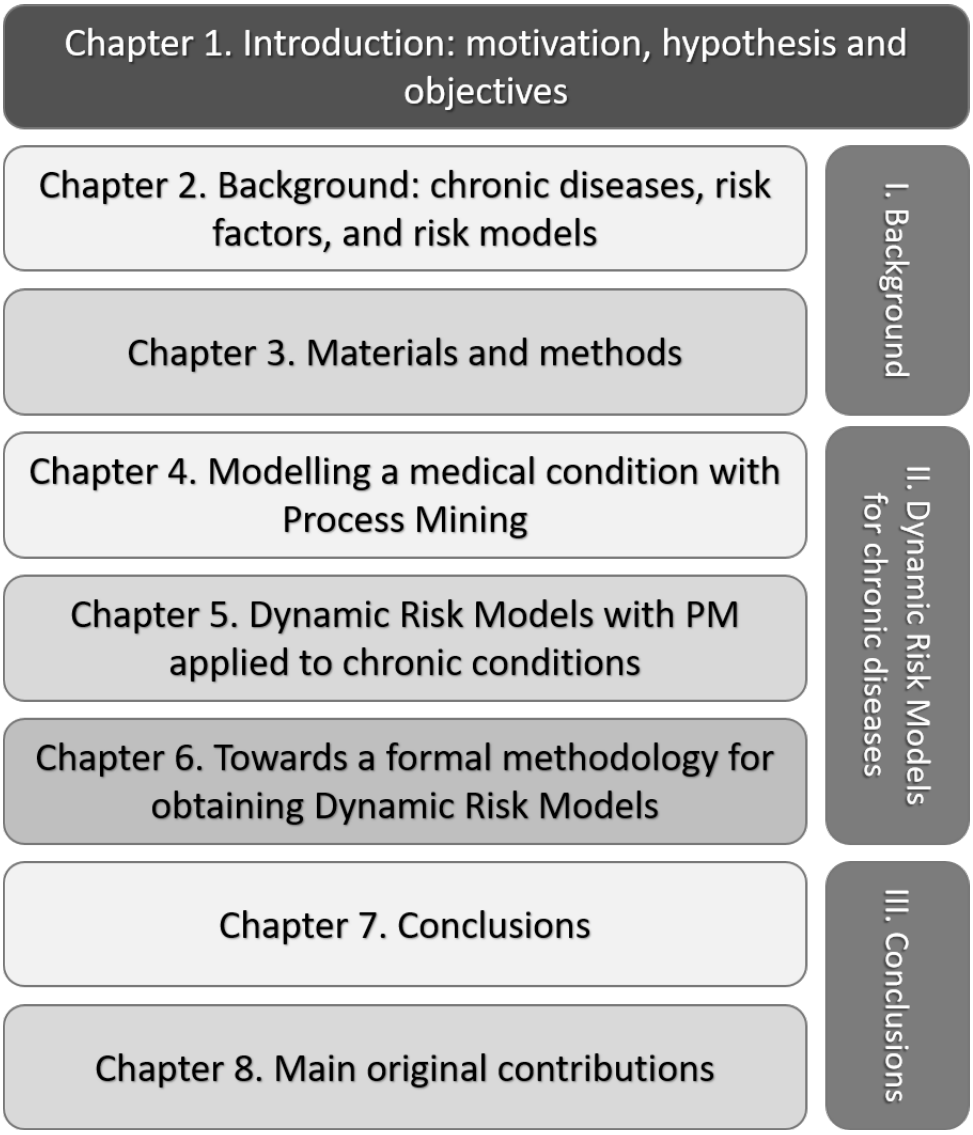

Chapter 1. Introduction: motivation, hypothesis and objectives
Chapter 2. Background: chronic diseases, risk factors, and risk models
Chapter 3. Materials and methods
I. Background
Chapter 4. Modelling a medical condition with Process Mining
Chapter 5. Dynamic Risk Models with PM applied to chronic conditions
Chapter 6. Towards a formal methodology for obtaining Dynamic Risk Models
II. Dynamic Risk Models for chronic diseases
Chapter 7. Conclusions
Chapter 8. Main original contributions
III. Conclusions

FIGURE 1.1: Structure of the thesis.

Chapter 2

Background

As introduced in the Chapter 1, chronic conditions are accustomed to being of long duration. In this context, the temporal dimension might be crucial for a complete awareness of a health problem. In a scenario where Digital Transformation is increasing its presence, clinical information collected and stored in EHR could suppose a great opportunity if we can incorporate them into a novel generation of risk models that will provide insights and knowledge from this massive amount of data. It might transform the way risk models and chronic conditions are currently understood. To pursue this aim, this chapter starts presenting the concept of chronic conditions, intending to comprehend them in a better manner, and to provide a general overview of how risk models for chronic conditions work now. The chapter also explores what limitations we encountered in the way risk models are currently interpreted and used by health professionals. Thus, the motivation of this doctoral thesis relies on the exploration of how some of these limitations can be resolved with a novel approach using Process Mining techniques that considers diseases as processes, including patients' behaviour and care procedures.

2.1. Understanding chronic conditions from a temporal perspective

A key consideration for understanding the chronic conditions and their consequences is looking them from the perspective of time. By time we mean temporal perception and dynamics of the disease. Many current models assume linear trajectory of living with a chronic disease, however, living with a chronic condition is an ongoing and continually shifting process with implications in several person's dimensions. Consequently, it is worth looking at chronic conditions and their risk factors from the temporal perspective.

2.1.1. Chronic diseases and risk factors

In order to approach a chronic condition from a temporal perspective, first we need to understand what a chronic disease is and what common characteristics they have. There are several definitions for chronic diseases, for example, the World Health Organization (WHO) defines them as *diseases of long duration and generally slow progression* [1]. Berstein et al. [11] defined them as having one or more of the following characteristics –they are permanent, leave residual disability, are caused by non-reversible pathological alteration, require special training of the patient for rehabilitation, or may be expected to require a long period of supervision, observation or care. For the purpose of this doctoral thesis and fusing this two well-known definitions, we consider a chronic condition as a *persistent disease or long-lasting in its effects, with a long period of supervision, observation and care*.

Chronic diseases, such as heart disease, stroke, cancer, chronic respiratory diseases, and diabetes, are by far the leading cause of mortality in Europe, representing 77% of the total disease burden and 86% of all deaths [1]. The financial costs associated with treating chronic diseases are extremely high, and given that the average age of European populations is increasing, chronic diseases will continue to place substantial pressure on national budgets [3]. Similarly, chronic diseases are among the most prevalent and costly health conditions in the United States. Nearly half (approximately 45%, or 133 million) of all Americans suffer from at least one chronic disease, having a great impact on health care costs [12]. Some of the chronic diseases with higher impact are coronary heart disease, stroke, many varieties of cancer, depression, diabetes, asthma, chronic obstructive pulmonary disease, or hypertension among others. Over 50 million people in Europe have more than one chronic disease, due to either random co-occurrence, possible shared underlying risk profile, or synergies in disease development [2]. Persons with chronic conditions are a large and growing segment of the population. Although chronic conditions are often associated with the older age population, evidence shows that 15 million of all deaths attributed to chronic diseases occur between the ages of 30 and 69 years [1], confirming that chronic conditions might affect people from all ages, but also representing an opportunity involving younger population in the management of their own conditions.

In this scenario, chronic diseases share common risk factors and conditions, even more, they are well-established and well-known. While some risk factors, such as age, sex, or genetic make-up, cannot be modified, called non-modifiable risk factors, many behavioural risk factors can be modified and are the same in men and women. Common modifiable risk factors are the unhealthy diet, physical inactivity, and the tobacco use. The causes of these risk factors are expressed through the intermediate risk factors of raised blood pressure, raised glucose levels, abnormal blood lipids, and overweight and obesity [4]. The recognition of these common risk factors and conditions is the conceptual basis for an integrated approach to a chronic disease. The intermediate or metabolic risk factors contribute to four key metabolic changes that increase the risk of suffering chronic diseases, these are raised blood pressure, overweight and obesity, hyperglycemia (high blood sugar levels), and hyperlipidemia (abnormally elevated levels of any or all lipids or lipoproteins in the blood). In terms of attributable deaths, the leading metabolic risk factor globally is elevated blood pressure, to which 19% of global deaths are attributed [13], followed by overweight and obesity, and raised blood glucose [1].

These three risk factors and their associated biometric measures are collected during clinical practice usually in a periodic manner and stored in EHR; so they are ready to be used and analysed. However, chronic conditions are not only due to these risk factors, they are the result of a combination of genetic, physiological, environmental, and behavioural factors, some modifiable and some not, that should be taken into consideration during their management, analysis, and treatment. Information collected in EHR usually include such data or part of them, however it is not completely analysed and consequently it is not considered when managing chronic conditions. Chronic conditions commonly require ongoing management over a period of years or decades, so individuals' behaviour should be taken into consideration. The younger segment of the population suffering from chronic diseases has a long period for dealing with them but they can also suppose an allied in the disease management, adopting and using a range of sensors or mobile personal devices for monitoring a set of factors. Despite sensors and related technologies already have some challenges, as precision, size, power consumption, communication, and privacy, they provide valuable information for disease management, and ultimately for improving patients' quality of life [14]. Since recently, the common fact of measuring physiological variables such as blood pressure or glucose levels was traditionally done by exams in a specialised health centre. Thanks to the development and introduction of a considerable set of sensors reading vital signs, such as blood pressure cuff, glucometer, heart rate monitor, including electrocardiograms, this situation has radically changed, allowing patients to take their vital signs daily at home [14]. This has a double potentiality, on the one hand, patients are aware of their vital signs and can better manage their conditions. On the other hand, these data will significantly complement standard tests included in EHR.

This emerging Digital Transformation supposes an exceptional opportunity for creating new models and extracting knowledge from data, using Data Mining paradigms [15]. Consequently, an in-depth analysis of these data is paramount to obtain the necessary knowledge that allows, not only to improve the quality of the provided care but also better management of diseases and to move towards a patient-centred and value-based healthcare model, within the personalised medicine paradigm. Personalised medicine promises prediction, prevention, and treatment of illness that is targeted to individuals' needs [16]. Furthermore, it is a demand within these new paradigms to analyse data in a dynamic and integrated way, instead of linear [16]. Another opportunity in the area is the development of new algorithms to support clinicians, new visualisation tools that understandably show processes and models for health professionals, and decision support tools, allowing more effective and precise management of diseases and treatments [17].

2.1.2. Risk Models: obtaining insight and information from data

In order to support clinicians in their daily practice, there are statistical methods that provide early warnings about a patient's risk for an adverse episode in the future, called *risk models*. They use the information routinely gathered in health consultations, such as the metabolic risk factors, and analyse them to support clinicians' decision-making. Concretely, in the preventive medicine approach risk models are associated with health diseases. They are statistical tools intended to offer *an individual probability for developing a future adverse outcome in a given period* [18]. Therefore, risk models are frequently used in clinical trials to determine the eligibility of the participants, also as a measure to improve cost-effectiveness in a preventive intervention, and as a decision support tool to facilitate personal healthcare decisions where the model should assign *each individual* near to her/his a true risk. Preventive risk models assign a distribution of future risks over an entire population, and because they are based on routinely gathered data (such as demographic, or measures about metabolic factors), they can *stratify* the entire population by all levels of risk. This is an important aspect of risk models, the use of routinely collected data which are captured for other purposes but can be utilised for computing risk models, therefore its implementation may require minor efforts regarding dataset infrastructures [19]. In fact, data routinely collected electronically in the health sector supposes one of the best sources of information, that, in combination with increasing computing power, opens a new world of analysis and application for improving patient care in the field of risk modelling [6].

Consequently, the use of risk models introduces many benefits as they support and complement clinical reasoning and decision-making in medicine. In this context, risk models are computed in a moment and have validity over time. Results from risk models understood as risk values of an individual patient, play an important role in the decision taken by health professionals, who decide treatments delivered to patients depending on them. Moreover, predictive risk models could also be focused on predicting more general adverse event, such as the utilisation of a specific service, trying to influence in the design of the entire health system [19]. Moreover, risk models may enable the stratification of the population with chronic conditions in order to develop new models of care based on their healthcare resources consumption [20].

In fact, they have been successfully applied to several healthcare domain, such as in the evaluation of developing complication in patients with diabetes, identifying the significant risk factors associated [21]; in the prediction of health costs and utilisation assessing concrete variables such as the Body Mass Index (BMI) [22]; complementing the screening referral decisions by identifying those patients at greatest risk of colorectal cancer [23], among others.

2.1.3. Risk models limitations

Although the main benefits of using risk and prediction models in the healthcare domain are clear, since they are now implemented have some limitations. It is worthy to explain this fact with a concrete example based on obesity chronic disease. Obesity disease implies a risk of suffering from other chronic diseases as a result of the excess weight, such as cardiovascular diseases, asthma, and musculoskeletal disorders [24, 25, 26]. When a patient is classified as obese, with a Body Mass Index (BMI) greater than or equal to 30 kg/m^2, the risk of comorbidities is considered as severe at this point [25]. However, this is not only a question of patient's current state; it is indeed more important to consider obesity onset, evolution, weight fluctuations, duration of obesity (known as the time since BMI was first known to be at least 30 kg/m^2), or even parental BMI to see comorbidities association and treatment [27, 28]. Notwithstanding, in real practise and from the actual static risk models, when a patient achieves a normal BMI after a weight lost, her/his risk is commonly re-computed to a normal risk situation, and usually, comorbidities disappear from the patient's context. In other words, the evolutionary perspective is not considered. Changes in the individuals are usually connected to behaviours, attitudes, and beliefs, meaning that people with the same disease and treated with the same treatment respond in different ways [29].

Therefore, one of the main risk models restraints relies on their own definition, committing static pictures of variables, without considering any dynamic perspective. The current understanding of risk models relies on static *snapshots* of variables or measures, rather than ongoing, dynamic feedback loops of behaviour considering changes and different states. Conventionally, modelling, assessment, and management of the risk of healthcare variables have been done from a static and time-invariant set of concepts, definitions, and propositions, assuming a linear relationship between variables. Risk and predictive models do not respond well to unexpected changes in patient's conditions, as they suit standard conditions rather than unusual or unpredictable ones [5]. Individual differences cause considerable variances in the execution of models. It is also hard to judge whether results obtained on a specific cohort can be effectively translated to other populations. Another of their main limitations is their *one-size-fits-all* approach. That is, using all available data to build a general model, and then with this model, predicting the individual probability of developing a disease. However, patients may have different medical conditions, different socio-economic characteristics, live in different environments, etc. Using a comprehensive model may miss some specific information that is important for individual patients. Thus, building a patient-centred and temporal-based model is important in the field of personalised medicine.

Consequently, a new approach for the risk models should consider the dynamic characteristics of the diseases, including disease's variability and dependencies with other conditions, such as comorbidities, social conditions, or age. Notwithstanding, in the concrete case of chronic diseases, they tend to be of long duration consequently physiological, environmental, and behavioural factors might vary, so this evolution should be also contemplated during their analysis, and treatment. On the other hand, the temporal perspective of the clinical information is crucial for exhaustive knowledge of a health process. Diseases are not changeless; they evolve towards different destinations, especially when talking about chronic health problems. For example, in reference [30], the results suggested that optimal blood pressure management in children with chronic kidney disease (CKD) slows progression to end-stage renal disease and that works focused only on baseline blood pressure measurement may underestimate risk than using time-fixed blood pressure. Likewise, the human being is not invariant, a person is changing throughout her/his biography in age, lifestyle, socioeconomic status, or intercurrent diseases, consequently this should be reflected in the risk models.

2.1.4. Standardisation of risk models

In the literature, there are several approaches to standardise risk models in medicine using time-stamped data. Knowledge-Based Temporal Abstractions (KBTA or TA) is one of them [31]. Overall, TA are methods used to achieve a switch from a qualitative time series description of raw data, to a qualitative interval-based representation of time series, intending to abstract high-level concepts from time-stamped data. TA has been used to approach health processes in some areas. In this work [32], the authors proposed using TA together with dynamic Bayesian networks in the prognosis of the risk coronary disease, predicting the risk of a particular patient by suffering a coronary heart disease event based on his/her past medical history.

Other works approached the use of TA for the assessment of costs related to Diabetes Mellitus [33], and the combination of TA and Process Mining techniques to derive a data-driven stratification model that supports the assignment of personalised care plans for the diagnosis and treatment of diabetes and for the anticipation [34]. This method has been also applied for defining typical medical abstraction patterns [35] that are high level descriptions of the temporal behavioural of medical biosignals. This other work [36] presented the use of TA and mining of physiological data streams to develop process flow mappings that can be used to update patient journeys in a neonatal intensive care setting. Previous works tried to generate an automatic summarising of the patient's current based on his/her data through temporal abstraction. However, the great majority of clinical variables, such as weight, blood pressure, or blood sugar, have numerical results, while TA techniques are based on discrete labels, and in consequence, these techniques could oversimplify important information from the analysis. Going a step forward, the work presented in [37] performed a dual approach, using TA in combination with Process Mining for blood pressure and temperature .

In this line, other authors suggested the importance of taking into account the full set of behaviours through real-time measurements to create models over time and, in consequence, inferring patterns, context, and states of patients, with the ultimate objective of developing personalised interventions [38]. Nonetheless, modelling methodologies rely on predictive strategies rather than on the evolution of patient measurements or pathways. In this line, it is needed to implement a Data-Driven approach capable of discovering patients' behavioural models as temporal and dynamic flows succeeding precision medicine paradigm [39]. With this objective, Data-Driven models are decisive for supporting the discovery of individuals' behaviour process [40].

2.2. Towards a *white-box* approach

In the Evidence-based Medicine paradigm [41], it is proposed the formulation of protocols and guidelines using the best existing evidence in the literature in combination with the running knowledge of the health professionals. However, excessively general protocols ignore patients that might have different responses to the same treatment. In this line, Personalised Medicine [42] looks for new treatments considering individuals. This new concept develops new strategies for analysing individual variability for improving the care each patient receives. An example of that is human genomics, which contemplates the genetic information in humans for selecting the best treatments. However, Personal and Precision Medicine is more than genomics, it could be also the utilisation of all available data from a single patient for building high computing systems that can support professionals in the selection of the best care for each case. This paradigm in combination with Integrated Care [43], looks for a complete view of the patient taking into account all the information available of the patient as a process.

All the available information from a patient could be easily found in the EHR, allowing an enormous amount of available data, witnesses of the patient's passage along with the received care. An in-depth analysis of these data is necessary to obtain the required knowledge supporting, not only the improvement of the the quality of the provided care but also to move towards a patient-centred and value-based healthcare model, within the personalised medicine paradigm. From an organisational point of view, the analysis of the established care circuits would allow detecting deficiencies, redundancies, critical points that might affect patient safety and satisfaction, and care pathways efficiency. Form the point of view of the provided care, the analysis of the information included in EHR could improve and personalised the treatments. Moreover, predictive modelling tasks for diseases using EHR information is an increasing area of interest both for researchers and clinicians. The EHR records are series of sequenced and temporal data representing patient visits, where each visit is a set of high dimensional clinical events. Therefore, EHR includes the temporal perspective, with high level-concepts and time-stamped data.

However, this high quantity of data provided by the EHR can not be efficiently comprehend by the clinical experts in an adequate and timely manner as it is stored as raw data. Therefore, the critical issue is not the availability of such huge amount of data, or even how to store and manage them, it is the use of intelligent systems for processing such data and presenting the essential underlying facts that are significant to the expert [44]. Accordingly, to support clinical experts to understand what is actually happening in a healthcare process or environment, it is needed to translate EHR raw data into knowledge. Nevertheless, this is not a meaningless issue. Greatest of Machine Learning models are based on mathematical algorithms that produce accurate models based on the data, and they usually offer an abstraction of what is actually happening in the process, and consequently the results are difficult to be interpreted by the health experts.

To support the human interpretation of the results of such tools, semantic [45], cognitive [46], and perceptual computing [47] paradigms work to produce actionable information. Whereas *Semantic Computing* offers a natural way for communicating the result to experts, associating a meaning to the data considering the available context; *Semantic Computing* is intended to mimic the human brain trying to build artificial experts that learn from experience by matching patterns and using the automatic learning algorithms. Going a step forward, *Perceptual computing* incorporates the concept of personalisation in cognitive spaces. One of the main challenges of these paradigms relies on their own definition as they are thought to be applied by replicating the human mind capabilities [47], that is not demonstrated to be possible within the current computing paradigms.

In this context, Data Mining techniques aim to extract useful knowledge from raw data. Raw data hold valuable information such as trends and patterns, which can be used to improve decision-making. However, Data Mining techniques are seen as *black-box* systems by health professionals. There is a wide range of *black-box* supervised Data Mining methods, which are capable of accurate predictions, but where obtained models are too complex to be easily understood by humans [7], this means that humans do not usually trust in their results as they do not have the rationale behind the decisions of the algorithm.

Moreover, in the concrete case of the healthcare context, the human mind replica has associated ethical, social, and moral aspects that cannot be leveraged by computers. Since imitating the human brain is not the best path, another approach could be to incorporate it as another system within the paradigms, instead of replicating or emulating their capabilities. This is precisely what Interactive Pattern Recognition [48] relies on, it uses the human mind as a central component in the learning system. Interactive frameworks are thought to implicate the experts in the process of automatic learning, supporting them in the process of cognition, and in the better understanding of the interventions, allowing the correction, and the selection of the best personalised solution based on their own experiences [8]. Thus, the objective of the Interactive Pattern Recognition systems is to ensure a close collaboration between the automatic learning algorithm and the human expert. Along this interaction it is achieved a double goal, on the one hand, this approach enables that the expert can better understand the results, and on the other hand, the professional can correct those models according to her/his own knowledge and experience, providing perceptual and cognitive models. For this, human understandable pattern recognition algorithms are needed [8].

Classical Data Mining techniques such as Neural Networks [49], Support Vector Machines [50], or hidden Markov models [51] are pattern recognition tools that provide highly accurate models, however they do not provide human understandable models that enable an interaction between the expert and the intelligent system. Consequently, other tools that empower the interactive paradigm are needed in this context. Process Mining technologies is one of these tools.

Process Mining [52] solutions can offer a clear care process understanding in a better way compared to other Data Mining techniques that are seen as *black-box* systems. Overall, Process Mining is based on syntactical Data Mining framework [53] that is thought to support process experts in the understanding of the process in a comprehensive, objective and exploratory way. It provides human understandable models enabling the application of the Interactive paradigm. In this way, the application of Process Mining techniques can reinforce health experts in the understanding of chronic diseases underlying processes by providing more understandable risk models.

2.3. Towards a new definition for risk models

Within this chapter, we have formally introduced the chronic conditions, their associated risk factors, and the benefits and limitations of risk models as they are currently defined and used. Therefore, we have stated the major restraint of committing risk models as static snapshots of variables and measures; without considering the dynamic perspective associated with the variables and diseases themselves. From this limitation, the doctoral thesis states the need of a new approach that incorporates the dynamic perspective of the diseases and data coming from EHR systems to the risk models. This novel approach to risk models should shape a chronic condition and create patients' behaviour based stratification using Process Mining techniques. Furthermore, a formal methodology supporting this modelling would allow health professionals creating these new risk models.

Thus, the subsequent chapter describes the main materials and methods used during the present work to solve this challenge, with a special focus on the Process Mining techniques and the active involvement of the healthcare professionals in the analysis of its results thanks to its graphical visualisation. Then, the different techniques are applied to model a medical condition and the basis for the patients' behaviour based stratification through the Dynamic Risk Models are shaped and put into practise throughout the following chapters.

Chapter 3

Materials and Methods

The main objective of this section is to describe what and how the research was approached in this work. Following the central motivation of the current work, which pursues to infer real processes for chronic diseases based on the available data from the medical data sets, together with the incorporation of the dynamic and behavioural perspectives to the risk models, this chapter states what materials and methods could support to reach such objective.

This chapter introduces the Interactive paradigm [48] that promotes the interaction of the human with the machine using the human brain as another computation node in the learning system. This means, the application of an Interactive approach requires the active involvement of the human expert in the process.

Unfortunately, this fact requires Data-Driven models understandable by the expert so they could interact with them and an interactive technique that supports its application. Process Mining [52] is a relatively new paradigm increasing this presence in the medical domain that supports the interactive approach. Accordingly, this chapter describes the Process Mining paradigm and introduces the concrete techniques and tools used in the research. Finally, trace clustering techniques are also described supporting the patients' stratification based on the behavioural aspects of the risk models.

3.1. Interactive approach

As explained in Chapter 2, Data Mining models usually have one major disadvantage in the healthcare domain, which is they are commonly considered by professionals as *black-boxes* [54]. Moreover, they often have a high learning curve as healthcare professionals do not understand what is behind the inferred models, as greatest of Machine Learning models are based on mathematical algorithms that produce accurate models but not understandable. Representing results in a human-understandable manner could definitively help to overcome part of the limitations but also incorporating them in the discovery process could suppose a considerable difference.

The results of these technologies are questioned by the professional who does not understand and trust them. Medical diagnosis model is responsible of human lives, therefore health professionals are not confident enough to treat a patient as instructed by a *black-box*. Moreover, in current medicine there is not new knowledge without medical understanding; in other words, clinicians should understand the disease process to add their evidence to the current medical knowledge. However, these results can not offer insights to medical doctors because they are not human-understandable. Besides, these inductive methods are based on statistical frameworks that produce accurate results only when the number of cases is adequate. Nevertheless, healthcare professionals are unlikely to need more support in the standard case because is usually covered by the standard treatment. They require help with treating rare cases and classic Machine Learning technologies have lower accuracy in these contexts [55]. An Interactive paradigm [48] is meant to provide a solution to this problem.

The *Interactive paradigm* defines this concept by integrating human activity into a process [48]. It assures a close collaboration between an automatic learning algorithm and the human. It provides models that professionals can use for a better understanding of the actual process but also improve those models according to human knowledge so new perceptual and cognitive models are provided. It is clear then that the application of an Interactive methodology requires the direct engagement of human experts in the learning process. This fact also enforces the acceptance of the entire methodology and its results by professionals.

Explainable models in Machine Learning provide results that are understandable for experts in the domain, in contracts with *black-box* models that are extremely difficult to explain and barely be understood even by domain experts [56]. However, they need a *translator* to provide knowledge to the expert, and this interaction is unidirectional, from the model to the expert. Therefore, the expert can understand the decision taken, but it is difficult to discover alternatives, or correct and improve the models. It is needed a bidirectional framework to allow a real interaction between the model and the healthcare professional. This could suppose an interactive process where the expert is aware of what is actually happening in the process and could modify it with her/his expertise and knowledge. This fact diminishes the interactive learning capabilities which is based on the corrections introduced by the expert to improve the system repeatedly.

3.2. Process Mining

Considering medical processes are hard to be designed by consensus of experts, the utilisation of data for creating medical processes is a recurrent idea in the literature [57], [58], [59]. Data-Driven methods are an example of feasible solution in this field, supporting medical experts in their daily decisions [60]. Behind this paradigm, there are frameworks specifically designed for dealing with process-oriented problems, as this is the case of Process Mining.

Process Mining provides tools, algorithms, and visualisation instruments to allow human experts obtaining information about the characteristics of the execution of a process, by analysing the trace of events and activities that occur in a concrete procedure, from a process-oriented perspective.

Process Mining technology is Data-Driven with a focus on understandability rather than accuracy. This is because Process Mining technologies are expert focused that means presenting understandable models is more priority than create accurate models, for that Process Mining technologies renounce accuracy if needed for creating human-readable models [9]. In consequence, they provide an excellent opportunity to extract knowledge from the Data-Driven world. This supposes the application of interactive models combining the best from the Data and Knowledge-Driven paradigms [48].

Process Mining has a close relationship with workflow technologies, as Process Mining algorithms usually represent their finding as workflows. Workflows are the most commonly used representation framework for processes. Clinical guidelines represent some decision algorithms using workflows, because of their simplicity and ease of understanding [61]. Process Mining algorithms use the events recorded in each process and represent them as a workflow. This workflow infers the original flow in an understandable and enriching manner and supports experts in the actual knowledge of what is happening. For that, this paradigm offers a high-level view to professionals, allowing them an enhanced understanding of the complete process.

Process Mining algorithms are usually divided into three groups [62]:

- Process Mining Discovery Algorithms: these are tools that can graphically create described workflows from the events recorded in a process [63]. Different Process Mining discovery algorithms have been used in healthcare scenarios [64], [65]. The selection of the adequate discovery algorithm depends on the quality of the available data, and the kind of the desired representation workflow.

- Process Mining Conformance Algorithms: these are algorithms that can detect if the flow pursued by a patient conforms with a defined process [66]. They, for example, can be used to measure the patient's adherence to a specific treatment and to allow the graphical representation of the moment where the patient is not fulfilling the treatment flow, supporting the clinicians in the improvement process of the patient's adherence. These techniques can also be used to compare processes, and to detect the differences in their executions. Moreover, conformance algorithms can compare workflows and show the differences in a graphical manner. It allows healthcare experts to quickly identifying changes in different processes. For instance, this technique has been used to detect behavioural changes over time in humans [48]. Another example is trace clustering techniques that could be seen as a Process Mining Conformance technique because they use distances among the models for grouping the traces [67].

- Process Mining Enhancement Algorithms: these are tools that extend the information value of a process model using colour gradients, shapes, or animations to highlight specific information in the workflow, providing an *augmented reality* for a better understanding of the process [68], [69], [70].

One of the main issues to be resolved when applying an interactive pattern recognition problem is that experts should understand the inferred model to present corrections and conclude knowledge from the models [48]. As explained, Process Mining technology is Data-Driven, but with a focus on comprehensibility. Following this idea, Process Mining can be defined as *a Syntactic Data Mining technique that supports the domain experts in the proper understanding of complex processes in a Comprehensive, Objective, and Exploratory way* [68]. Thus, Process Mining is a powerful solution supporting healthcare professionals in obtaining individual healthcare processes, and therefore, one of the most suitable technologies for the purpose of this work.

3.2.1. Process Mining: extracting knowledge from EHR

The term Electronic Health Records (EHR) is widely used and refers to the concept of a comprehensive, cross-institutional, and longitudinal collection of a patient's health and care data [71]. It, therefore, includes data that are not only relevant from the clinical perspective or treatment but also to health in general. In this context, the patient could play an active role by accessing, adding, and managing health-related data.

Process Mining [52] uses existing information in clinical databases and EHR to create human-understandable views that support healthcare stakeholders in enhancing their perception of the clinical process. It offers a better understanding of a care process than other Data Mining techniques as explained in Chapter 2. Health processes are structured multidisciplinary care protocols which detail essential steps in the care of patients within a specific clinical problem [5]. In this line, care pathways are complex processes including each stage of the management of a patient with a specific condition over a given period and include progress and outcome details. In that way, care pathways should be understood as a patient's overall journey, instead of isolated functions independently. This is precisely what Process Mining does.

Commonly, Process Mining technologies use a log of actions recorded on a temporal basis to infer workflows that explain the whole process in a human-understandable manner. Traditional Process Mining is based on the use of transactional logs as samples. The information available on those transactional logs is composed of events that should include information about the starting time and, sometimes, the finishing time. This paradigm is called Event-Based Process Mining [72].

Keeping in mind that one of the main objectives of Process Mining is to infer knowledge from data, Process Mining understands data as recorded event logs, where each event refers to a case, an activity, and a point of time to discover, monitor, and improve real processes (see Figure 3.1). In Figure 3.1, each *Event* contains timestamp information about a patient's healthcare episode. A set of events corresponding to the same patient is called *Case*. And finally, a *Log* is a set of cases. As explained, Process Mining Discovery algorithms produce human-understandable flows, as shown in the first graphic of Figure 3.1, whereas the second graph represents the same flow after applying an enhancement algorithm. By using colour gradients for providing information about the duration of the actions or the number of transitions between actions, it becomes possible for example, to visually represent deadlocks and bottlenecks in the system, as well as easily detect the most common paths and activities performed by patients.

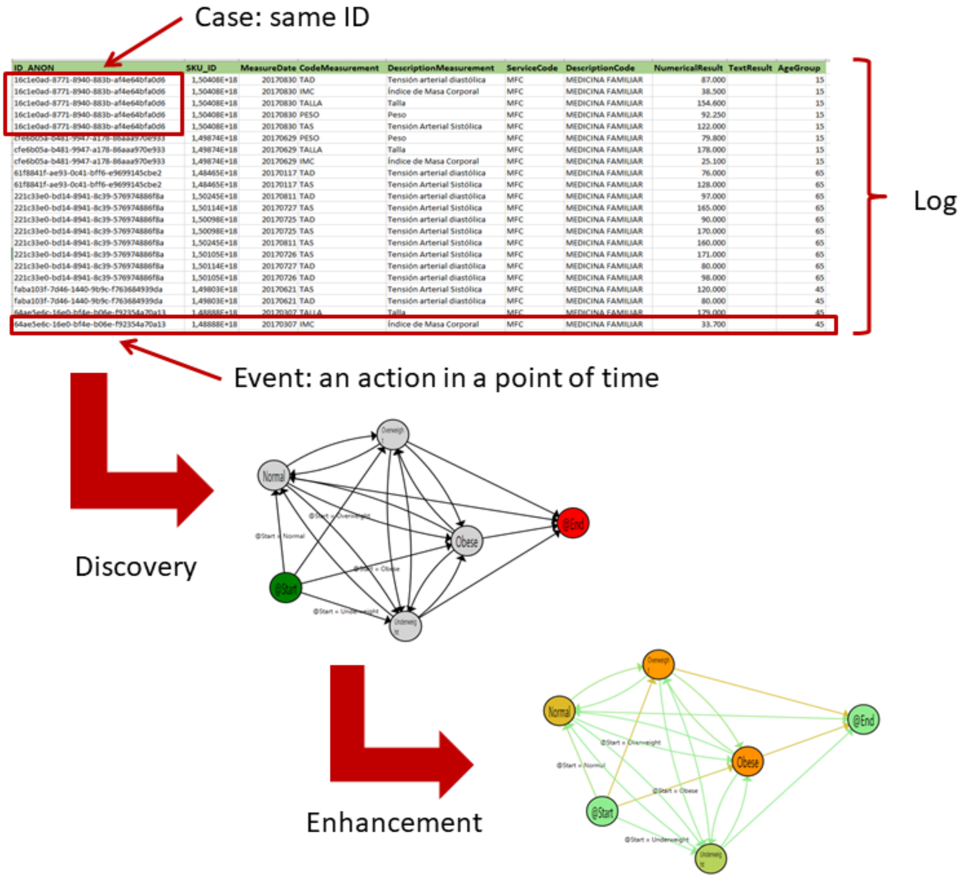
Even if Process Mining is designed to be general-purpose, the healthcare domain represents a segment with significant case studies, as suggested by the review in the area proposed in [73], based on 1278 articles. Process Mining applied to the healthcare sector has its own and relevant challenges. Care pathways are often long and resource-demanding, which complexity is related to the high number of healthcare professionals involved. Similarly, the healthcare processes usually comprise multidisciplinary teams to select the best possible treatment among the different options, based on a diversity of evidence, such as medical visits, imaging data, or laboratory tests. Furthermore, patients perform a role themselves because of their actions, values, beliefs, or fears. All previous variables might cause a deviation from the standard process. Another relevant point to investigate concerns the patients' individual differences in diagnostic and treatment care. Individual differences cause numerous variances in the execution of the healthcare processes. Consequently, it is crucial to consider that medical processes are decidedly dynamic, eminently complex, multidisciplinary, and often ad-hoc [74].

In this context, Process Mining techniques have been successfully used in the healthcare domain from different perspectives and approaches. The work presented in [75] utilised EHR data for evaluating the hospital processes using a Process Mining technique. Specifically, it assessed the effectiveness of the changes in the hospital facility environment before and after the construction of a new building. Another study [76] demonstrated the applicability of Process Mining in healthcare using a practical case of a gynaecological oncology process from three different perspectives, the control flow, the organisational, and the performance. The same authors used Process Mining techniques to extract process-related information for stroke patients in [77]. The authors of this article [55] exposed the possibilities of Process Mining in the characterisation of the emergency process for stroke patients, and how Process Mining technology can highlight the differences between the stroke patients' flow compared with other patients' flow in emergency.

Besides, diseases are not static: they evolve over time in different directions. Process Mining for healthcare can construct individual behaviour models [78], including on the one hand the personal preferences including the social, mental, and health determinants, and on the other hand the variability and evolution of diseases over time. Likewise, the work presented in [79] approached the analysis of the user behaviour using Process Discovery techniques to derive activity models from sensor activation logs in an intelligent environment. With this objective, Data-Driven models are fundamental for supporting the discovery of the patient behaviour process [40].

Within this context, the information included within the EHR [71] could be a perfect candidate to be used with Process Mining techniques to extract knowledge. In EHR, an event represents each entry in the registry for a concrete patient, such as a visit, a laboratory analysis, a recorded variable, etc. All events for the same patient perform a case, easily traceable thanks to the unique patient identifier, and finally, all cases construct the Process Mining Log, as shown in the table included in Figure 3.1.

Going a step forward, Process Mining can also provide a solution for the Data-Driven discovery risk models based on the patients' behaviour and evolution. Risk values of individuals can be interpreted as events of the patient behavioural risk process. With these events, we can create Process Mining Logs that can be used to discover the flow followed by a risk model in the patient's healthcare process. These views could help healthcare professionals understanding behaviours and risk models in a better manner. Consequently, they could extract new evidence based on the correlation of the dynamic behaviour and the adverse outcomes suffered by the patient.



ID_ANON	SKU_ID	MeasureDate	CodeMeasurement	DescriptionMeasurement	ServiceCode	DescriptionCode	NumericalResult	TextResult	AgeGroup
16c1e0ad-8771-8940-883b-af4e64bfa0d6	1,50408E+18	20170830	TAD	Tensión arterial diastólica	MFC	MEDICINA FAMILIAR	87.000		15
16c1e0ad-8771-8940-883b-af4e64bfa0d6	1,50408E+18	20170830	IMC	Índice de Masa Corporal	MFC	MEDICINA FAMILIAR	38.500		15
16c1e0ad-8771-8940-883b-af4e64bfa0d6	1,50408E+18	20170830	TALLA	Talla	MFC	MEDICINA FAMILIAR	154.600		15
16c1e0ad-8771-8940-883b-af4e64bfa0d6	1,50408E+18	20170830	PESO	Peso	MFC	MEDICINA FAMILIAR	92.250		15
16c1e0ad-8771-8940-883b-af4e64bfa0d6	1,50408E+18	20170830	TAS	Tensión Arterial Sistólica	MFC	MEDICINA FAMILIAR	122.000		15
cfe6b05a-b481-9947-a178-86aaa970e933	1,49874E+18	20170629	PESO	Peso	MFC	MEDICINA FAMILIAR	79.800		15
cfe6b05a-b481-9947-a178-86aaa970e933	1,49874E+18	20170629	TALLA	Talla	MFC	MEDICINA FAMILIAR	178.000		15
cfe6b05a-b481-9947-a178-86aaa970e933	1,49874E+18	20170629	IMC	Índice de Masa Corporal	MFC	MEDICINA FAMILIAR	25.100		15
61f8841f-ae93-0c41-bff6-e9699145cbe2	1,48465E+18	20170117	TAD	Tensión arterial diastólica	MFC	MEDICINA FAMILIAR	76.000		65
61f8841f-ae93-0c41-bff6-e9699145cbe2	1,48465E+18	20170117	TAS	Tensión Arterial Sistólica	MFC	MEDICINA FAMILIAR	128.000		65
221c33e0-bd14-8941-8c39-576974886f8a	1,50245E+18	20170811	TAD	Tensión arterial diastólica	MFC	MEDICINA FAMILIAR	97.000		65
221c33e0-bd14-8941-8c39-576974886f8a	1,50114E+18	20170727	TAS	Tensión Arterial Sistólica	MFC	MEDICINA FAMILIAR	165.000		65
221c33e0-bd14-8941-8c39-576974886f8a	1,50098E+18	20170725	TAD	Tensión arterial diastólica	MFC	MEDICINA FAMILIAR	90.000		65
221c33e0-bd14-8941-8c39-576974886f8a	1,50098E+18	20170725	TAS	Tensión Arterial Sistólica	MFC	MEDICINA FAMILIAR	170.000		65
221c33e0-bd14-8941-8c39-576974886f8a	1,50245E+18	20170811	TAS	Tensión Arterial Sistólica	MFC	MEDICINA FAMILIAR	160.000		65
221c33e0-bd14-8941-8c39-576974886f8a	1,50105E+18	20170726	TAS	Tensión Arterial Sistólica	MFC	MEDICINA FAMILIAR	171.000		65
221c33e0-bd14-8941-8c39-576974886f8a	1,50114E+18	20170727	TAD	Tensión arterial diastólica	MFC	MEDICINA FAMILIAR	80.000		65
221c33e0-bd14-8941-8c39-576974886f8a	1,50105E+18	20170726	TAD	Tensión arterial diastólica	MFC	MEDICINA FAMILIAR	98.000		65
faba103f-7d46-1440-9b9c-f763684939da	1,49803E+18	20170621	TAS	Tensión Arterial Sistólica	MFC	MEDICINA FAMILIAR	120.000		45
faba103f-7d46-1440-9b9c-f763684939da	1,49803E+18	20170621	TAD	Tensión arterial diastólica	MFC	MEDICINA FAMILIAR	80.000		45
64ae5e6c-16e0-bf4e-b06e-f92354a70a13	1,48888E+18	20170307	TALLA	Talla	MFC	MEDICINA FAMILIAR	179.000		45
64ae5e6c-16e0-bf4e-b06e-f92354a70a13	1,48888E+18	20170307	IMC	Índice de Masa Corporal	MFC	MEDICINA FAMILIAR	33.700		45

FIGURE 3.1: Process Mining rationale.

3.2.2. Interactive Process Mining

As said, the Interactive approach has clear advantages over other methodologies due to its integration with experts. However, to successfully apply this paradigm, it is required to have a human-understandable focus. It allows professionals to analyse and correct the evidence inferred by algorithms. Notwithstanding, most Pattern Recognition algorithms are machine-oriented and are not human-understandable. Therefore, it is necessary to find an adequate framework that allows the application of the Interactive methodology. The characteristics of Process Mining framework are ideal for being applied in combination with the Interactive paradigm in the understanding of healthcare processes.

As said, one of the main challenges of using an Interactive paradigm is that it requires human-understandable models. Classical Machine Learning tools, such as Neural Networks [80], Support Vector Machines [50], or hidden Markov models [81], are addressed to learn the best accurate models. However, the internal rules that are behind the models are not formulated for being human-comprehensible. Consequently, it is needed to select the appropriate tools and algorithms that aim at the best accuracy and human readability.

Following the visualisation techniques approach, Process Mining can build graphical human-understandable models without an intermediate translation language. On the one hand, it allows the direct understanding of the medical processes, and on the other hand, it enables the straight modification of processes and permits the objective measurement of the effects of the changes. Moreover, this approach supports the medical expert in the models' interpretation and the behaviour of their patients. Besides, it allows them to modify the models according to their experience and measure the effects in a proactive and connected manner, empowering them using fully bidirectional interactive systems.

In this line, the work presented in [9] proposes the application of Process Mining techniques over the Interactive paradigm. Using feasible Process Mining algorithms it could be possible to infer the lifestyle and healthcare processes adhere to patients. This information is appropriately displayed as formal workflows. Health professionals could filter and evaluate these workflows by exploring new medical evidence, using Process Mining discovery and enhancement algorithms. Within the Interactive paradigm, the healthcare experts could also correct possible errors exercising their knowledge, and create formal models or protocols of evidence.

Using this approach, healthcare experts could have an enhanced view of the patient's care process, highlighting the most compelling issues in the flow. To do this, they could apply conformance algorithms with formal scientific evidence, showing deviations of the process followed by the user with the ideal protocol. Furthermore, it is possible to analyse the patient's individualised behaviour and compare it with past inferred workflows. By analysing these views, it is possible to measure, for example, changes in treatment adherence by using individualised Process Mining conformance algorithms or even detect behaviour changes due to psychological illness [78]. Moreover, comparing patients' flows (including behaviour and responses) could allow health professionals to discover similar patients with comparable responses to a specific treatment.

This methodology allows professionals to infer initial formal processes from the available data. Besides, it is self-adapted to the considered population. In each iteration, the system improves the healthcare protocols or models, thanks to the expert involvement, as they could actively provide their knowledge. In the last instance, this could suppose an improvement in the patients' quality of life. By avoiding the *black-box* concept, healthcare professionals could correct the models in each iteration, extract evidence from the results presented by intelligent algorithms, and be confident on them.

3.3. Process Mining tools

Process Mining techniques offer a unique opportunity in the healthcare domain to extract existing information in clinical data sets and improve clinical processes' understanding. Furthermore, using Process Mining with the Interactive paradigm could leverage its use by healthcare professionals to a new level. In the current framework, we can find a set of commercial Process Mining techniques and algorithms, that can be applied to an event log to generate models, tables, and data for analysis, with different results, capabilities, and characteristics in the healthcare domain. However, it is paramount to consider the concrete particularities of the healthcare sector when selecting the most appropriate tool [68].

Classical techniques have been incorporated into a range of Process Mining tools and algorithms for general purposes and concretely for healthcare processes. The work presented in [65] includes a complete review of the Process Mining tools and algorithms most utilised in case studies in the healthcare sector. The study highlighted ProM[1] as one of the most used Process Mining tool, followed by Disco[2], RapidProM[3], and RapidMiner[4], and PMApp. For our problem, PMApp facilitates producing interactive dashboards that respond to the selection of arrows and nodes by capturing Graphical User Interface (GUI) events. It also allows the user to create custom forms and algorithms for discovery, custom filters, and enhancement maps [55]. It permits the traceability of all learning processes, so each activity is continuously associated with single events. In PMApp, it is also possible to render maps that can enhance the discovered model using colour gradients. With this feature, it is possible to render specific maps that highlight particular situations that depend on a customised formulation represented by nodes. All previous characteristics facilitate that health professionals comprehend the processes in a better manner. Furthermore, PMApp framework has been widely tested in real healthcare scenarios, such as in the analysis of the follow-up protocols of patients with diabetes [82, 33]; the measurement and the discovery of the individualised behaviour of older adults at risk of dementia [78]; the characterisation of emergency flows, or for measuring organisational changes effects [55], among other works. As Interactive Process Mining is tool independent other frameworks could be used for the experimentation of this work, however, for the reasons explained PMApp was the framework selected for this work.

PMApp implements the Parallel Activity-based Log Inference Algorithm (PALIA) [83]. PALIA Algorithm infers TPAs (Time Parallel Automaton) [84], which are the mathematical basis of Life Activity Protocols. The PALIA algorithm is detached into five different phases:

- The *Parallel Acceptor Tree Algorithm* step that constructs a graph tree with the corresponding samples, considering both the beginning and the end of the procedure, and their parallelism. As a result, a basic TPA is built admitting only the entry sample.

- During the *Onward Merge* phase all the equivalent branches are merged. The algorithm validates the posterior branch of each node, fusing nodes and transitions when equivalent. Two branches are equal when all the nodes and transitions utilise the same tokens for the same processes.

- The *Parallel Merge Algorithm* step fuses the nodes that are consecutively together and represent the same event. As a result, a corrected TPA is return with the fused nodes and corrected arcs.

[1] http://www.promtools.org/doku.php

[2] https://fluxicon.com/disco/

[3] http://www.rapidprom.org/

[4] https://rapidminer.com/

- In the remaining two steps, corresponding with the *Delete Repeated Transitions* and *Delete Unused Nodes* respectively, the PALIA Algorithm deletes repeated transitions and unused nodes. The outcome is a TPA modelling the system.

The PALIA algorithm uses syntactical pattern recognition techniques to learn TPAs and allows to infer parallel structures representing workflows to solve Activity-Based Process Mining problems.

Concretely, for the experimentation of this doctoral thesis, the implementation of the PALIA algorithm provided by the PMApp tool was used.

3.4. Towards the Interactive paradigm

Within this chapter, it has been described how the use of Process Mining in combination with the Interactive paradigm supposes the application of Machine Learning algorithms in a new manner that could offer an understandable representation of healthcare processes by the professionals in the field. Process Mining technology is Data-Driven with a focus on comprehensibility rather than in the accuracy. However with the application of the Interactive paradigm, models converge faster than conventional ones, moreover as they can be corrected their potential errors is zero. Hence, although Process Mining could lose accuracy in the base algorithm, this is solved by the expert intervention. Furthermore, the Process Mining techniques have been successfully used in the healthcare domain from several perspectives and strategies that make them an ideal candidate to explore their applicability as a solution for Data-Drive discovery of dynamic risk models using EHR recorded information. In this approach, risk values for individuals could be studied as events of the patient behavioural risk process.

For the experimentation, it was selected PALIA algorithm provided by PMApp tool, also tested in real healthcare scenarios, presenting promising results and arising interesting characteristics, such as the possibility of render maps for enhancing the discovered model using colour gradients, the possibility of creating custom dashboard based on the problem to be studied, the statistical concepts included or the feature of clustering.

In this work, we have used real data from different data sets for the three main distinct experiments performed. Consequently, the following three chapters include the description of the data sets, the event log collection, and the pre-processing process particularised for each experiment.

Chapter 4

Modelling a medical condition with Process Mining

As stated in Chapter 2, in the preventive medicine the risk models are statistical tools intended to offer *an individual probability for developing a future adverse outcome in a given period* [18]. However, the **temporal** perspective neither the **patient particularities** are usually considered in this approach. On the one hand, diseases are not static; they evolve towards different destinations, especially when talking about chronic health problems. Similarly, the human being is not invariant; a person is changing throughout her/his biography in age, lifestyle, socioeconomic status, or inter-current diseases among other that affect their health status and their management. On the other hand, patient characteristics should be contemplated for discovering more accurate stratification groups.

Hence a risk model dynamically approached should examine the temporal perspective and maximise the process value based on each group condition and characteristics. This chapter presents the novel definition of *Dynamic Risk Models* and how Process Mining in combination with trace clustering techniques could tackle this new definition. This chapter is enclosed within the objectives O1 and O2 of this work.

4.1. The need for a new behavioural view of risk models

The use of risk models introduces many benefits as they support and complement clinical reasoning and decision-making in medicine, in fact they are playing increasingly important roles as clinical care becomes more tailored to individual characteristics and needs. In the classic approach, a risk model represents the probability of an adverse outcome within a future time period, and then subjects are grouped according to the magnitudes of their assigned risks. In this regard, there is increasing interest in discovering more precise stratification groups that permit a better understanding of the clinical cases [85]. Individual differences cause great variances in the execution of models, consequently patients' characteristics should be considered when approaching this stratification but also including disease variability and dependencies with other conditions, such as comorbidities, social conditions, or age.

Moreover, risk models have associated the underlying disease processes, independently if the focus is on the patient or the professional. They are based on processes, showing the behaviour of the disease and guiding the procedures. In consequence, a process view could be a good approach for addressing risk models, in other words, considering the dynamic perspective.

However, as explained in Chapter 3, in medicine it is needed a theoretical awareness about the disease itself, and to have an updated and precise information about how the process is really happening. It is not just about representing data in a dynamic manner but also a question of what data and how to represent them. Consequently, it is not the same knowing the current state of an individual regarding the process of a disease, that having a temporal representation about previous states that resulted to the current one, and what factors might have affected this evolution. Following the example introduced in Chapter 2 about obesity, the duration of obesity is as important as the current BMI state, as well as the temporal obesity related events that might occur during the patient's journey. In addition, health professionals should be able to inquire what and how they want to know about processes and diseases, and usually these questions are complex and not formalise for Machine Learning systems. Examples of those questions could be, *At what point does a person who accumulates fat begin to have other complications?, Has a person who is losing weight the same comorbidities as a person who is gaining weight, even if they have the same BMI?*

4.1.1. Towards a Data Science framework for Dynamic Risk models: Process Mining & Trace Clustering

Process Mining can construct individual and human behaviour models [78]. This could allow to include not only the individual preferences, and social, mental and health determinants, but also the variability and evolution of a disease over time in the risk models. On the other hand, trace clustering techniques are unsupervised Data Mining solutions that are able to group traces that have similar behaviour, maximising differences with the rest of groups. They could be seen as a Process Mining Conformance technique because it uses distances among the models for grouping the traces [67], supporting us in the dynamic approach for risk models based on the stratification groups that permit a better understanding of the clinical cases. In the framework of the present work, it is clear that extracting information and knowledge from data and discovering patients with different risk behaviours is relevant Modelo. To this end, Process Mining combined with trace clustering techniques can be a solution for that question [67].

Clustering algorithms require the definition of a conventional measure of the similarity between traces, usually called *distance*. The selection of an adequate distance is fundamental for achieving the best grouping for each problem. In our case, the objective is to find the best distance that minimise the difference between traces depending on the risk behaviour evolution for grouping the similar cases, and in consequence, obtaining the risk models that represent different behaviours.

Classical vector-based distances, like Euclidean [15], approach the entry corpus as a geometric vector, assuming that there is no order in the samples. These distances cannot explain the syntactic behaviour of the process and, for that, are not adequate for the problem stated in this work. Besides, distances based on process numerical abstractions can provide a particularly interesting stratification of the processes, but can not maximise the differences in the risk evolution [86], [87].

From the information theory paradigm, there are solutions for measuring similarity between sequences. In the literature, the most common available solutions

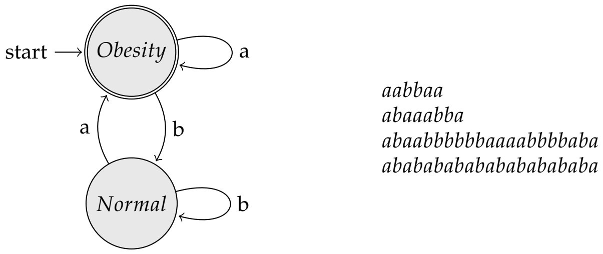


FIGURE 4.1: Example of *Miracle Diets* Risk Pattern

follow the *Edit Distance Paradigm* [88]. Edit Distance Paradigm quantifies the dissimilarity between two series, based on the minimum number of operations required to transform a sequence into another. Occasionally, in some Edit Distance Algorithms is possible to assign different costs to operations depending on the problem to solve. In these cases, the dissimilarity is computed with the aggregation of all the costs produced by the operations applied. There are several algorithms in the literature following this paradigm, such as the classic algorithms like Hamming Distance [89], or the Levenshtein Distance [90], that have been used to measure the differences between traces. Following previous philosophy, but, applying it in the Process Mining field, it is possible to use trace alignments for evaluating the similarity between two traces [91].

To support health professionals in the discovery of new evidence, it is desirable to provide distances that widen the similarity within the traces from an interactive point of view [48]. For that, is crucial a direct relationship between the final models presented to health professionals, and the traces behind the model.

Considering the example illustrated in Figure 4.1, it formulates an hypothetical risk model for a concrete set of patients alternating their BMI risk condition between *Obesity* and *Normal* stages. This risk model could represent patients using a *Miracle Diet*. In the example, value *a* is associated to *Obesity* state, since value *b* is associated to *Normal* state. Sequences in the figure *aabbaa*, *abaaabba*, *abaabbbbbbaaaabbbbaba*, *ababababababababababa* are possible traces of the risk model, representing changes in the person's weight. Using Edit distance algorithms, these traces have high differences. Indeed, out of the model, traces can have a lower distance than traces accepted by the model. For instance, *aca*, where *c* is another BMI state, could be more similar to *aba* than *abababababababababa*. Consequently, previous distances might not be adequate for trace clustering in our problem.

In the literature, other works try to solve this problem by comparing topological information of traces after a discovery. In *Topological Distances*, the process discovery algorithm creates a model for each, such as in [78] and [82] that used Topological Distances. The discovery process performs a generalisation for creating each model in the trace. After that, the Edit Distance paradigm allows comparing the nodes and arrows as graphs. The distance between two traces is computed using the added and deleted nodes and arcs. It enhances the similarity between two traces that refer to the same model, compared to the usual trace Edit Distances for discovery issues. At this point, Topological Distances could use two different concepts to compute the edit distance, time, and weight.

On the one hand, *Weighted Topological Distance (WTD)* maximises similarity in the topology structures of the inferred workflow. On the other hand, *Time Topological Distance (TTD)* takes into account time spent in each activity for computing distances. In this distance, not only nodes are compared, but also time spent in nodes. Consequently, TTD enlarges similarity in cases where patients have the same flows, using similar time in each stage. Previous characteristics made Topological Distances a perfect candidate for grouping patients with similar behaviours in the field of risk models.

Intending to use topological distances for discovering risk models, it is necessary to select a proper Clustering algorithm. There are several Clustering algorithms in the literature [15]. The most common algorithm used is k-Means. It builds k partitions from the entry sample, although its main disadvantage is that the k parameter is fixed, and this forces the expert to primarily decide the number of groups. Quality Threshold Cluster (QTC) is another Clustering algorithm that requires a *quality threshold* to determine the maximum distance among traces in the cluster. This algorithm is computationally more expensive than the k-Means however, it is extended enough for the problem stated in this work, the detection of patients with different risk behaviours for diverse chronic diseases. PMApp tool incorporates both functionalities, allowing the use of topological distances with the QTC algorithm.

4.2. Approaching Dynamic Risk Models

What clinical experts and health professionals are mainly and foremost keen on is monitoring how patients flow through the graph of a risk model, to analyse its accordance and easily identify those groups of patients that did not follow it, with the objective of comprehending the reasons and the related implications. Another relevant point to investigate concerns the patients' individual differences in diagnostic, treatment and care. As explained in Chapter 2, patients' differences cause great variances in the performance of the healthcare processes.

Based on the previous discussion, we present a definition for the Dynamic Risk Model as *the behavioural categorisation of a disease considering the evolution of the associated risk models from a dynamic perspective, permitting a better understanding of the patients' groups progression*. To implement a model in the healthcare domain with such characteristics, it is needed a framework that allows inferring a healthcare process from the available clinical information for a concrete disease, and including patients' characteristics and evolution. The framework should also cover the representation of this information in an appropriate and understandable manner from the point of view of the healthcare professionals, who should be understood not only as the last consumers of the Dynamic Risk Models but also as producers of information thanks to their knowledge and experience. Although a time-series of risk factors measures could not be considered as a complete process, in fact it is an abstraction of the process that provides a very interesting high view that could support the expert in an intuitive understanding of the process.

According to the proposed definition of Dynamic Risk Models and the abstract description of the desired framework, Process Mining and trace clustering techniques reinforce in this procedure. On the one hand, by using existent information in clinical databases to create human-understandable views that support healthcare professionals in enhancing their perception of the clinical process. And on the other hand, by stratifying patients with similar behaviours. Therefore, the next step was to design an experiment that allows us answering if it is possible to obtain a Dynamic Risk Model associated to a disease in the manner we have defined and using the aforementioned Process Mining and trace clustering techniques.

For this purpose, an experiment methodology was defined, represented in Figure 4.2. The experiment proposes two different flows to validate the approach and compare results. On the one hand a classical analysis based on statistical information, and on the other hand, the one that would permit to obtain the Dynamic Risk Model. To answer if it possible to characterise a medical condition process dynamically, we decided to focus the experiment on a well-defined medical condition for a well-known and controlled population. In this case, we chose the characterisation of the malnutrition risk of older adults in a nursing home. This would allow to validate the Dynamic Risk Models from a clinical perspective based on the selected sub-population and the relevant information regarding the considered condition. Thus, the first step of the experimentation comprised the selection of the controlled sub-population. Then, the second step looked for the availability of the needed information from the clinical database, considering both approaches, the classical and the Process Mining one. In this regard, Process Mining needs timestamp information to create the corresponding event, so this second step of the experimentation should consider the extraction of the clinical information with timestamp, such as variables, visits or diagnosis, together with the socio-demographic ones that describes the population to be considered. Therefore, the third step deals with the extraction of the relevant information for the considered use case. The last stage of the proposed experiment examines the data from the two perspective, a classical analysis based on an statistical approach (flow 1), and one using Process Mining and trace clustering algorithms to obtain the Dynamic Risk model for the risk of malnutrition (flow 2). The final step considered the comparison of both results. The subsection below (section 4.2.1) describes the experiment for the concrete case of a population at risk of malnutrition.

4.2.1. Dynamic Risk Model for malnutrition

Malnutrition is one of the major geriatric syndromes and frailty factor [92]. Therefore, it is important that health professionals can assess and follow-up the nutritional status in a proper manner. There are several methods to assess the nutritional status of a person, but it is outstanding to use a simple, reliable, easy-to-use, fast, economic, consistent, and sensitive ways to identify all or almost all older patients at risk of malnutrition. The Mini Nutritional Assessment (MNA®) is a validated nutrition screening and assessment tool that can identify geriatric patients age 65 and above who are malnourished or at risk of malnutrition. The MNA was developed nearly 20 years ago and is the most well-validated nutrition screening tool for the elderly [93]. Originally comprised of 18 questions, ans it is easily completed within 10 to 15 minutes time, but in some situation there is a need of shorten this time. For that reason Rubenstein and colleagues [94] developed a six question MNA short-form by identifying a subset of questions from the full MNA that had high sensitivity, specificity and correlation to the full test [95].

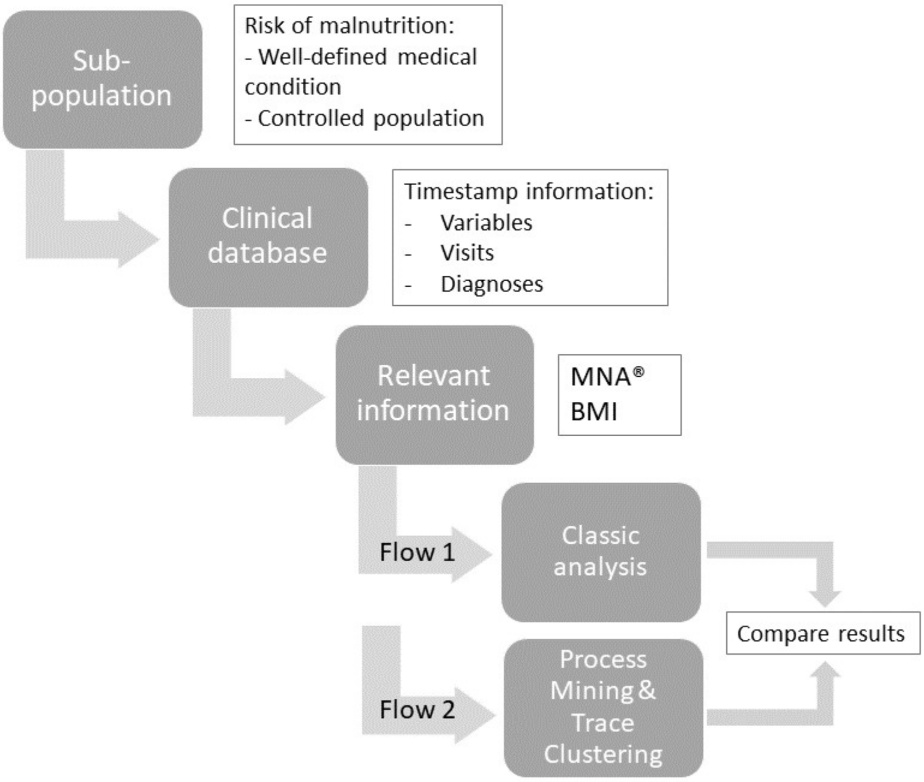


FIGURE 4.2: Experiment methodology for the malnutrition study.

The MNA short-form test, hereinafter MNA, is composed by simple measurements and brief questions collecting data about: anthropocentric measurements (weight, height and weight loss), global assessment (questions related to lifestyle, medication and mobility), dietary questionnaire (questions related to number of meals, food and fluid intake, and feeding autonomy) and subjective assessment (self-perception of health and nutrition). Table 4.1 includes the six questions included in the MNA together with the responses options. The total score provides the nutritional classification, a total score of MNA < 8, $8 - 11$, and > 11 indicates malnutrition, risk of malnutrition, and no malnutrition, respectively.

Following the experiment proposed in Figure 4.2 the second step comprised the selection of the clinical database with relevant information for the considered problem. For that purpose, we used real data from a nursing home flow of residents who were admitted between January 2015 and December 2016. The log information was acquired from the nursing home information system and an Android based app, named NutriPro App, containing the MNA questionnaire and a set of recommended nutritional interventions based on the questionnaire score [96]. A single-cohort study was driven from June 2016 to December 2016 in the nursing home centre. Nutritional status was assessed by means of the app based two times: at the baseline and after six months.

As said, all data was collected through NutriPro App and the nursing home information systems. Enrolled subjects gave their informed consent to participate in the study and ethical access to the confidential patient data was obtained through the ethical process reached in the framework of the project in which the research was done. The evaluation methodology also included the assessment of the NutriPro App as a screening tool implementing the MNA test, the impact of the intervention program and the impact of the NutriPro App on the professionals work. Observable variables are included in Table 4.2.

A total of 154 subjects were screened at the baseline and at month six of the study. The first data of these subjects was collected between June and July, 2016. Table 4.3 provides the descriptive analysis of the subjects included in the study.

The third phase of the experiment deal with the selection of the relevant information for the considered condition, in this case the risk of malnutrition. In this regard, the F1 and F2 questions of the MNA test (see Table 4.1) are anthropometric measures, as they involve the Body Mass Index (BMI) and the calf circumference of the person. Thus, the BMI measures together with the MNA could be collected in a screening programme dedicated to detect population at risk of malnutrition in a temporal basis, therefore were the relevant variables considered in our experimentation.

A classic approach to assess malnutrition using theses measures usually treats them as static ones, with no information about patients' evolution and pathway. Within the study, it was performed an analysis of the collected data that comprised a static cohort of the BMI and the MNA nutritional status. At baseline we had 154 subjects, from which 24 were malnourished (15.5% MNA$\leq$7), 64 were assessed as being at risk of malnutrition (41.9% 8$\leq$ MNA$\leq$11) and 66 were scored as normal nutritional status (42.6% MNA$\geq$12). During the study lifetime 49 patients were exitus (31.8%), so at month six we had data from 105 subjects. Table 4.4 includes the data collected at baseline and month six for the 105 residents who finished the study.

TABLE 4.1: Mini Nutritional Assessment MNA®

Screening
A Has food intake declined over the past 3 months due to loss appetite, digestive problems, chewing or swallowing difficulties? 0 = severe decrease in food intake 1 = moderate decrease in food intake 2 = no decrease in food intake
B Weight loss during the last 3 months 0 = weight loss greater than 3 kg 1 = does not know 2 = weight loss between 1 and 3 kg 3 = no weight loss
C Mobility 0 = bed or chair bound 1 = able ti get out bed / chair but does not go out 2 = goes out
D Has suffered psychological stress or acute disease in the past 3 months? 0 = yes 2 = no
E Neuropsychological problems 0 = severe dementia or depression 1 = mild dementia 2 = no psychological problems
F1 Body Mass Index (BMI) (weight in kg) / (height in m)^2 0 = BMI less than 19 1 = BMI 19 to less than 21 2 = BMI 21 to less than 23 3 = BMI 23 or greater If BMI is not available, question F1 can be replaced with question F2
F2 Calf circumference (CC) in cm 0 = CC less than 31 3 = CC 31 or greater

TABLE 4.2: MNA experiment: description of data

Observable variables
Demographic data: gender and age
Anthropocentric data: weight and height
Mobility status: in bed, wheel chair, normal
Medication
Pathology
Specific diet

TABLE 4.3: Characteristics of the MNA study sample

	Women	**Men**	**Total**	p
Subjects (n)	110	44	154	
Age (years)	84.5±12.8	81.0±9.9	83.5±12.3	0.109
Weight (Kg)	60.0±13.9	70.4±16.2	62.9±15.1	<0.001
Height (cm)	150.1±16.3	165.3±8.7	155.3±10.0	<0.001

TABLE 4.4: MNA Nutritional status of patients during the study.

MNA Nutritional Status	**MNA0 (M0)**	**MNA1 (M6)**
Malnourished	14 (13.3%)	11 (10.5%)
At risk of malnutrition	52 (49.5%)	45 (42.9%)
Normal nutritional status	39 (37.1%)	49 (46.7%)

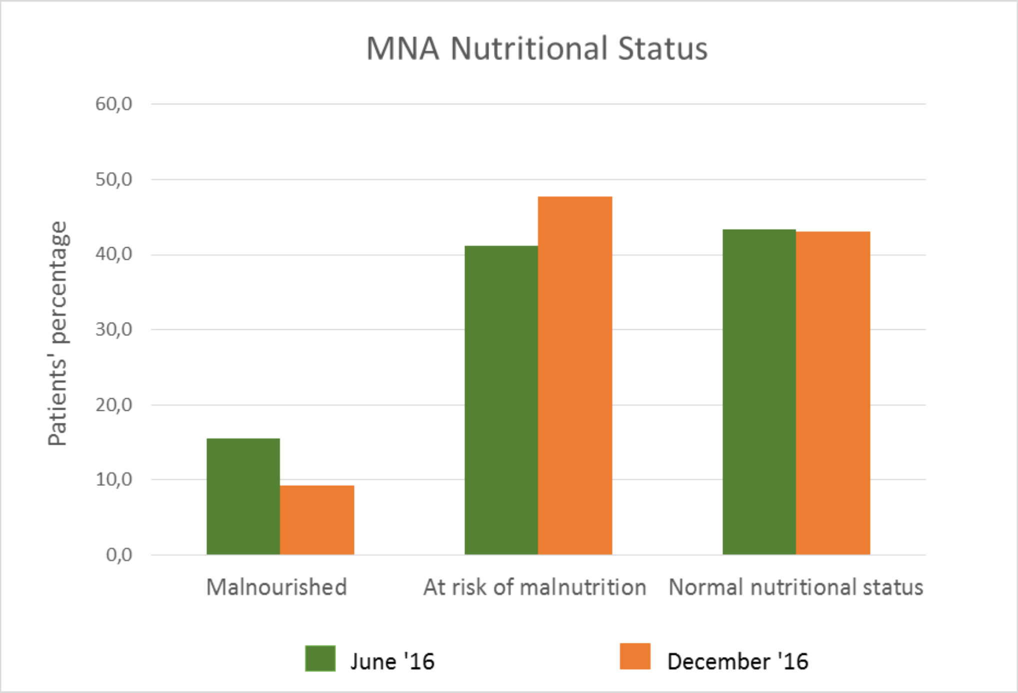


FIGURE 4.3: MNA status at baseline and six months later.

Following with a classic approach, Figure 4.3 shows the scores in MNA obtained in subjects evaluated at the baseline (June'16) and after month six (December'16).

Based on the first assessment, healthcare professionals working in the nursing home received some recommendations to improve the nutritional status of the older adults participating in the study.

After the six months, it was observed a significant decreased in malnourished scored residents, and the consequent increasing in the percentage of subjects scored as in risk of malnutrition by the MNA results. However this static analysis did not reflect the behaviour of the patients, the possible relationship between the BMI and the MNA scores, or if patients could be modelled based on their evolution to personalise the given recommendations and to improve their nutritional status. Moreover, BMI shows weakness in the limited ability to measure small changes over time as comprehensive assessment tool of nutritional status, it does not take into account the changes in body composition that occur with age. Additionally, this analysis from a static perspective does not include variability over time, and does not show any behaviour among the different patients or groups of patients, in fact health professionals do not know what is actually happening between several MNA or BMI measures.

Based on the proposed experimentation (Figure 4.2), the second flow considered the analysis of the available data using Process Mining and trace clustering algorithms to obtain a possible Dynamic Risk Model for malnutrition. Similarly to the classical analysis, from the initial 154 subjects, we discharged the 49 patients who exited during the study duration, together with nine more residents who moved to

TABLE 4.5: Data description for the Process Mining analysis for malnutrition.

Column Name	Data type	Example
Patient ID	Global unique identifier	56fab0adbaa18e943657ac2b
Gender	String	Male
Nursing Home ID	String	NAME
Birth Date	Date	12/27/2016
Weight	Float	98.2
Weigh measure date	Date	02/04/2016
Height	Integer	170
Body Mass Index	Float	33.98
Number of MNAs	Integer	3
First MNA date	Date	06/22/2016
First MNA score	Integer	5
Patient age at first MNA	Integer	73
Last MNA date	Date	09/20/2016
Last MNA score	Integer	6
Patient age at last MNA	Integer	74

another nurse home during the six months of the study. Therefore, data from 96 participants (67 female and 29 male) were analysed. The event log used for the Process Mining analysis included the information described in Table 4.5.

We performed some data processing to adapt data to the needed Process Mining format. We considered the BMI as the main event for the analysis, and the MNA result as the aggregated data. Therefore, the BMI value was considered as the activity or the action, the starting of the action was associated with the *Weight measure date*, and the finish time was established as the next BMI measure date.

We used two different cut-off points to determine the activity name, the WHO criteria [97] and the age-related criteria developed in 1989 by the Committee on diet and Health in the USA [98]. This was because of the controversial about whether use of this international BMI classification criteria is the most appropriate for monitoring the nutritional status of older subjects [99]. Consequently, we used both in our experimentation to validate if the cut-off points are also affecting the distribution of the processes. Table 4.6 includes the BMI classification for both criteria.

With this information, PALIA algorithm, implemented in PMApp, was executed for both BMI classification criteria to obtain the patients' pathway based on their BMI behaviour. Figures 4.4 and 4.5 represent the results obtained with the graphical representation of all patients' process regarding their BMIs during the study, not as statics but as patients' path, including the average duration in each node and transition [100]. Nodes labelled as *@Start* and *@End* represent the starting and ending points within the process, respectively. Nodes and arrows were coloured to enrich them and easy the overall interpretation as follows, nodes were coloured by the average time spent in the action (node), and edges were painted with a gradient symbolising the number of persons that proportionally follow the corresponding transition,

TABLE 4.6: BMI: WHO and age-related classification and cut-off points.

BMI WHO classification		BMI Age-related classification	
BMI	**Nutritional status**	**BMI**	**Nutritional status**
Below 18.5	Underweight	Below 24	Below the norm
18.5 - 24.9	Normal weight	24.0 - 28.9	Normal weight
25.0 - 29.9	Overweight	Above 29	Above the norm
Above 30	Obesity		

both from green (minimum value) to red (maximum value), using the gradient scale shown in Figure 4.6.

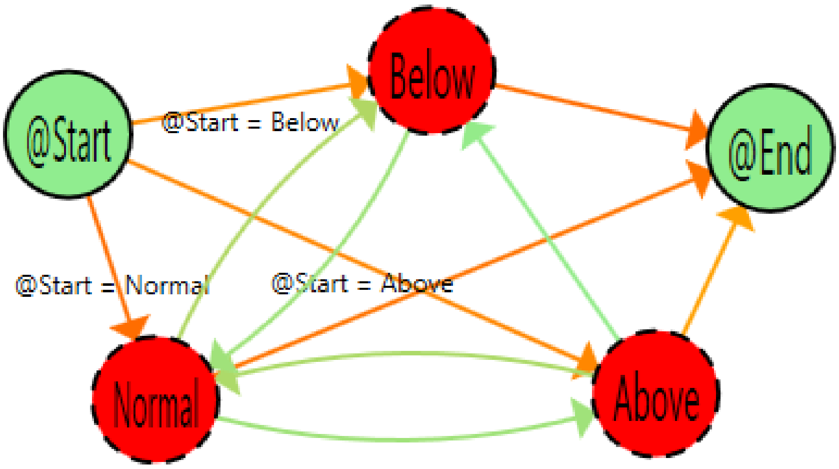


FIGURE 4.4: Malnutrition process using age-related classification of BMI for a nursing home population.

Focusing on the flow for the BMI age-related classification (Figure 4.4), it was observed that the time spent in the three stages, *Below*, *Normal*, and *Above* were almost equal, showing the nutritional status per se did not provide enough insights about the process or the patients' behaviour. Among the three stages there was population classified as normal, at risk or malnourished by the MNA results. Similarly, the flow for the WHO BMI classification (Figure 4.5) showed the same path but considering another BMI cut-off points, with a greater variability due to the tinny cut point of the criteria, demonstrating how the results were affected by the cut-off points considered, and the variability that this might introduce. However, in both cases, it was interesting to discover similarities among the population considered to establish common patterns that helped to classify the population and validate interventions.

Therefore, once the dynamic perspective was incorporated to the model, the focus was on the second part of the Dynamic Risk Models definition, *they present the risks based on the stratification groups that permit a better understanding of the clinical*

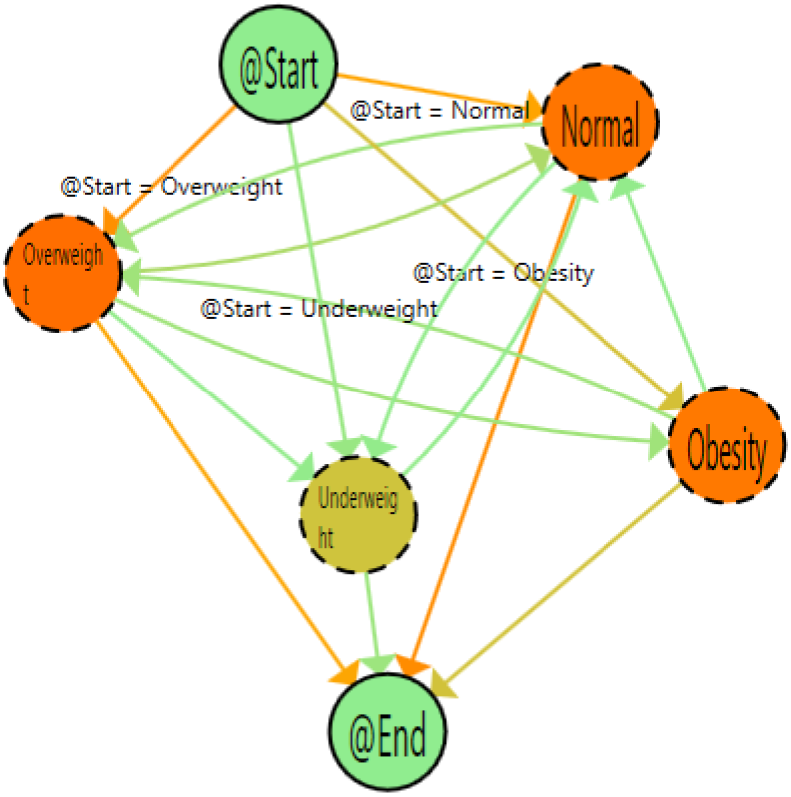


FIGURE 4.5: Malnutrition process using WHO classification of BMI for a nursing home population.

cases, by applying the trace clustering techniques using the QTC algorithm implemented in PMApp. The objective was to stratify the considered population into different conducts regarding their BMI and MNA results. The QTC algorithm requires a quality threshold to decide the maximum distance among traces in the cluster, where a minimum distance (0.0) means equal flows, and a maximum distance (1.0) produces traces with no common activities. Several threshold distances were used with the objective of obtaining the most significant groups of patients, balancing between the number of groups and the behaviour showed within each group, looking for clear patterns regarding the BMI evolution. Most significant result were obtained for a threshold of 0.2 with eight groups representing the most common behaviours for the BMI WHO classification criteria, and five groups for the BMI age-related classification criteria. Results for the the BMI age-related classification criteria are presented in Figure 4.7 and Table 4.7, including the different discovered models for the 95 patients plus one patient considered as outlier, because these patients could not be included in any of the found models. Table 4.7 also includes the distribution among the different groups considering the gender, and the nutritional status based on the first and last MNA scores.

Looking into the models, model A0 (see Figure 4.7a) with 33 patients characterises patients following a common behaviour of being above a normal weight

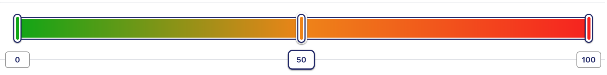


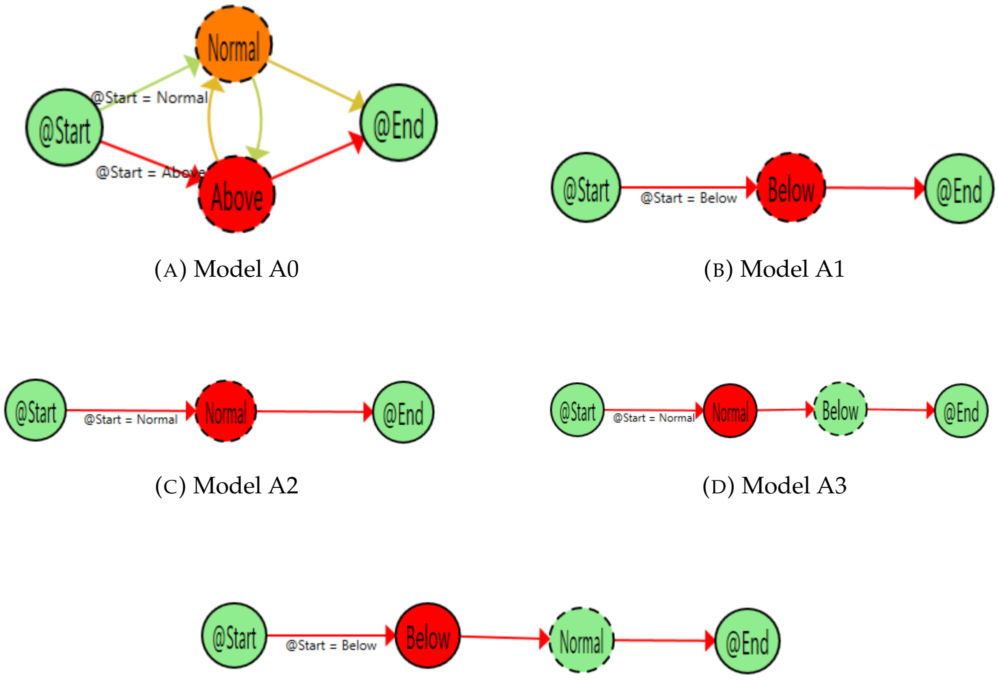
FIGURE 4.6: Gradient scale key for model representation: green to red.

TABLE 4.7: Models for BMI and MNA nutritional status results, age-related classification criteria. Presented in [100].

Group	MNA Assessment								
	MNA0			MNA1			Gender		
Model A0 N=33 (34.4%)	Normal Risk	20 13	60.61% 39.39%	Normal Risk Malnourished	22 9 2	66.67% 27.27% 6.06%	Female Male	23 10	69.70% 30.30%
Model A1 N=26 (27.1%)	Risk Malnourished Normal	16 9 1	61.54% 34.62% 3.85%	Risk Normal Malnourished	17 6 3	65.38% 23.08% 11.54%	Female Male	19 7	73.08% 26.92%
Model A2 N=24 (25.0%)	Normal Risk Malnourished	12 11 1	50.00% 45.83% 4.17%	Normal Risk	12 12	50.00% 50.00%	Female Male	16 8	66.67% 33.33%
Model A3 N=8 (8.3%)	Risk Malnourished	6 2	75.0% 25.0%	Risk Malnourished Normal	4 3 1	50.0% 37.5% 12.50%	Female Male	5 3	62.50% 37.50%
Model A4 N=4 (4.2%)	Risk Normal	3 1	75.0% 25.0%	Normal Risk	3 1	75.0% 25.0%	Female	4	100%

or changing between normal weight and above the norm during the study. It is hard to mention that red transitions show the average number of patients following this path, indicating the most common behaviour was patients starting and ending above the norm. Model A1 (see Figure 4.7b) represents patients with a common behaviour of being below a normal weight all the study duration, although the number of patients classified as malnourished decreased, comparing the results from the first and the last MNA, since the first MNA detected 34.62% of patients suffering malnutrition, while the second one identified 11.54% of patients in this situation (see Table 4.7). Model A2 (Figure 4.7c) includes normal weight behaviour, it means patients that were in a normal weight all the study lifetime, even though their MNA assessment varied among the three nutritional status (see Table 4.7). Model A3 (Figure 4.7d) represents a common pathway for patients losing weight, as patients within this cluster evolved from a normal weight to a weight below the norm, with MNA results of being at risk or malnourished. Contrary, model A4 (Figure 4.7e) shows common behaviour of gaining weight, including patients that followed the pathway from below to normal weight, supported by the fact that most of the patients included in this group are reversing their risky situation, as MNA results show (see Table 4.7).

Following the analysis, Table 4.8 and Figure 4.8 include the results corresponding with the eight behavioural groups for the BMI WHO classification criteria. Concretely, Table 4.8 represents the distribution among the different groups considering



(A) Model A0

(B) Model A1

(C) Model A2

(D) Model A3

(E) Model A4

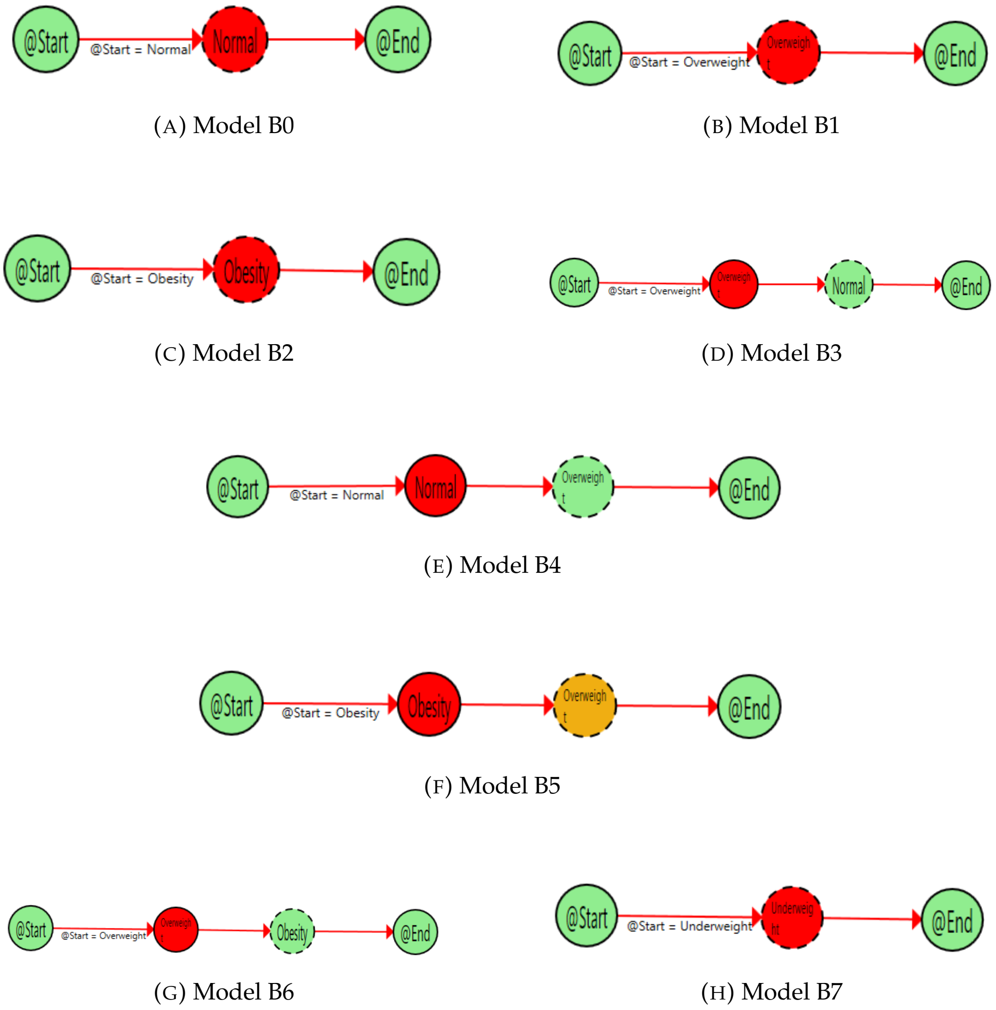
FIGURE 4.7: Behavioural Models for BMI - age-related classification criteria. Presented in [100].

the gender and the nutritional status based on the scores from the first and second MNA test. Whereas Figure 4.8 shows the discovered models for the 96 patients, of which 92 have been represented by the models and four were grouped as outliers, as the clustering algorithm could not include them into any of the models found, as they did not follow a similar behaviour (see Figure 4.9).

Considering each model, model B0 (see Figure 4.8a) represents patients with a common behaviour of being in normal weight, despite their scores for the two MNAs, that show 75% of them were classified as at risk of malnutrition in the first MNA, and 60% in the second. Regarding model B1 (see Figure 4.8b), it includes patients with a stable weight pattern of having overweight during the study lifetime, even if some of them were classified at risk of malnutrition, concretely the 34.78% at the first MNA, and 39.13% at the second one, that was a considerable number within the group. Model B2 (see Figure 4.8c) showed patients following also a stable weight behaviour of obesity, surprisingly one of three in this group were classified at risk of malnutrition based on their score in both MNA tests, with no improvements or changes both in their weight nor their nutritional status. This group shows a different intervention should be put in practise with these patients as no behavioural changes were perceived in their risky situation. Models B3 and B4 include the behaviours for patients losing weight, differentiating between those moving from overweight to a normal weight at the end of the study in model B3 (see Figure 4.8d), and those losing weight from obesity to overweight in model B4 (see Figure 4.8e). Two other groups represent the contrary behaviour, it means gaining

TABLE 4.8: Models for BMI and MNA nutritional status results, WHO classification criteria. Presented in [100].

Group	MNA Assessment								
	MNA0			MNA1			Gender		
Model B0	Risk	21	75.0%	Risk	17	60.71%	Female	21	75.0%
N=28	Malnourished	6	21.43%	Normal	9	32.14%	Male	7	25.0%
(29.2%)	Normal	1	3.57%	Malnourished	2	7.14%			
Model B1	Normal	15	65.22%	Normal	12	52.17%	Female	16	69.57%
N=23	Risk	8	34.78%	Risk	9	39.13%	Male	7	30.43%
(24.0%)				Malnourished	2	8.7%			
Model B2	Normal	11	68.75%	Normal	11	68.75%	Female	10	62.5%
N=16	Risk	5	31.25%	Risk	5	31.25%	Male	6	37.5%
(16.7%)									
Model B3	Risk	6	66.67%	Risk	5	55.56%	Female	5	55.56%
N=9	Malnourished	2	22.22%	Normal	2	22.22%	Male	4	44.44%
(9.4%)	Normal	1	11.11%	Malnourished	2	22.22%			
Model B4	Risk	3	60.0%	Normal	3	60.0%	Female	4	80.0%
N=5	Normal	2	40.0%	Risk	2	40.0%			
(4.2%)									
Model B5	Risk	3	60.0%	Risk	3	60.0%	Female	4	80.0%
N=5	Normal	2	40.0%	Normal	2	40.0%			
(4.2%)									
Model B6	Normal	2	50.0%	Normal	4	1000%	Female	3	75.0%
N=4	Risk	2	50.0%				Male	1	25.0%
(4.2%)									
Model B7									
N=2	Malnourished	2	100%	Risk	2	100%	Female	2	100%
(2.1%)									



(A) Model B0

(B) Model B1

(C) Model B2

(D) Model B3

(E) Model B4

(F) Model B5

(G) Model B6

(H) Model B7

FIGURE 4.8: Behavioural Models for BMI - WHO classification criteria. Presented in [100].

weight. Model B5 shows a population gaining weight from a normal situation to a overweight state (see Figure 4.8f), and model B6 from overweight to obesity state (see Figure 4.8g). Finally, it was discovered another stable pattern for patients with an underweight situation during the study duration, represented in model B7 (see Figure 4.8h), supported by the MNA results that showed a malnourished and risky situation for all in the first and second screening, respectively.

4.3. A novel perspective on Risk Models

The results from this experiment permitted to obtain the Dynamic Risk Model associated to a population at risk of malnutrition considering their BMI and MNA test results [100]. The model clearly showed that BMI could be considered as a dynamic process, and how this process really works. Moreover, the behavioural models found new evidences that the first or classic study could not show. This

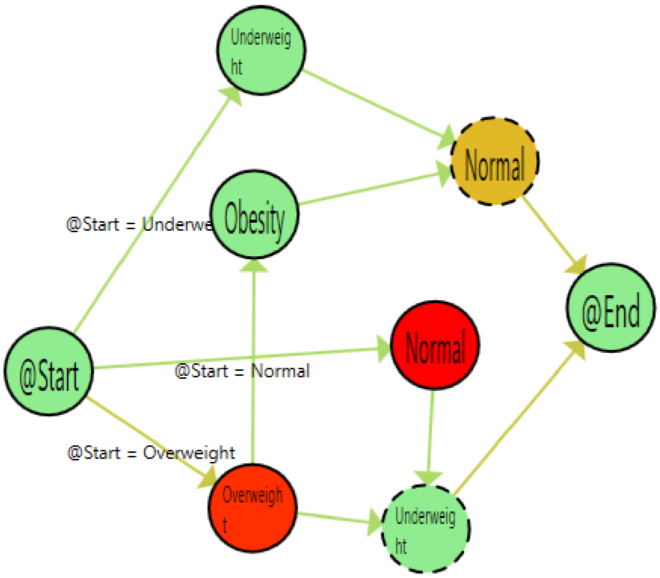


FIGURE 4.9: Outliers group, BMI - WHO classification criteria.

new approach permitted the healthcare professionals working in the nursing home and involved in the study to differentiate among patients based on their weight behaviours, and consequently to personalise nutritional interventions based on the different groups of patients. Furthermore, the dynamic perspective of the discovered flows allowed the validation of the nutritional interventions. On the other hand, the aggregation of the dynamic models in combination with the MNA scores stand that generally malnutrition state was reversed for those patients underweight but not always for patients above a normal weight, obese or overweight. This was new evidence for the healthcare professionals participating in the study, as they did not aware of this fact until the evaluation of the aforementioned results.

The analysis of the result also allowed to detect patients with a MNA score of being malnourished within all models, showing how malnutrition is related not only with weight but also with other factors, such quantity and quality of nutrients and the level of activity, that should be considered when designing an successful intervention. As proof, malnutrition status and BMI depend on the cut-off points selected, what might cause variability in the results and less understandability. However, the discovered models already included the variability over time and the evolution patterns, allowing to see the overall process and flow, and helping health professional understanding them. The models also recognise the patients' evolution according to the BMI trend instead of by cuts, and here it is precisely where Process Mining could provide an actual value for the identification of the older adults who will benefit from this screening.

From the analysis of these outcomes, we can see how Process Mining techniques could enable the analysis of patients' patterns, combining BMI measures with MNA results. The use of Process Mining in combination with trace clustering techniques facilitated the generation of different groups of patients with clear behaviours representing a new Dynamic Risk Model. Obtained graphical models were easy to understand by the health professionals. Even more, the models supported the generation of groups to represent the course flow followed by patients regarding weight changes, differentiating certain behaviour's from others. It is worth to mention that the sample size was considerable small to be significant and the study period was also short for the considered variables, however it permitted to demonstrate the Process Mining potentiality to continue the research in the field and to extend the analysis to other health conditions, concretely in the field of chronic diseases that is presented in the following chapters of this work.

Chapter 5

Dynamic Risk Models with PM applied to chronic conditions

In the previous chapter, it was presented how Process Mining techniques could be used to generate dynamic models to achieve a better understanding of the care processes associated to a medical condition. Combining BMI measures and MNA results, the outcomes generated different behavioural groups regarding the weight changes [100].

Going a step forward, the aim of this chapter is to examine if the behavioural modelling of a chronic health condition using Process Mining techniques is possible using the information collected and stored in the EHR. Chronic conditions comprise several characteristics that made them a perfect candidate to study the possibilities of discovering understandable models that explain the underlying process, obtaining what we defined as Dynamic Risk Models. Therefore, the work presented in this chapter is framed in the context of the objective O2.

5.1. Dynamic Models Supporting Chronic Disease Management

Chronic conditions are *diseases of long duration, with a long period of supervision, observation and care, that are non-reversible*, as they were defined in the framework of this doctoral thesis (see section 2.1.1). To illustrate the problem and solution to tackle in the case of chronic diseases, it was necessary to personalise the experiments to specific chronic conditions to investigate whether an approach similar to the one followed in the previous chapter (see Chapter 4), could be used for representing Dynamic Risk Models that characterise them based on the evolution of the considered condition using Process Mining and trace clustering techniques. The main idea was to model a chronic disease process abstraction through a human-understandable graphical representations that could support healthcare professionals in a better comprehending of the chronic disease processes, as it considers the disease's variability over time and the patient nature. For this reason, it was followed a bottom-up process abstraction, where a simple risk model was considered, and then the methodology was expanded to more complex variables.

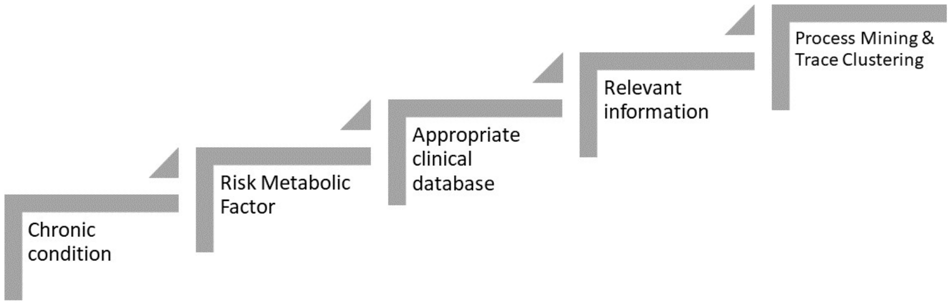

Chronic condition
Risk Metabolic Factor
Appropriate clinical database
Relevant information
Process Mining & Trace Clustering

FIGURE 5.1: Experiment framework for chronic conditions.

5.1.1. Experiment framework

Following the experimentation presented in section 4.2, but adapting it to the concrete case and characteristics of chronic conditions, we proposed a new experimentation framework in Figure 5.1. Overall, the first step comprised the selection of the corresponding chronic condition to be analysed, secondly the associated risk metabolic factor that should be considered for the chronic condition. During the third step, it is searched and selected the most appropriate clinical database where to extract the relevant information for the chronic condition, and the last step includes the Process Mining and trace clustering analysis to discover the *Dynamic Risk Models*.

Going into depth in each stage of the framework process, we needed to select the chronic condition over which to apply the experiment. As introduced in Chapter 2, the global leading chronic diseases are elevated blood pressure or hypertension, overweight and obesity, and raised blood glucose or hyperglycemia. Measures for these three conditions are usually taken in routine clinical practise and collected in the hospital databases, thus easily available for our experimentation. Moreover, their measures –blood pressure, weight, and blood sugar– are classically approached by static snapshots while they have imply a dynamic component, and they are always relevant for the management of the associated chronic condition. Concretely, blood sugar is usually examine in overall routine laboratory analysis, whereas blood pressure is also often taken in ordinary visits, although the patient is not suffering any specific disease. This fact suggests, in principle, a wide availability of information for both variables in clinical databases. Thus, we selected Hypertension and Hyperglycemia, and their associated risk metabolic factors –blood pressure and blood sugar– for the experiments in this second phase of the doctoral thesis, corresponding with the first and second steps of the experiment procedures. As already mentioned in Chapter 2, metabolic risk factors associated to chronic conditions are used to define the risk models. It is therefore correct to assume the same approach in our case, considering both the chronic disease itself and the associated risk factors, in our case case, the pair hypertension with the blood pressure, and hyperglycemia with the blood sugar.

The third step corresponded with the selection of the suitable clinical database. Within the clinical information, data stored in the EHR supposes a very valuable

TABLE 5.1: Data sample size description for hypertension and hyperglycemia experiments.

Age Group	Population	Total %
15	498	1%
20	1838	3.66%
25	2075	4.13%
30	2752	5.48%
35	3919	7.81%
40	4209	8.39%
45	3821	7.61%
50	3692	7.36%
55	3499	6.97%
60	3509	6.99%
65	3879	7.73%
70	4345	8.66%
75	3699	7.37%
80	3381	6.74%
85	1967	3.92%
90	1193	2.38%
95	863	1.72%
100	521	1.04%
>100	536	1.07%

and detailed information about each patient regarding the journey through the care she/he receives in a timestamp manner. In collaboration with a tertiary hospital in Valencia, we had access to a real retrospective data from the EHR to demonstrate the possibility of generating Dynamic Risk Models for chronic diseases. The data was extracted from 2012 to 2017, from 50,196 unique patients as described in Table 5.1. The ethical approval to the confidential patient data was obtained through the ethical process reached by the framework of the project in which the research was done. As all data were anonymised prior to the extraction by the hospital IT department, and as it was used retrospective data, therefore, the information consent was not needed in this case. Extracted data enclosed information from the primary care service, emergency, outpatient, and morbidity diagnosis service, as described in Table 5.2. The hospital experts provided the data in several Comma-Separated Values (CSV) files, concretely one CSV file per table included in Table 5.2, where values were represented in a set of rows and columns.

For the fourth and fifth steps different experiment flows were performed for the two considered chronic condition, sub-sections below (sections 5.1.2 and 5.1.3) describe the particularities of each experiment and the discovered Dynamic Risk Models for both conditions.

TABLE 5.2: Database information for hypertension and hyperglycemia experiments.

Table	Description	Unique Patients/ Observations	Period
Patients Anonymize	Patients general information: age, identifier, some diagnoses	50,196	-
Primary Care	Primary consultations' data: variables and annotations	17,853/215,523	2017
Hospital Admissions	Type of admission, ICD9[a], Diagnostics, DRG [b], date	10,403/180,797	2012–2016
Emergency	Severity, admission service code, destination service, date	34,054/180,797	2010–2017
Outpatient	Provision type, date	6667/706,888	2012–2017
Morbidity Diagnoses	ICD9 [a] code, diagnose date	48,080/1,048,575	2012–2017
Laboratory	Laboratory measures: date, id, description, result, units	50,196/18,182,239	2012–2017

[a] International Statistical Classification and Related Health Problems, [b] Diagnosis-Related Group.

TABLE 5.3: Patients Anonymize dataset description for hypertension and hyperglycemia experiments.

Column Name	Data Type	Example
ID_ANON	Global unique identifier	000269d4-b40a-df4f-a1c0-56db3f989ad2
Age Group	Integer–group of age by 5 years	40
Overweight	Integer: 1/0, overweight diagnose	0
Obesity	Integer: 1/0, obesity diagnose	1
Unspecified Overweight/Obesity	Integer: 1/0	1

5.1.2. Dynamic Risk Model for Hyperglycemia

Raised blood glucose or hyperglycemia is directly related to diabetes. Diabetes is a chronic and metabolic disease characterised by elevated levels of blood glucose, which leads over time to serious damage to the heart, blood vessels, eyes, kidneys, and nerves [101]. Hyperglycemia is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems. Fasting hyperglycemia or fasting blood glucose is defined as when you do not eat for at least eight hours, and the standard for measuring blood glucose is *mg/dL* which means milligrams per decilitre. The expected values for normal fasting blood glucose or fasting plasma glucose (FPG) concentration are between 70 mg/dL (3.9 mmol/L) and 100 mg/dL (5.6 mmol/L). When FPG is between 100 to 125 mg/dL (5.6 to 6.9 mmol/L) changes in lifestyle and monitoring glycemia are recommended. If FPG is 126 mg/dL (7 mmol/L) or higher on two separate tests, diabetes is diagnosed [101]. The difficulty with defining normality mirrors is related with the definition of the diagnosis based on cut-points for intermediate hyperglycemia that is, placing a specific cut-point on a continuous variable. Furthermore, other factors such as age, gender, and ethnicity are relevant to defining normality. Since there are insufficient data to accurately define normal glucose levels, the term *normoglycemia* [102] should be used for glucose levels associated with low risk of developing diabetes or cardiovascular disease, which represents levels below those used to define intermediate hyperglycemia [102].

The fourth step of the experimentation included in Figure 5.1 comprises the selection of the relevant information from the database, in our case from the data stored in the EHR. This process included the needed information for the creation of the hyperglycemia models, therefore patient characteristics such age and global identifier were extracted from *Patients Anonymize* set, and the FPG values were obtained from the *Laboratory* set, the description of these dataset is included in Tables 5.3 and 5.4 respectively.

Previous to continue with the experiment procedures, the standard deviation and the average of the measures included in the *Laboratory* dataset for the FPG were

TABLE 5.4: Laboratory dataset description for hypertension and hyperglycemia experiments.

Column Name	**Data Type**	**Example**
ID_ANON	Global unique identifier	000269d4-b40a-df4f-a1c0-56db3f989ad2
Test Request Date	String	20170830
Test Result Date	String	20170830
Test Id	Integer—test identifier	561
Test Description	String—measure description	Lipid index
Test Result	Float—test result	22.2
Test Units	String—code of the units	mg/dL
Age Group	Integer—group of age by 5 years	45

TABLE 5.5: FPG measures median and average.

Measure	**Value**
Average	41.53
Standard Deviation	24.5
Variation coefficient	0.59

calculated to evaluate the validity of the considered dataset. Table 5.5 includes the standard deviation, the average and the variation coefficient for FPG measures. The variation coefficient shows the extend of variability of data in a sample in relation to the mean of the population. The higher the coefficient of variation, the greater the level of dispersion around the mean, contrary, the lower the value of the coefficient of variation, the more precise the estimation. A variation coefficient of 0.59 indicates a 24 days of dispersion in 42 days of sampling rate, that is acceptable in this case.

Once the relevant data was extracted, the following step (see Figure 5.1) involved the application of the Process Mining and trace clustering techniques to discover the Dynamic Risk Model associated with hyperglycemia. Figure 5.2 describes the needed procedures to deal with this step. Overall, the exclusion criteria were defined and applied. In this case, patients with less than three observations for FPG during the period were excluded from the experiment, void FPG results were deleted, and the rest of measures from the laboratory dataset were discarded. Then, the Process Mining Log was created, following some filters were applied to adequate the Log for the construction of the appropriate model. After that, the Process Mining discovery was applied, and finally enhancement actions were carried out to enrich the model. Following paragraphs describe in detail the aforementioned procedures.

As shown in Figure 5.2, exclusion conditions were applied to the *Laboratory* dataset during the first stage of the experimentation procedures. After deleting patients with less than three observations during the period under study, entries with a void test result, and the other measures were also discharged. We obtained a dataset with 25,992 patients and 328,545 observations for the experimentation (from the initial 50,196 unique patients with 18,182,239 observations). The next action was the creation of the Process Mining Log. This process requires to build the log with the set

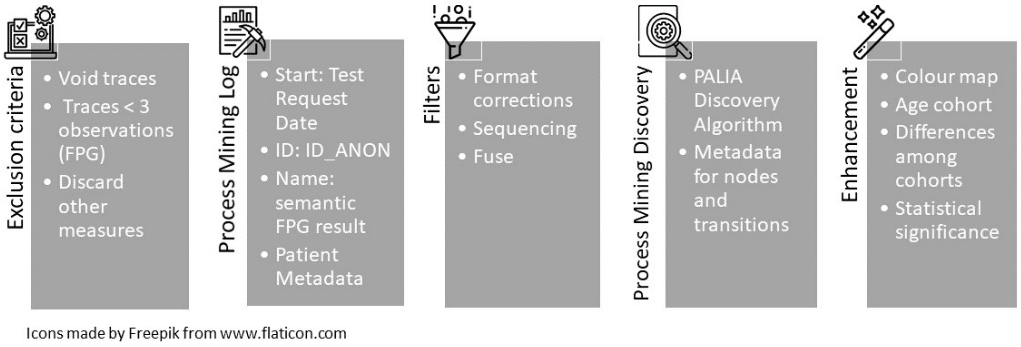

Exclusion criteria
• Void traces
• Traces < 3 observations (FPG)
• Discard other measures
Process Mining Log
• Start: Test Request Date
• ID: ID_ANON
• Name: semantic FPG result
• Patient Metadata
Filters
• Format corrections
• Sequencing
• Fuse
Process Mining Discovery
• PALIA Discovery Algorithm
• Metadata for nodes and transitions
Enhancement
• Colour map
• Age cohort
• Differences among cohorts
• Statistical significance
Icons made by Freepik from www.flaticon.com

FIGURE 5.2: Hyperglycemia Dynamic Risk Model experimentation particularisation.

of events containing the timestamp information about patients. For each event we considered the *Start* timestamp from the field *Test Request Date* (see Table 5.4), the identifier corresponding with the corresponded with the *ID_ANON* (see Table 5.3), and the name based on the semantic results as Named events, defined by the clinicians according to the mapping of the process, for the FPG results. The semantic values addition provides a semantic vision that facilitates the understanding of the chronic condition process semantically, this means to associate a semantic value to a numeric one. These values are disease depending, so in the case of hyperglycemia, the measurement of glucose in the blood remains the mainstay of testing for glucose tolerance status, this could be obtained by laboratory measures. We followed the current WHO diagnostic criteria for diabetes type 2 [102, 103]. The Diabetes semantic results were *Diabetes* for FPG $\geq 126\ \ \text{mg/dL}$; *Intermediate Hyperglycemia* for values of FPG between 100–125 mg/dl; and *Normal* for FPG less than 100 mg/dL. Finally we included some metadata associated with the patient characteristics, such as the *Age Group*, and *Overweight* and *Obesity* fields (see Table 5.3). The result of this process was the creation of the Process Mining Log.

Once the Process Mining Log was created, we proceeded with the application of the appropriate filters to adequate the model. Filtering procedure comprised format corrections, sequencing of traces assuming the ending of the current trace is the beginning of the next one, and equal traces fusing. The next step handled the application of the Process Mining Discovery algorithm. In our case, PALIA discovery algorithm was applied to obtain the process model for the Log. The result is shown in Figure 5.3.

In our approach, the metadata behind the model is also paramount to conceive valuable dynamic models, for example two patients could have the same FPG measure in a moment, *Intermediate Hyperglycemia*, but their timing and frequency could be completely different, therefore it was essential to analyse differences in the flow to comprehend the dynamic characteristics of the model. For this reason, within this step the Log was processed to compute the metadata associated with the model. PMApp supports metadata correlated to models in several ways, concretely in this work we used metadata computed to nodes and edges with statistical information, so we could appreciate how the executions of the models were performed. This

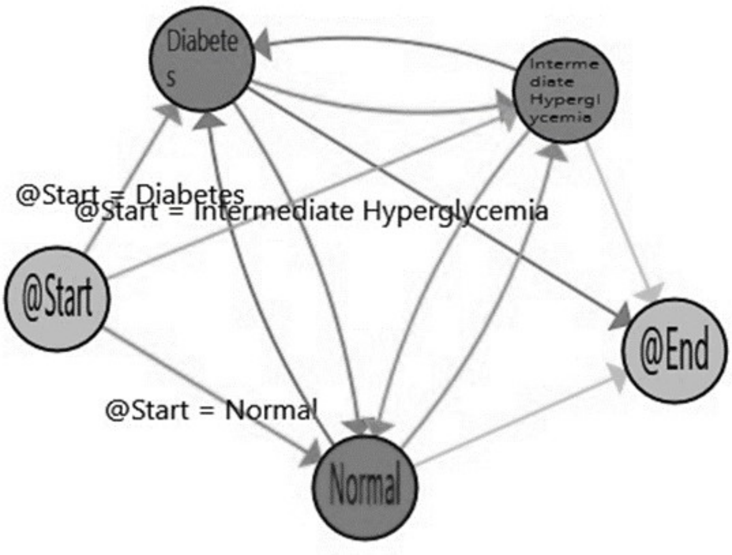


FIGURE 5.3: Dynamic fasting plasma glucose.

statistical information contains the execution number, the duration average, the duration median, the duration aggregation, the case number, and the duration by case. PMApp also manages storing the relationship between the topological structures of the model with the log events, as it is possible to navigate from the model to the individual.

Using the metadata, it was possible to enhance the model represented in Figure 5.3 and to improve its understandability and professionals' confidence on it. Firstly, we applied a colour map for nodes and transitions. Concretely, nodes were coloured by the average time spent in the stage, and edges were painted with a gradient symbolising the number of patients, that, proportionally follow this transition, both from green (minimum value) to red (maximum value), using the gradient scale represented in Figure 5.4 for both nodes and transitions.

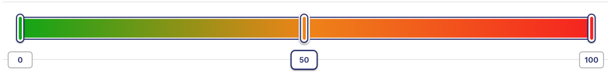


FIGURE 5.4: Gradient scale key for model representation from green to red.

After this enhancement action, the model resulted in an enriched and easier version to interpret, represented in Figure 5.5, the Dynamic Risk Model for FPG. It shows the considered population flow [29], where the *Normal* stage for FPG is the most prevalent on average. However, population spent a considerable time in *Intermediate Hyperglycemia* and *Diabetes* stages. It is not only important the time spent in each stage but also the transitions among stages, as this can suppose the difference between a well-controlled glucose status or not. The reddest transitions correspond with the higher number of patients, that proportionally followed a concrete path,

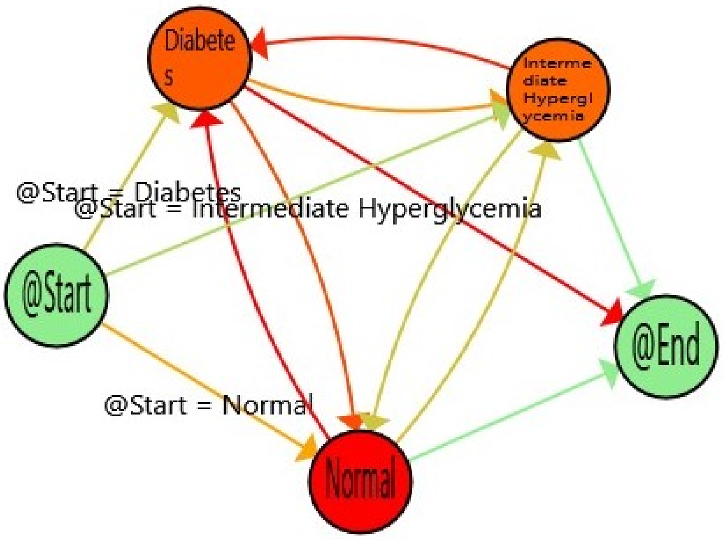

Diabetes
Intermediate Hyperglycemia
@Start = Diabetes
@Start = Intermediate Hyperglycemia
@Start
@End
@Start = Normal
Normal

FIGURE 5.5: Enhanced Dynamic fasting plasma glucose. Presented in [29].

and in this case Normal to Diabetes transition is the most followed one.

As explained in Section 1, some factors, such as age, gender, and ethnicity, are relevant for stating normal glucose level, therefore they could be consider for enhancing the model. Concretely the age group was added to the Log as metadata, accordingly it was used to obtain more relevant views for the health experts. To do so, we considered a three age cohort, *Young Adults* for 17–39, *Middle age Adults* for 40–59, and *Older Adults* for 60–100 [104]. The resulting augmented Dynamic Risk Model is included in Figures 5.6–5.8 for the three age cohorts, respectively.

With increasing age, we can observe how the time spent in *Intermediate Hyperglycemia* and *Diabetes* stages is higher, with the corresponding decrease in the time consume in the *Normal* stage. The flows also show how the most prevalent path is the one in which the population finalises in the *Diabetes* stage in the three groups. An in depth analysis can be done exploring the nature of the differences among the different cohorts using the enhancement possibilities of PMApp.

The model could also be enhanced highlighting the differences between two processes regarding the nodes and edges and the degree of the differences. We compared the older adults group with the other two age groups. The enhancement map compares two groups and colour the negative differences in red, where the saturation of the colour represents the difference's degree, from white for the positive to red for the negative difference, using the gradient scale represented in Figure 5.9.

These enhanced models are showed in Figures 5.10, and 5.11. Concretely, Figure 5.10 represents the differences between the older adults (≥ 60) group and the middle age group (40-59) flows, we could observe differences in the time spent in the *Diabetes* and *Intermediate Hyperglycemia* stages and the paths between them. In this population, age was an indicator of uncontrolled hyperglycemia and higher risk. This fact was supported by the next model in Figure 5.11 when differences were

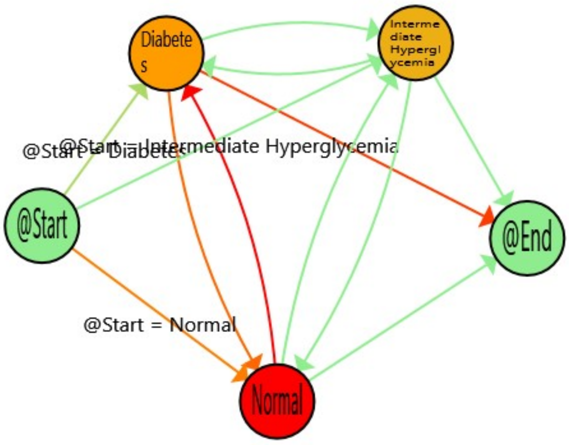

Diabetes
Intermediate Hyperglycemia
@Start = Diabetes
@Start = Intermediate Hyperglycemia
@Start
@End
@Start = Normal
Normal

FIGURE 5.6: Dynamic FPG for young adults. Presented in [29].

even deeper, represented by the reddest colour, when compare the older adult population (≥ 60) with the young adults one (17-39).

The comparison of groups is a very useful tool that could help health professional to discover and understand the nature of the differences among groups. In medicine, a classic trust measure to evaluate different medical processes is to show differences among them known as statistical significance. Basically, it helps to quantify whether a result is likely due to chance or to some factor of interest, thus it was a perfect candidate to be used for enhancing the discovered models. Most of the literature focuses on the p-Value for measuring the statistical significance [105]. PMApp implements the statistical significance using the p-Value for comparing nodes that refer to the same activity. For each execution associated with each activity is got the set of times and is applied the Kolmogorov-Smirnov Test to evaluate the normality of the distribution of the time values. If the two distributions reach the normality test, then it is used a T-student Test for the p-Value computation. If not, it is assumed the distributions are not normal and the Mann-Whitney-Wilcoxon Test is performed. If both situations, for a p-Value lower than a given threshold, it is concluded that the distributions are significantly different. Following the literature, the threshold was set to 0.05 [55]. This technique can be used to highlight the differences with statistical significance between two flows of the model. This approach can not only discover when a process is different but also in which parts of the models the differences lie.

The enhanced models in Figures 5.10, and 5.11 show that differences between age groups existed, however is paramount to know if these differences were statistical significant. Applying this technique, Figures 5.12 and 5.13 show where models for middle age and young populations differ with respect to the older adults one. Nodes highlighted in yellow mean that a statistically significant difference between the two cohorts exist. Concretely, Figure 5.12 represents how the older adult population significantly spent more time in the *Diabetes* stage, whereas they were less

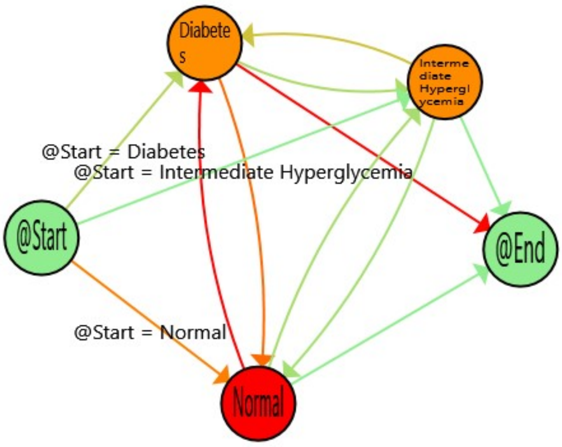

Diabetes
Intermediate Hyperglycemia
@Start = Diabetes
@Start = Intermediate Hyperglycemia
@Start
@End
@Start = Normal
Normal

FIGURE 5.7: Dynamic FPG for middle age adults. Presented in [29].

time in the *Normal* stage. Comparing the young population with respect to the older adult one, Figure 5.13 shows that these differences were even more considerable statistically speaking, as the young population substantially consumed less time in the *Diabetes* and *Intermediate Hyperglycemia* stages, and more time in the *Normal* state. With these enhanced models, we could confirm how age is affecting the FPG flow in the average time spent in each state of the model. These findings are supported by the literature, works such as [106], [107], [108] studied age-related changes in glucose and the distribution of FPG in adults, illustrating some people progressively lose the ability to regulate glucose levels as they did when they were younger, and concluding a positive correlation of age with worsening glucose tolerance.

The Dynamic FPG for the considered population confirmed main literature conclusions regarding age-related changes of FPG in adults, representing, in a comprehensive manner the different FPG flows in different age groups. This fact endorses the clinical validity of the model [29], [109]. Moreover, this approach is easy and relatively quick to apply, and could be adapted to any population. The model could incorporate other possible risk factors, such as BMI, blood pressure, or smoking, allowing the understanding of the risk factors and their interactions.

5.1.3. Dynamic Risk Model for Hypertension

Hypertension, also known as high or raised blood pressure (BP), is a condition in which the blood vessels have persistently raised pressure [110]. Based on WHO information, hypertension is a serious medical condition and can increase the risk of heart, brain, kidney, and other diseases. It is a major cause of premature death worldwide, and an estimated 1.13 billion people worldwide have hypertension [110]. Blood pressure is based in two numbers, systolic blood pressure (SBP) representing the pressure in blood vessels when the heart contracts or beats. And

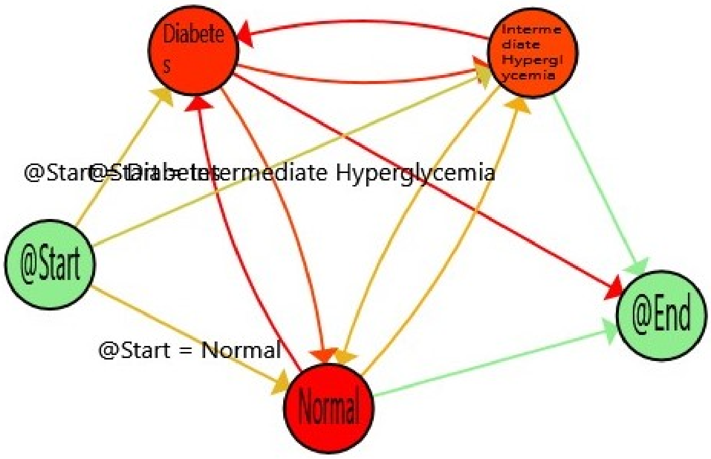

Diabetes
Intermediate Hyperglycemia
@Start@Start=Diabetes@Start=Intermediate Hyperglycemia
@Start
@End
@Start = Normal
Normal

FIGURE 5.8: Dynamic FPG for older adults. Presented in [29].

FIGURE 5.9: Gradient scale key for model representation from white to red.

the diastolic blood pressure (DBP) representing the pressure in the vessels when the heart rests between beats. Hypertension is diagnosed if, when it is measured on two different days, the SBP readings on both days is 140 mmHg or more, and/or the DBP readings on both days is 90 mmHg or more or taking anti-hypertensive medication [111].

Based on the framework for the experimentation included in Figure 5.1, the third and fourth steps deal with the selection of the appropriate clinical database and information based on the chronic condition under study and its associated risk factor. Similarly as in the previous model, relevant data for the creation of the hypertension dynamics were selected from the tertiary hospital database (see Table 5.2). Accordingly, patients' characteristics and the global identifier were extracted from the *Patients Anonymize* set see (Table 5.3), whereas SBP and DBP were obtained from the *Primary Care* dataset which description is included in Table 5.6.

Previous to continue with the experiment procedures, the standard deviation and the average of the measures included in the *Primary Care* dataset for the SBP and DBP were calculated to evaluate the quality of the sampling. Table 5.7 includes the standard deviation, the average and the variation coefficient for SBP and DBP measures, considering they are always taken together.

The fifth step of the framework considered the application of the Process Mining and trace clustering techniques, that again was particularised for the hypertension case as Figure 5.14 describes. The experiment includes the specific procedures to obtain the Dynamic Risk Model associated with hypertension. Comprehensively,

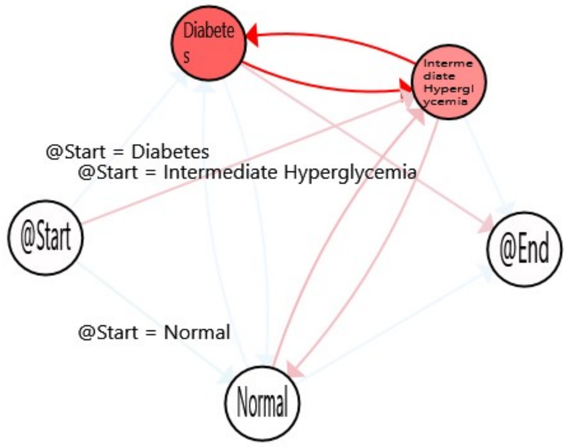

Diabetes
Intermediate Hyperglycemia
@Start = Diabetes
@Start = Intermediate Hyperglycemia
@Start
@End
@Start = Normal
Normal

FIGURE 5.10: Enhanced model: differences between Middle-Older Adults in FPG flow. Presented in [29].

TABLE 5.6: Primary Care dataset description for hypertension experiment.

Column Name	Data Type	Example
ID_ANON	Global unique identifier	000269d4-b40a-df4f-a1c0-56db3f989ad2
Measure Date	String	20170830
Code Measurement	String—type of observation	Weight, Height, SBP, DBP,. . .
Numerical Result	Float—measurement's result	87.5
Text Result	String—void numerical result	Yes/No
Age Group	Integer—grouped by 5 years	45

the first action deal with the concrete exclusion conditions of the experiment, then the Process Mining Log was built, following some filters were applied to adequate the Log to the considered model, so the Process Mining discovery algorithm could be applied, and lastly enhancement actions were performed to enrich the model.

Exclusion criteria were established for void SBP and/or DBP, patients with less than four observations during the period for SBP and DBP. Therefore, the rest of the information from Table 5.6 was discarded. From the initial 17,853 initial unique patients, we got a dataset from 3,575 subjects.

Following the experimentation process, the Process Mining Log creation with the set of event was needed. Each event was composed by a *Start* timestamps, a name, an identifier, and correlated metadata. The starting of the event corresponded with the *Measure Date* from the *Primary Care* dataset. We used named event, consequently the name of the event corresponded with the semantic result. Concretely, the semantic value was added for the SBP/DBP combination following the American Heart

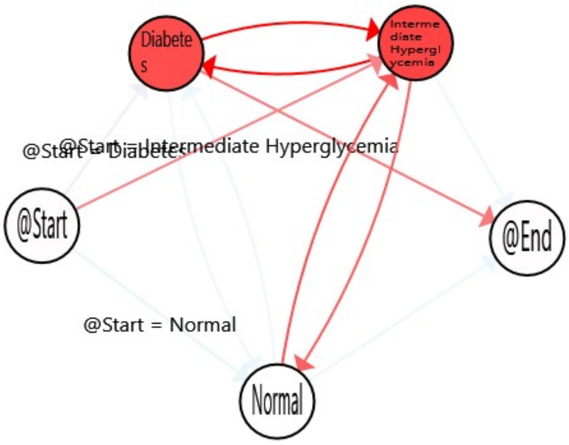

Diabetes
Intermediate Hyperglycemia
@Start = Diabetes
@Start = Intermediate Hyperglycemia
@Start
@End
@Start = Normal
Normal

FIGURE 5.11: Enhanced model: differences between Young-Older Adults in FPG flow. Presented in [29].

TABLE 5.7: SBP and DBP measures median and average.

Measure	Value
Average	42.03
Standard Deviation	24.37
Variation coefficient	0.58

Association (AHA) guidelines [112]. Semantic results for hypertension were *Normal* for SBP numerical result $<$ 120 mmHg and DBP numerical result $<$ 80 mmHg; *Elevated* for SBP between 120–129 mmHg and DBP $<$ 80 mmHg; *Hypertension stage 1* for SBP between 130–139 mmHg or DBP 80–89 mmHg; and *Hypertension stage 2* for SBP $\geq$ 140 mmHg or DBP $\geq$ 90 mmHg. The identification of the trace coincided with the *ID_ANON*, while the trace data, considered as the set of metadata related to the same case, included the *Age Group*, and *Overweight* and *Obesity* fields, all from the *Patients Anonimyze* dataset (Table 5.3).

Once the PM Log was generated, some filters were applied to shape the Log in an appropriate manner. It included some format corrections, concretely to the Measure Date, the Test Request Date, and the Numerical Results fields. After that, traces were sequenced assuming the end of the current traces is the beginning of the next one, and finally equal traces were fused. At this point, PALIA discovery algorithm was applied to obtain the process behind the Log with metadata computed to nodes and transitions, like in the previous model. Because of BP is a continuous variable that fluctuates in response to various physical and mental changes, we decided to combine trace clustering techniques with the discovery algorithm. It allowed to stratify

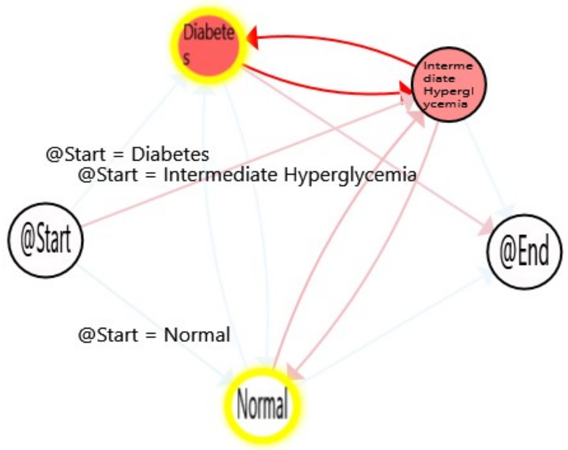

Diabetes
Intermediate Hyperglycemia
@Start = Diabetes
@Start = Intermediate Hyperglycemia
@Start
@End
@Start = Normal
Normal

FIGURE 5.12: Enhanced model: statistical significance between Adult-Older Adults FPG flow. Presented in [29].

the population with similar BP behaviour based on the semantic value of BP, using WTD and QTC. We performed several experiments with different values for the quality threshold and the similarity trying to obtain the most significant results adjusting between the number of groups and the behaviour represented within each group. Most valuable results were achieved for a quality threshold of 0.15 and 0.02 similarity. For the 3,575 subjects, 13 groups were obtained modelling the different BP flows, plus an outliers group for the 545 patients without a clear behaviour. The characteristics of these models regarding the groups population and the percentage of the total are listed in the Table 5.8.

Finally a colour map was used for enhancing the model to improve the comprehensibility. The colour map was applied to the average time spent in each node, and the number of patients proportionally following a transition, using the gradient scale represented in Figure 5.4. We grouped the different models in two patterns, differentiating between stable models and models showing blood pressure variability. The graphical representation for the stable patterns is included in Figure 5.15. These models comprise stable normal BP pattern in Model 1 (Figure 5.15a), stable hypertension stage 1 behaviour in Model 5 (Figure 5.15b), and stable hypertension stage 2 pattern in Model 8 (Figure 5.15c). In this three models there were no changes in the BP stage during all the study duration.

The rest of the discovered models present a blood pressure variability in the flows, varying in manner and duration among the different groups. We focused on the time spent in each stage to classify the different groups with respect the blood pressure stage where patient spent most of time during the considered period. Therefore, we grouped the various models into normal BP, elevated BP, hypertension stage 1, and hypertension stage 2, as predominant stages. Following this idea, in Figure 5.16 we included the population with a normal BP most of the period, but with

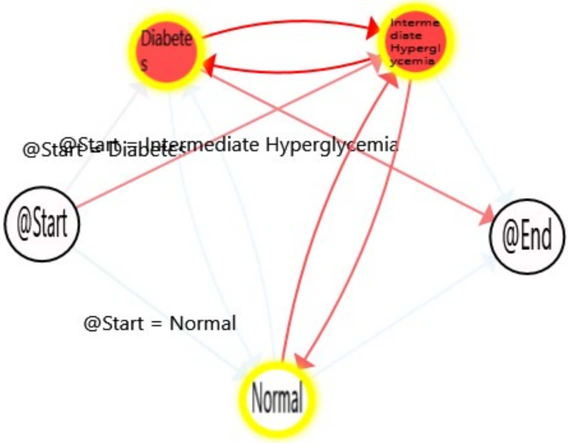
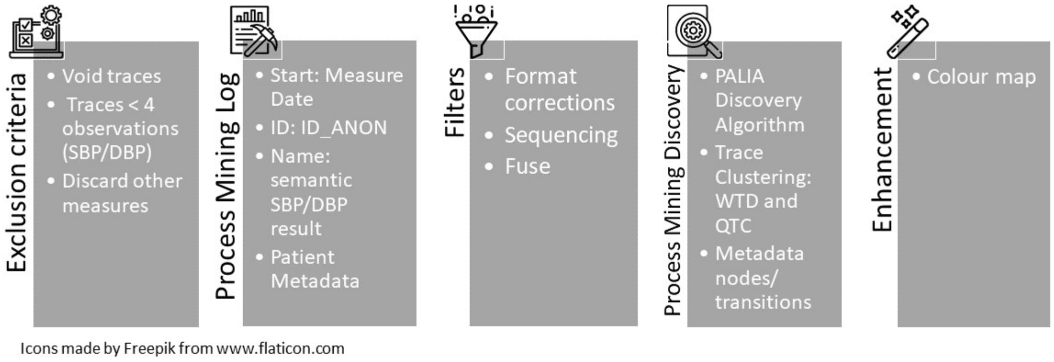

Diabetes
Intermediate Hyperglycemia
@Start = Diabetes
@Start = Intermediate Hyperglycemia
@Start
@End
@Start = Normal
Normal

FIGURE 5.13: Enhanced model: statistical significance between Young-Older Adults FPG flow. Presented in [29].

significant changes in their blood pressure. Concretely in Model 4 (Figure 5.16a), considerable changes among the different BP stages are observed. This variability is less significant in the other two models, in Model 7 changes mainly occurred between normal and elevated stages. Whereas Model 11 shows less variability but the time pent in normal stage and hypertension stage 2 is comparable.

Analysing the variability among models within elevated BP included in Figure 5.17, we can differentiate between Model 3 where the population mainly vary between elevated and normal BP stages (Figure 5.17a), and Model 6 where transitions mainly occurred between elevated and hypertension 2 stages (Figure 5.17b).

Looking into detail groups that comprised most of the time in hypertension stage 1, we found Model 2 represented in Figure 5.18a and Model 12 in Figure 5.18b.



Exclusion criteria
- Void traces
- Traces < 4 observations (SBP/DBP)
- Discard other measures

Process Mining Log
- Start: Measure Date
- ID: ID_ANON
- Name: semantic SBP/DBP result
- Patient Metadata

Filters
- Format corrections
- Sequencing
- Fuse

Process Mining Discovery
- PALIA Discovery Algorithm
- Trace Clustering: WTD and QTC
- Metadata nodes/ transitions

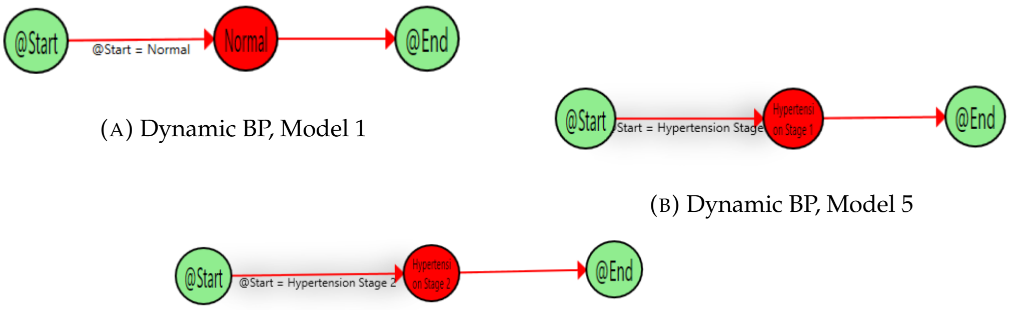
Enhancement
- Colour map

Icons made by Freepik from www.flaticon.com

FIGURE 5.14: Hypertension Dynamic Risk Model experimentation.

TABLE 5.8: Characteristics of the Dynamic Hypertension groups.

Behaviour	Group Name	Population	Total
Stable patterns	Model 1	335	9.4%
	Model 5	185	5.2%
	Model 8	118	3.3%
Mostly Normal	Model 4	275	7.7%
	Model 7	154	4.3%
	Model 11	94	2.6%
Mostly Elevated	Model 3	290	8.1%
	Model 6	159	4.4%
Mostly Hypertension stage 1	Model 2	310	8.7%
	Model 12	82	2.3%
Mostly Hypertension stage 2	Model 0	810	22.7%
	Model 9	110	3.1%
	Model 10	108	3.0%



(A) Dynamic BP, Model 1

(B) Dynamic BP, Model 5

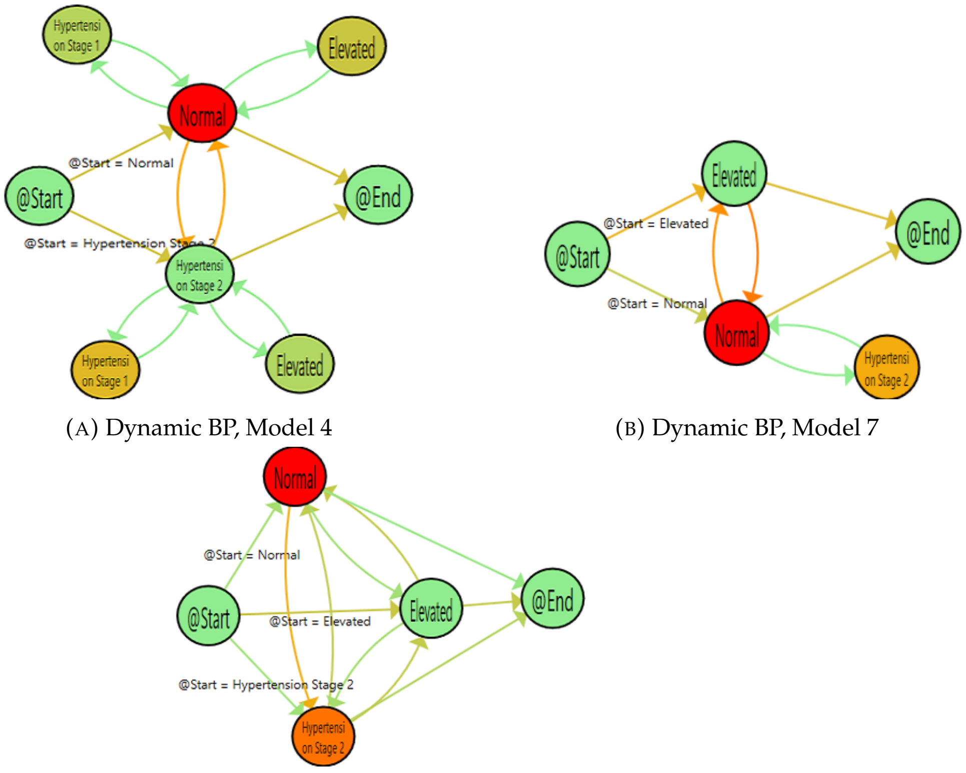
(C) Dynamic BP, Model 8

FIGURE 5.15: Dynamic Risk model of hypertension: stable patterns. Presented in [29].

Model 2 is characterised by a population that spent most of the time in hypertension stage 1 but comparable with the time spent in elevated state, with considerable fluctuations with the hypertension stage 2. Whereas in Model 12, population spent most of the time in hypertension stage 1 with fluctuations among the other states.

Lastly, we considered groups that spent most of time in hypertension 2 stage in Figure 5.19. These three models show a considerable variability, especially in Model 0 (Figure 5.19a) and Model 9 (5.19b) where BP state fluctuates among the different stages, whereas most of the population included in Model 10 (Figure 5.19c) changed between hypertension stages 1 and 2.

Blood pressure variability (BPV) has been studied and documented in the literature, concretely [113], [114], [115] stated how BPV increases with increasing BP level, regarding short-term variations, that occur throughout the day. On the other hand,

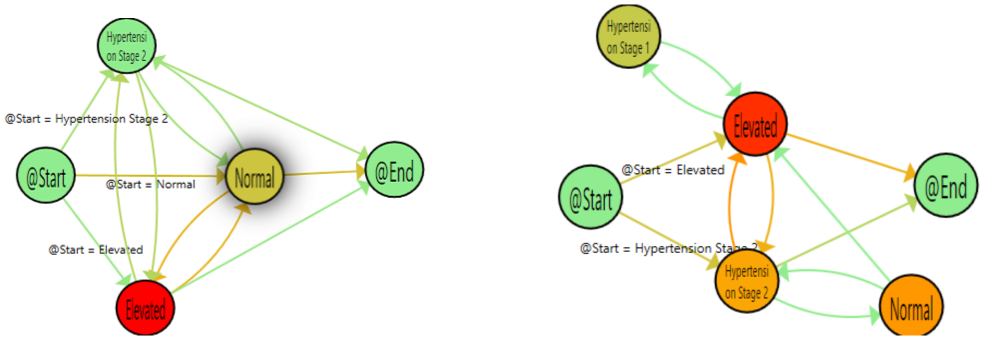


(A) Dynamic BP, Model 4

(B) Dynamic BP, Model 7

(C) Dynamic BP, model 11

FIGURE 5.16: Dynamic Risk model of hypertension: mostly normal. Presented in [29].

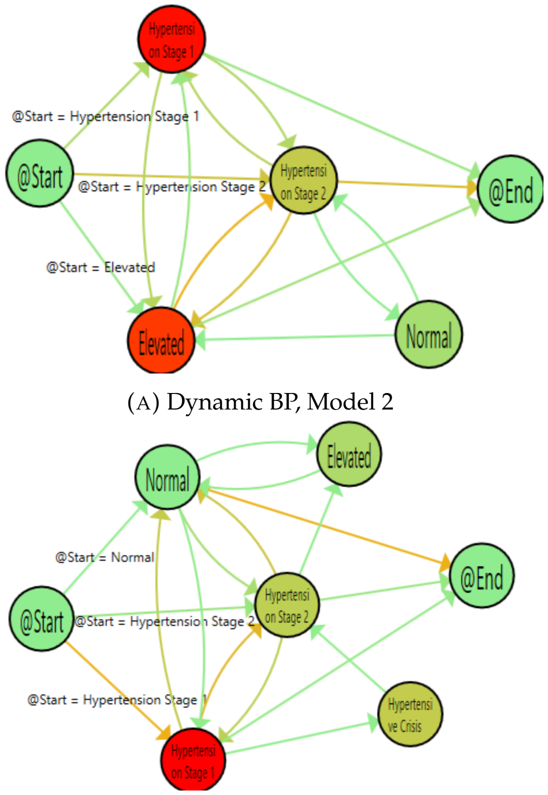


(A) Dynamic BP, Model 3

(B) Dynamic BP, Model 6

FIGURE 5.17: Dynamic characterisation of hypertension: mostly elevated. Presented in [29].

long-term variations meaning BP tendency to be different between days, months and seasons. Moreover, [114] stated the frequent inability of BP to remain controlled from one visit to another even during antihypertensive treatment. This fact is clearly observed in our Dynamic Risk Model, where 2,392 patients plus the 545 outliers



(A) Dynamic BP, Model 2

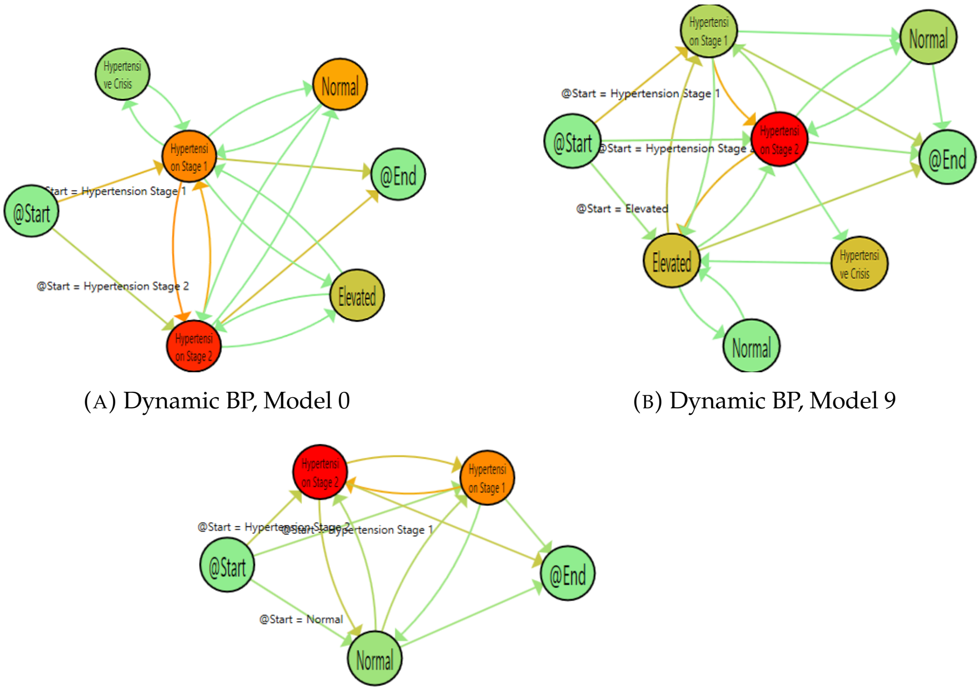
(B) Dynamic BP, Model 12

FIGURE 5.18: Dynamic characterisation of hypertension: mostly hypertension stage 1. Presented in [29].

show this variability, representing the 82.15% of the considered population. Moreover, some groups also support the idea of BPV increase with increasing BP level, such as Model 9 (Figure 5.19b) that shows a great variability for a population with a hypertension stage 2, or Model 0 (Figure 5.19a) with a variation among all the BP states. Besides, Model 2 (Figure 5.18a) represents this fluctuation for hypertension stage 1. Contrary, Model 7 (Figure 5.16b) and Model 11 show less variability for a population mostly in normotension. However, Model 4 (Figure 5.16a) represents a population mainly in normotension but with considerable variability, suggesting a greater effort should be placed to achieve consistent BP control between visits. Increased short-term and long-term BPV is associated with the development, progression, and severity of other diseases [116], our models support the idea that high blood pressure could also be the result of increased BPV, and not only of elevation of mean blood pressure values [29].

5.2. Towards a common framework

Using retrospective data from a tertiary hospital we have been able to create two Dynamic Risk Models for two of the most common and prevalent chronic diseases – Hyperglycemia and Hypertension. These two Dynamic Risk Models have presented



(A) Dynamic BP, Model 0

(B) Dynamic BP, Model 9

(C) Dynamic BP, Model 10

FIGURE 5.19: Dynamic characterisation of hypertension: mostly hypertension stage 2. Presented in [29].

novel information to health experts from data collected and stored in daily clinical practise with insight and comprehensible views about the processes behind the diseases. With this approach BP and FPG have been treated as continuous variables instead of based on static cut-off points. Furthermore, we were even able to combine the results with other variables such as the age to generate different cohort and discover differences with statistical significance among them.

In the course of the two performed experiments we have followed a set of steps dealing with some common aspects such as the extraction of the relevant variables from the dataset, the creation of the events and the trace data to generate the Process Mining Log, and the application of a set of filters with different purposes. Similarities have been established within the two generated Dynamic Risk Models, therefore the following step was to establish a framework to apply Process Mining techniques in the context of discovering Dynamic Risk Models for chronic diseases.

Chapter 6

Towards a formal methodology for obtaining Dynamic Risk Models

Heretofore, it has been conducted two case studies that involved the application of Process Mining and trace clustering techniques for the production of valuable and understandable representations of the real flow of a chronic diseases considering both the dynamic and the behavioural perspectives. Concretely, we obtained two Dynamic Risk Models for hypertension and hyperglycemia, following concrete procedures through the experiments and trying to answer specific questions. The idea now is to consolidate and establish this approach. In this context, a methodology that allows appropriately presenting data, and involving health professionals in the loop is needed. With this purpose, the focus should be not only on the steps to perform for obtaining the model and the appropriate tools and procedures that allow the active involvement of the health experts, but also on the definition of the characteristics the methodology should guarantee. Thus, this chapter is related to the objectives O2 and O3 of this work.

6.1. Towards a new approach for analysing risk models

Chronic conditions represent one of the principal challenges for health systems today, their appearance is usually gradual, their development progressive and their treatment complex [117]. This has led to increased efforts to improve the care provided and the knowledge from a multidisciplinary approach. Within this context, it comes a series of questions about what is the best manner to tackle challenges, moreover, using the information related to the patients' health episodes stored in the health information systems, it is possible to support questions posed by experts in the field. Some examples of queries within the chronic conditions field posed by healthcare experts could include: What are the most commonly followed processes in a concrete chronic disease? Are there any differences between care processes followed by different patients' groups? [118]. Can we create digital twins models based on the patient characteristics to anticipate patient outcomes and disease progression or identifying personalised therapies? [119]

Our approach in the past sections of this work (see Chapters 4 and 5)) were centred in the use of Process Mining as the main tool for obtaining Dynamic Risk Models processes abstractions for chronic diseases from data included it the health information system from a hospital. However, these risk models should support questions posed by the health experts in the field of chronic conditions to help them to obtain further knowledge about them.

Consequently, the objective of this chapter is to propose a methodology based on Process Mining that allows answering questions posed by health experts in the filed of chronic conditions, in the form of Dynamic Risk Models as they were defined in Chapter 4, as the graphical representations of a disease process considering its evolution from a dynamic perspective and the individual characteristics of the patients, presenting the risks based on the stratification groups that permit a better understanding of the clinical cases. In this sense, the healthcare domain has its own particulates and implies a theoretical knowledge about the process under study, needed to posed the interesting and needed questions, for that, the health experts involvement is paramount. Moreover, we have to assume not all factors would be answered or represented by a Dynamic Risk Model as we have described it, and consequently, the methodology could not be able to answer every posed questions.

Based on this, the methodology should allow two main fundamentals, on the one hand, to describe and incorporate the inquires according to the users' needs, considering they are mainly the clinical experts in the chronic diseases field. And on the other hand, that the obtained results are understandable for them. Moreover, to establish a methodology that allows obtaining Dynamic Risk Models associated with chronic conditions, it is also needed to consider two main issues. First, the nature of the diseases over the Dynamic Risk Model would be formulated, and secondly the characteristics of the risk to be dynamically modelled, since it is worth to mention that not all factors would be answered or represented by a Dynamic Risk Model.

Based on these three needs, to incorporate clinical experts' needs, to provide understandable results, and to consider the nature of the chronic disease itself during the formulation, we can establish the preferable basis of the methodology upon two pillars: *interactive inputs*, and *understandable outputs*, represented together with how they are combined to reach the concept of *Interactive KPIs*, in Figure 6.1.

According to the established needs, the methodology departs from the idea of supporting health experts in the formulation of questions with clinical relevance that allow them to gain understanding and knowledge about the underlying processes. For this to end, it should work with **interactive inputs**, so the questions are posed by the health professionals in the field, allowing them to inquire what and how they want to know about a disease, instead of presenting them the results that experts in the Process Mining field consider important without the clinical experts involvement. Proposing questions beyond the clinical experts might result in Dynamic Risk Models with no clinical meaning, or analysing the incorrect variables, and in consequence, with no utility. In this scenario, both medical and Process Mining experts need to work closely. Moreover, it is essential to consolidate health professional knowledge, not only at the questions definition stage but also within the rest of the process. Machine Learning systems are not error-free, so human intervention is needed to verify and correct the results [120]. As explained in section 3.1, the Interactive paradigm defines this concept by integrating the human activity into the process [48]. It assures a close collaboration between an automatic learning algorithm and the human in order to, not only provide models that professionals can use for a better understanding of the actual process but also correct those models according to human knowledge so new perceptual and cognitive models are provided.

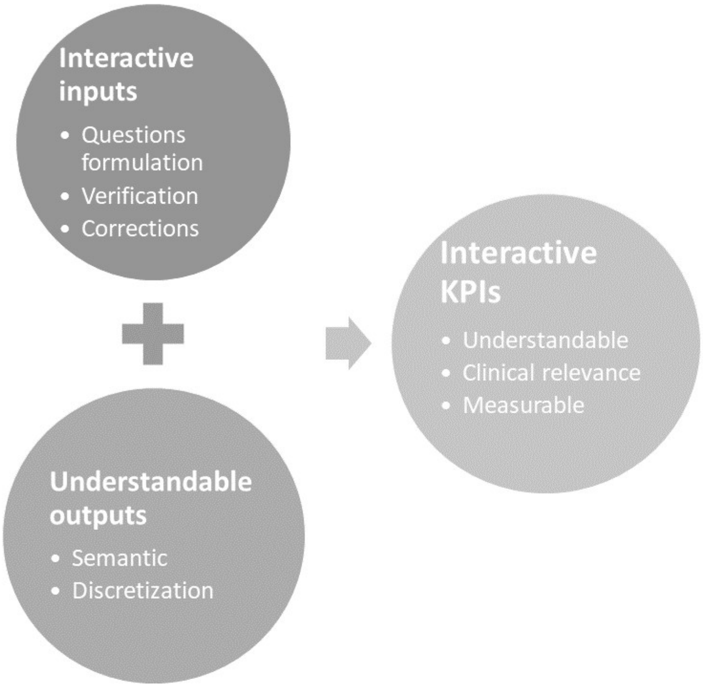


FIGURE 6.1: Formal basis of the methodology.

Thus, the Interactive paradigm perfectly matches with the idea of the definition of the questions and the interactive inputs, as long as the process of formalising questions should consider incorporating human experts in their definition but also in the verification and refinement of the obtained results, in an interactive way. Besides, the results provided to the posed questions should be human-understandable, measurable, and with clinical meaning and value. Additionally, the definition of a work methodology to follow with the final objective of transforming formulated questions into models would ease the process and guide clinical experts in their definition and interpretation. Moreover, when we contemplate that not all queries in any clinical domains would be regarded as candidates for its translation into Dynamic Risk Models.

However, the clinical experts' involvement is not only needed in the overall process of the question formulation, verification, and corrections, but also as consumers of the results. The Dynamic Risk Models discovered thought the methodology should be understood by the health professionals in other to gain awareness of the disease process and to generate new knowledge, is therefore paramount to produce **understandable outputs** to achieve this. Recapitulating the work presented in Chapter 4 and Chapter 5, both documented how paramount is the health experts' involvement in the experimentation not as mere spectators but as a fundamental tool following the Interactive paradigm and integrating human activity into a process [48]. As explained in Chapter 5, semantic values associated to numeric ones add a semantic vision that facilitates the understanding of the chronic condition process by the health professionals in a natural manner. Moreover, these semantic values should be added following experts in the chronic conditions field knowledge. For example, it is much simpler to understand that a person is overweight than a person who has a BMI of 26.5. However, it is not only a matter of understandability; it is also a discretisation issue in which a continuous variable is transformed into a discrete one. This process should also be performed and supervised by clinical experts. In this manner, they can decide the most appropriate interval or cut-off points applying different definitions or concepts based on their knowledge and the available literature. Following the same example, a BMI of 26.5 could be assumed as overweight for an adult population [97] but as a normal weight for an older adult population [98].

Integrating both basis, the methodology presents the results as **Key Performance Indicators** (KPIs). The concept of KPIs comes from the business sector, and is commonly used for providing a measurable value that demonstrates how effectively a company is achieving key objectives, in other words a measure of the level of performance of a process. It could be seen as a method for formalising complex questions. KPIs are usually defined to be SMART (Specific, Measurable, Achievable, Relevant and Timely). It means, the measure has a *Specific* purpose for the business, it is *Measurable* to really get a value of the KPI, the defined norms have to be *Achievable*, the improvement of a KPI has to be *Relevant* to the success of the organisation, and finally it must be *Time* phased, which means the value or outcomes are shown for a predefined and relevant period. KPIs are used for the analysis and evaluation of processes, and they are usually addressed using numerical indicators that represent the status of the process in a moment in time, and are also compared with previously defined values that represent the expected value in the execution of the processes. This concept could support the formalisation of health professionals' questions in a formal, flexible and open way. Thus, it is desirable to incorporate this paradigm to the process of obtaining Dynamic Risk Models for chronic diseases in a measurable manner. Going a step forward, [121] proposes the term **Interactive Process Indicators** (IPIs), combining the previous idea with the interactive framework to create human-readable and contextualised KPIs. IPIs support the expert in the navigation behind the model discovering its features and specificity, letting them to understand, measure and optimise the underlying processes. The main idea behind is how Process Mining technologies can be used for creating indicators with clinical meaning for and with the healthcare professionals. The capabilities of Interactive Process Mining technologies can support not only the characterisation of general process-based KPIs, which show how the process is executed, but also the analysis of individual and personalised aspects of the processes as they include information form the individuals. Therefore, IPIs are not numbers such in the case of the KPIs but advanced views in the form of enhanced processes that provide a human-understandable view that supports the expert in the better perception of the processes for an advanced assessment, in our case Dynamic Risk Models for a chronic disease.

6.1.1. Question based methodology

Then, the methodology should be framed in a formal question based schema that covers the pillars established in the previous section, being interactive and allowing the incorporation of human knowledge. Mainly, there are three standard Process

Mining methodologies available in the literature[122], Process Diagnostics Method (PDM) [123], L^* life-cycle[124] , and PM^2 [10]. All of them propose a set of stages for providing best practices in the application of Process Mining technologies to an existing problem. However, for our problem, the only methodology that encourage an iterative vision is PM^2 [10]. PM^2 is a general purpose methodology based on six stages to translate concrete research questions into findings. Usually Process Mining applications are adaptions on PM^2 incorporating the basics needs in each case.

In the literature, in the case of healthcare, there are some techniques attempting to provide answers to questions posed by health experts, concretely [125] proposed a question-driven methodology specific for emergency rooms using Process Mining. This was the first interactive Process Mining methodology presented for medical domain based on PM^2. The methodology provides the necessary tools to guide the search for answers to frequently-posed questions by emergency rooms experts. The method follows and adapts the guidelines for Process Mining projects proposed in [52] to be question-driven. Specifically, it involves domain experts in the procedures, a crucial need identified in the previous section (see Figure 6.1) and proposes six clear steps of how to apply Process Mining to analyse emergency rooms processes ([125]):

- **Stage 1: Data extraction**. This first step of the methodology proposes to identify and extract the needed data from the Hospital Information System (HIS), build a data model for the underlying problem, name the events or activities, create any specific field and verify the quality of the extracted data.

- **Stage 2: Event log creation**. The second phase of the methodology aims to create the event log for the question posed by the emergency rooms expert, building an event log to be used in the following stages.

- **Stage 3: Filtering**. This stage enables the event log to be refined in line with detailed characteristics in accordance with the analysis sought.

- **Stage 4: Data analysis**. This step includes the analysis of data about how the process has been performed, as stored in the different event logs. This analysis involves the data analysis techniques and the corresponding tools and the application of statistical analysis and data mining.

- **Stage 5: Process Mining**. This phase deals with all the necessary steps for the application of Process Mining techniques and algorithms, including selecting the appropriate tool and identifying and applying the adequate methods.

- **Stage 6: Results evaluation**. During the last stage, results are presented to emergency rooms experts, to gather information about the answers provided to their posed questions, and about the clinical impact of the data and models obtained.

Within this methodology experts in the field play a crucial role from the beginning and through the overall stages. They are key input during the data extraction and collection process, in terms of the questions' definition, the establishment of the values for filtering the data, during the analysis stages, and of course during the verification and evaluation of the results, helping them to acquire additional knowledge about the process. This is directly related with the basis needed and presented for our methodology in Figure 6.1. Therefore, this methodology will be the formal framework over which our procedure for deploying Dynamic Risk Models for chronic conditions will be formulated in the following section.

6.2. Interactive and question based methodology for deploying Dynamic Risk Models

Using the methodology explained in the previous section as a formal framework, we have adapted and particularised it in the field of chronic diseases and Dynamic Risk Models. Thus, this section proposes the *interactive and question based methodology for deploying Dynamic Risk Models for chronic conditions using Process Mining*. We introduce a six-step methodology designed to support medical and Process Mining experts in the understanding of chronic disease's processes, bringing them a straightforward guideline of how to apply interactive Process Mining to the analysis of chronic disease's underlying processes. In this scenario, the medical expert contributes with her/his deeper understanding of the chronic disease formulating questions, but also within the verification and correction of the results during the different stages of the methodology. On the other hand, the Process Mining expert adds her/his awareness on how to apply Process Mining techniques for obtaining the corresponding Dynamic Risk Models in the form of IPIs. The six steps are as follows: (1) defining the corresponding question; (2) analysing the risk factors and variables that could be used for answering the posed question; (3) verifying data quality and availability; (4) formalisation of the IPI associated with the posed question; (5) applying stratification analysis; and (6) validating the results with the experts in the field. All the stages are presented in Figure 6.2 and explained below.

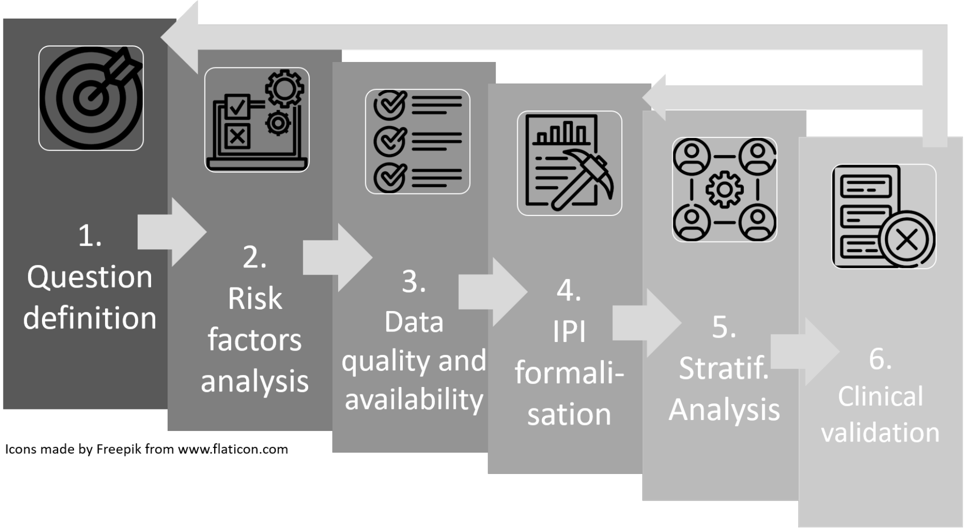


FIGURE 6.2: Interactive and question-based methodology for deploying Dynamic Risk Models.

- **Questions definition**

  The first step of the proposed methodology deals with the formulation of the questions regarding the chronic condition the healthcare experts in the field want to analyse in a dynamic manner. Posed questions should be balanced between straightforward questions that could be directly answered with a classic analysis of risk models, and too complex or impossible to establish a correlation between the available data and the question to be solved. The posed questions will be the point of departure for the rest of the methodology.

- **Risk factors analysis**

  Once the question or questions to be answered have been identified, the second stage includes the analysis of what data will be used. As explained in Chapter 2, chronic diseases share common risk factors that can be modified and are common for women and men. They are the intermediate or metabolic risk factors of blood pressure, raised glucose levels, abnormal blood lipids, overweight, and obesity. Therefore, the associated risk factors' analysis is crucial to answering the questions. This step also considers other variables or measures needed for answering the questions, such social-demographic data, vital signs, diagnostics, or medication among others. During this stage the close collaboration between the clinical and the Process Mining experts is crucial, identifying the needed variables, and consulting the literature.

- **Data availability and quality**

  This stage considers the posed question together with the specific risk factors and data from the two previous phases and builds the event log for the following stages, for this, some main tasks need to be considered.

  - Data availability - During this activity it should be identified the needed dataset and the required permissions to access them. A unique identifier across the different dataset is needed for mapping the different patient episodes and information. It is also important to assure the availability of a timestamp for each event, so we have the moment when the event take place. This task also analyses if all needed data could be extracted, and if there is missing data in the data sources, analyse if even so, it is feasible to perform the analysis. At this point, if important data is missing from the available dataset it could be necessary to go back to the first or second step and redefine to question or re-analyse the risk factors and needed variables. For example, sometimes a variable could be inferred from another, such it is the case of BMI, if we do not have direct access to it, we can calculate it from the pair weight and height, if available.
  - Build the model - The model represents all the data required to answer the posed question. All necessary data might be extracted and stored for the following stages.
  - Data quality - As described in [125], there are some event quality issues, such as incorrect time stamps, incorrect format of data, missing events or activities, repeated events, and others. They should be considered prior to the event log creation. The involvement of clinical experts is needed at

this point to decide if incorrect or inaccurate values might be corrected or discharged.

- Event Log - A specific event log needs to be created for each question to be answered, therefore, each event log is handled by the question that requires a response. Process Mining techniques assume the existence of an event log as an input, and it represents the actual execution of a process. As explained in Chapter 3, an event log is composed by traces, and each trace is represented by the ordered sequence of events that have occurred during the execution of that particular episode. Therefore, an event log can be seen as a collection of traces and a trace can be seen as a sequence of events. At this point, it may be also interesting to propose and create specific fields for the analysis. For example, we can add a new field for *polypharmacy*, using the medication information included in the database and the healthcare expert knowledge. In the event log creation task is decisive the collaboration between the clinical and Process Mining experts, as it is needed an in depth understanding of the process behind the chronic condition to include all the relevant information to answer the posed question.

- **IPI formalisation**

  This stage applies the experimentation based on Process Mining techniques to obtain the disease process from a dynamic perspective. Resulting then in the formalisation of the IPI for the posed question. In accordance with the standard flow proposed and described by [126] for the application of Interactive Process Mining, the main tasks involved in this stage are described below [127].

  - Log filtering and processing - The filtering task deals with the creation of a specific event log based on the characteristics of the posed question to be answered. Usually, filters will include, exclude, or group events from the event log among others. This activity will produce a refined Process Mining log in line with what the experts need regarding the question. These filters can be developed using algorithms that process the log and select only the patients (traces) that have the specified characteristics. For example, filters by date or time, or by gender could be applied among many others. Furthermore, not only filters can be implemented in this phase, but also, processing algorithms that can correct or simplify the log according to the information that the clinical expert provides. For example, algorithms that fuse sequential events that are equivalent, that group or rename specific events, that consider the finishing of one event as the start of the next one, or any other specific algorithm.
  - Process Mining Discovery - Once the log is prepared, the process model is inferred by using a Process Mining Discovery algorithm.

  - Model Processing - After the discovery, this step is comprised of the computation of the metadata associated with the model, this is data that provides information about the own data. This metadata is used to characterised the event log, identifying the frequency of activities, the distribution of cases over time and variants of process execution, among others. The analysis of metadata can support health experts in the understanding of the dynamic characteristics of the processes. For example two processes could have the same events, but their timing and frequency are different, and the analysis of these differences is important to assess experts in the divergences or similarities among processes.
  - Enhancement and graphical representation - Until this point the methodology was focused on accessing, collecting, and processing data, but it remains one of the main objectives in an interactive methodology, which is to present information to the clinical experts. This stage supports this goal. It provides an augmented model view highlighting specific situations using the available metadata information. The selection of the highlighting tools depends on the characteristics of the process to be presented and the question to be answered. This step improves the final representation and supports the usability, utility and reliability on the technology.
  - IPI - Finally, the result is the Process Indicator of the Dynamic Risk Model that will be presented to the clinical expert for its validation in the last step of the methodology.

- **Stratification analysis**

  The fifth step obtains the disease process considering the individual patients' characteristics as described in the Dynamic Risk Models definition, grouping patients based on their different behaviours. This stage would be applied if it is relevant for the understanding of the process and the diseases, as not all posed questions will derived in the analysis of the patients' characteristics by clusters. As explained in Section 4.1.1, trace clustering techniques are unsupervised Data Mining solutions that are able to group traces that have similar behaviour, maximising differences with the rest of groups to determine different patterns. Thus, trace clustering techniques are used in this phase as stratification filters to extract sub-logs from the main log, representing subpopulations depending on a specific characteristic. Applying trace clustering techniques the methodology can present the risks based on the stratification groups, allowing a better understanding of the clinical cases. Clinical experts might explore new ideas to stratify patients trying to discover different patients' behaviours. This stage culminates with the *Dynamic Risk Model*.

- **Clinical validation**

  The results should be validated with the existing literature and the clinical experts in order to know whether they provide the information to answer the questions posed by the experts in the chronic conditions field at the beginning of the methodology. This step might result in new questions or in the refinement of the initial ones.

In summary, the proposed methodology implements six steps covering all the needed actions to bring Dynamic Risk Models to health professional, by supporting and accompanying them in the formulation of questions related to chronic conditions that might benefit from a dynamic perspective.

6.3. Putting into practise the proposed methodology

This section provides an example of the application of the proposed methodology for answering two specific questions regarding one concrete chronic condition, obesity.

Use Case

The use case relates to obesity and overweight chronic disease as a sensitive problem for health public authorities in Valencia city (Spain). The data collected correspond to 2017 from a tertiary hospital in Valencia, and the aim of the case study was to demonstrate the utility and relevance of the proposed methodology to provide answers by the corresponding IPIs thanks to the proposed methodology defined in Figure 6.2. Following is describe the tasks undertaken during the different stages of the methodology to obtain the corresponding IPIs.

Questions definition

Obesity is itself a chronic disease, the prevalence of which is reaching pandemic proportions worldwide. Obesity was recognised as a disease by the American Medical Association (AMA) in 2013 [128]. Accordingly, WHO defines obesity and overweight as *abnormal or excessive fat accumulation that presents a risk to health* [97]. This excess fat is associated with a clear increase in health risk, moreover, being overweight and obese are risk factors for numerous chronic diseases, including type II diabetes, cardiovascular diseases, respiratory diseases, infertility, some types of cancer, and psychological disorders and social problems that have a very negative impact on the quality of life of people with obesity and overweight [129]. In addition, overweight and obesity are among the five greatest risks of mortality in the world [130].

On the other hand, obesity has a clear dynamic component. However, classic approach to characterise obesity and overweight are based on static data, such as BMI, body fat percentage (BF%), or waist circumference among others. This is, measurements made by a healthcare professional at a certain moment, and which are used to determine whether the patient is overweight or obese at that time. This fact presents two main problems, firstly, concerns related to the data itself, such as their availability and quality. To obtain a diagnosis of obesity based on the measures described, it is necessary to have the data in the clinical history of the patients, in order to be able to carry out a systematic detection of the disease, which does not always happen. And even when the data are available, their quality prevents them from being used for a systematic diagnosis at the population level. And, secondly, these measures do not take into account the variability over time, neither of the measure in question nor of the health status of the person, nor it is related in the diagnosis of the disease with the presence of comorbidities and pathology associated with obesity, its consequences and evolution. This was a compelling context in which to formulate new questions for applying the proposed methodology and obtaining the underlying and real process associated with obesity that could support new insights and knowledge of the disease. In this context and in collaboration with the clinical experts of the hospital there were formulated two questions: *Could obesity be re-defined as a dynamic process?*, *How dynamic obesity is related with its comorbidities over time?* [131].

Risk factor analysis

BMI is a simple index of weight-for-height that is commonly used to classify overweight and obesity in adults. It is calculated by dividing a person's weight in kilograms by the square of his/her height in meters (kg/m^2). WHO also establishes a normal BMI range as 18.5 to 24.9, while a BMI greater than or equal to 25 kg/m^2 and below 30 kg/m^2 is considered to be overweight, and similarly, a BMI greater than or equal to 30 kg/m^2 is classified as obese [97]. Regarding its comorbidities, we chose hypertension as the close relationship between excess adipose mass and hypertension is well documented [132]. As it was presented in section 5.1, hypertension is diagnosed if, when it is measured on two different days, the SBP readings on both days is 140 mmHg or more, and/or the DBP readings on both days is 90 mmHg or more or taking anti-hypertensive medication [111]. Therefore, BMI and its related measure weight and height, and its relation with hypertension, and its measures SBP and DBP, were chosen to characterised obesity and one of its comorbidities respectively, following posed questions.

Data availability and quality

Working with the two posed questions together with the specific risk factors and associated measures, the main objective of this step was to build the event log. For this, the different tasks were performed. First, data were extracted from the HIS of the hospital. The ethical approval to the confidential patient data was obtained through the ethical process reached by the framework of the project in which the present research was done. As all data were anonymised prior to the extraction by the hospital IT department, and as it was used retrospective data, the information consent was not needed in this case. Extracted data enclosed information from the primary care service and patients characteristics, as described in Table 6.1. The hospital experts provided the data in several Comma-Separated Values (CSV) files, concretely the set of rows and columns as described in Tables 6.2 and 6.3 for both dataset. Data from these two dataset built our model for the two posed questions.

The following step was to check the quality of the available data before the event log creation. Data quality tasks performed included date and value formats, as follows: establishing the desired date format (dd/mm/yyyy) for the *Measure Date* field (see Table 6.2), and correcting the format for the *Numerical Result* field (see Table 6.2) converting it from integer to decimal and changing the , by ., for example correcting 87, 500 to 87.5. Each report includes detailed information about each activity or event considered in the analysis, as follows: a unique episode ID, the activity name, a timestamp, and, optionally, a series of attributes about the activity or event.

TABLE 6.1: Hospital Information System description.

Table	Description	Unique Patients/ Observations	Period
Patients Anonymize	Patients general information: age, identifier, some diagnoses	50,196	-
Primary Care	Primary consultations' data: variables and annotations	17,853/215,523	2017

[a] International Statistical Classification and Related Health Problems, [b] Diagnosis-Related Group.

TABLE 6.2: Primary Care dataset description for obesity experimentation.

Column Name	Data Type	Example
ID_ANON	Global unique identifier	000269d4-b40a-df4f-a1c0-56db3f989ad2
Measure Date	String	20170830
Code Measurement	String—type of observation	Weight, Counselling, SBP, DBP,. . .
Numerical Result	Float—measurement's result	87,500
Text Result	String—void numerical result	Yes/No
Age Group	Integer—grouped by 5 years	45

TABLE 6.3: Patients characteristics dataset description for obesity experimentation.

Column Name	Data Type	Example
ID_ANON	Global unique identifier	000269d4-b40a-df4f-a1c0-56db3f989ad2
Age Group	Integer—group of age by 5 years	40
Overweight	Integer: 1/0, overweight diagnose	0
Obesity	Integer: 1/0, obesity diagnose	1
Unspecified Overweight/Obesity	Integer: 1/0	1

TABLE 6.4: BMI and SBP/DBP measures standard deviation and average.

	BMI	SBP/DBP
Measure	**Value**	**Value**
Average	45.96	42.03
Standard Deviation	22.36	24.37
Variation coefficient	0.49	0.58

These attributes include demographic information about patients that help to understand the process. Therefore, the event log was created considering all the relevant episodes for 2017: the unique identifier for the patient (*ID_ANON*), the activity name to the *Code Measurement*, the corresponding timestamp was composed by the *Measure Data* of the measure, and we also included the *Age Group* as demographic information.

Before continue with the subsequent steps of the methodology, we calculated the median and the average of the measures included in the *Primary Care* dataset for the BMI and SBP/DBP to evaluate the validity of the considered dataset. Table 6.4 includes the median, the average and the variation coefficient for BMI and SBP/DBP measures. As explained in the previous chapter (see Chapter 5), these values indicate that the sampling is acceptable so we can use it in the experiment.

IPI formalisation

The goal of this stage was to implement the experimentation based on Process Mining techniques according to the characteristics of the posed questions, the redefinition of obesity as a dynamic process, and its dynamic relationship with hypertension. Following the methodology the first task should involve the log filtering and processing to adapt the log to the characteristics of these questions. For the first posed question we needed the weight and height measures, as the objective was to obtain BMI behaviours. Accordingly, the first filter used aimed at selecting the episodes for weight and height measures. Regarding the second posed question particularise to the dynamic relationship of obesity with hypertension as comorbidities, we selected episodes with SBP and DBP. Then, we excluded the rest of episodes with other *Code Measurement* from the event log. As we had two posed questions, from this point we performed the methodology in separate flows for formalising the two different IPIs, as explain in the two following subsections, *IPI1 - Dynamic Obesity Risk Model* and *IPI2 - Dynamic relationship between obesity and hypertension*.

IPI1: Dynamic Obesity Risk Model

For the concrete case of the first posed question, we assigned the semantic result of the BMI as the activity name –*Underweight, Normal, Overweight, and Obesity* following BMI cut-off points specified by WHO for adults [97], this is: *Underweight* for BMI <18.5; *Normal* for BMI between 18.5-24.9; *Overweight* for BMI between 25.0-29.9, and *Obesity* for BMI ≥ 30.0.

Furthermore, based on the exclusion criteria established by the clinical experts, we excluded episodes for patients with less than four weight measures in the period, and entries with a void numerical result. We also applied processing algorithms to simplify the log according to the insights provided by the clinical experts, concretely we sequenced the traces, assuming the end of the current trace is the beginning of the next one, and fused them, merging consecutive traces with the same BMI value. At this point the event log was generated for 2,260 unique patients.

Following to the filtering and processing, and the generation of the event log with the desired characteristics, we inferred the process model itself by applying a Process Mining Discovery algorithm. In our case, PALIA discovery algorithm was applied to obtain the process model for the event log, concretely, the implementation provided by PMApp toolkit based on PALIA Suite tool [133]. At this point, the result neither was sufficient understandable, nor could be used for identifying patients' behaviours regarding their BMI. Therefore, we continued with the following stage of the methodology, the stratification analysis before the enhancement and the final graphical representation of the IPI.

Stratification analysis

The classic classification of obesity considers four main stages –*Underweight, Normal, Overweight, and Obesity*– therefore the idea was to obtain a close number of groups representing concrete BMI behaviours so we can redefine obesity. For this purpose, it was necessary to implement the fifth step of the proposed flow, and to use trace clustering techniques to stratify the population based on their behaviour. PMApp tool with WTD and QTC was used to obtain those groups (see section 4.1.1 for more details). Several experiments were performed with different quality threshold and similarity values to obtain the most meaningful fine-tuning results between the number of groups and the behaviour observed in each of them. Most significant results were obtained for a quality threshold of 0.12 and 0.01 similarity. Under these premises, nine groups were discovered (listed in Table 6.5) plus another one with 60 outliers (see Figure 6.4).

At this point we took the last task of the IPI formalisation back, the enhancement and graphical representation of the IPI. In this case, we used a colour map with a gradient scale for nodes and edges. For the nodes, the map was applied using the median time of stay in the node, whereas for the edges it was applied using the number of patients that proportionally followed the edge. The gradient colour goes from green (minimum) to red (maximum) as represented in Figure 6.3.

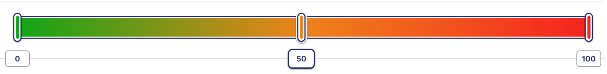


FIGURE 6.3: Gradient scale key for model representation from green to red.

The nine groups are listed in Table 6.5 from the most populated group to the less one, together with the novel semantic definition for obesity represented in each one of the clusters. The semantic definition is based on two concepts, the evolution and

TABLE 6.5: New semantic definition for dynamic obesity.

Group name	Population	% Total
Cluster 0: *Stable High Risk Model*	742	32.83%
Cluster 1: *Stable Increased Risk Model*	683	30.22%
Cluster 2: *Increasing Risk Model*	269	11.90%
Cluster 3: *Increasing High Risk Model*	204	9.03%
Cluster 4: *Decreasing Risk Model*	105	4.65%
Cluster 5: *Unusual Weight Behaviour 1*	57	2.52%
Cluster 6: *Decreasing to Normal Risk Model*	53	2.35%
Cluster 7: *Unusual Weight Behaviour 2*	47	2.08%
Cluster 8: *Stable Normal Risk Model*	40	1.77%

the associated risk of comorbidities with the BMI state [127]. Thus, observing the nine generated clusters showed in Figures 6.5, 6.6, 6.7, and 6.8, it is distinguished a two-stratification level. On the one hand, we detected an evolutionary norm showing a *Stable, Increasing or Decreasing* tendency and, on the other hand, the BMI state. We have associated each BMI to a risk situation [132], therefore underweight and normal BMI have associated a *Normal Risk*, overweight was linked with an *Increased Risk*, whereas obesity state is correlated with a *High Risk*. Combining these two patterns we classified the population into the groups included in Table 6.5. For instance, cluster 2 defined as *Increasing Risk* represents an increasing pattern over time, for those gaining weight from overweight to obesity class I, and consequently, increasing their risk to a higher situation. Cluster 0, *Stable High Risk* includes the population with an stable pattern over time and a high risk of comorbidities due to their obesity situation.

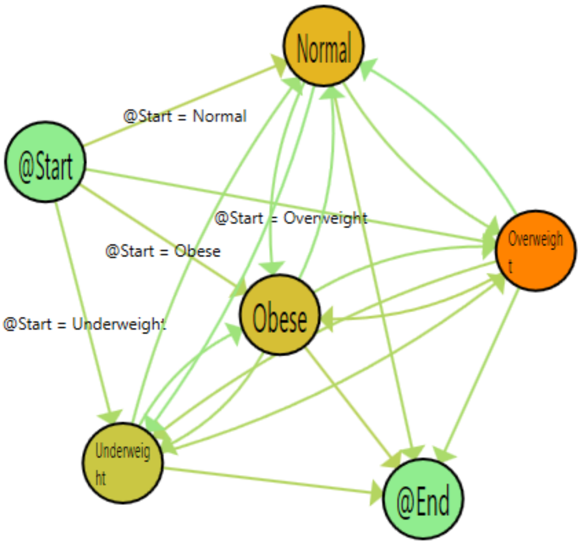
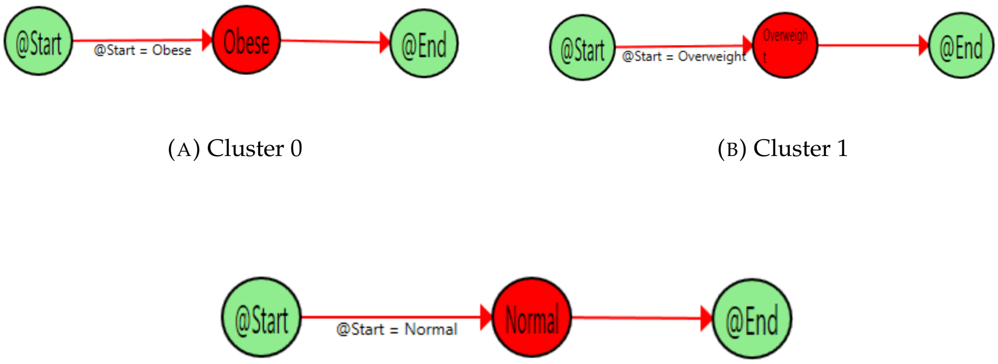


FIGURE 6.4: Dynamic obesity: outliers group.



(A) Cluster 0

(B) Cluster 1

(C) Cluster 8

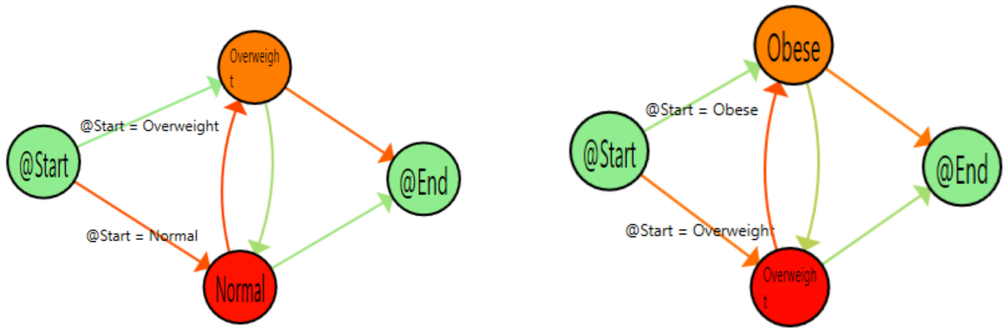
FIGURE 6.5: Models for BMI - Stable patterns. Presented in [29].

Considering the first level of stratification, we found that stable behaviours were the most common as 64.82% of the considered population showed a stable BMI behaviour (Figure 6.5), followed by increasing behaviour (20.93%) (Figure 6.6), and decreasing one (7%) (Figure 6.7), highlighting the fact that gaining weight is much more prevalent than losing weight. However, the most noteworthy issue was that only 1.77% of the population presented a stable and normal weight. We also found two other clusters, included in Figure 6.8, representing unusual models that are outside the new semantic definition of obesity, because at first sight they do not represent any behaviour, although they need to be validated with clinical experts in the last stage of the methodology.

Analysing each behavioural group in detail and within the *stable patterns*, it included the three groups shown in Figure 6.5. The group *Stable High Risk Model* represented in Figure 6.5a includes the obese population all the period maintaining an associated high risk. Similarly, those included in the *Stable Increased Risk Model* group showed in Figure 6.5b were overweight which is associated with an increased risk situation compared with a normal one. Finally, the group named as *Stable Normal Risk Model* shows a stable normal BMI associated with a normal risk of comorbidities (Figure 6.5c).

The next most common behaviour group includes the *increasing patterns* with two models as shown in Figure 6.6. The first model, *Increasing Risk* contains a population increasing their risk as they are mostly moving from a normal weight to an overweight situation (see Figure 6.6a). The second model named as *Increasing High Risk* represents the population mainly gaining weight from overweight to obesity and therefore increasing their risk to a higher one (see Figure 6.6b). Within the models we could also observed the time spent in each state, letting us to infer that they were not isolated measures, the population spent a considerable time in the final states of the models, confirming the tendency.

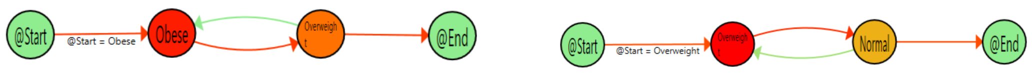
Finally, the two models for the *decreasing patterns* are shown in Figure 6.7. The first model *Decreasing Risk* represents a population who is decreasing their associated risk to an increased situation compared with the normal one (see Figure 6.7a)



(A) Cluster 2

(B) Cluster 3

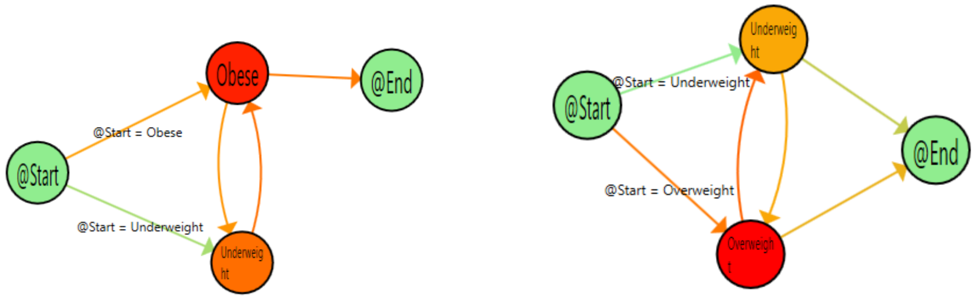
FIGURE 6.6: Models for BMI - Increasing patterns. Presented in [29].



(A) Cluster 4

(B) Cluster 6

FIGURE 6.7: Models for BMI - Decreasing patterns. Presented in [29].



(A) Cluster 5

(B) Cluster 7

FIGURE 6.8: Models for BMI - unusual patterns. Presented in [29].

as they were losing weight from obesity to overweight state. The second model *Decreasing to Normal Risk* shows a population that is normalising their associated risk as they abandoned the overweight state in favour of a normal weight (see Figure 6.7b). However, the enhanced models highlight that time spent in the firsts states –obese and overweight– was considerable and benefits from the weight loss might not be as expected regarding the associated risk. In these models is where the potentiality of the novel re-definition is more appreciated, as we are not only considering the final BMI stage but the whole process, therefore it can be analysed the time spent in each stage, the transitions,and when these transitions occurred, among others.

TABLE 6.6: Dynamic Obesity Risk Model groups.

Group name	Population	% Total
Model 1 - Stable High Risk	742	32.83%
Model 2 - Stable Increased Risk	683	30.22%
Model 3 - Increasing Risk	269	11.90%
Model 4 - Increasing High Risk	204	9.03%
Model 5 - Decreasing Risk	105	4.65%
Model 6 - Decreasing to Normal Risk	53	2.35%
Model 7 - Stable Normal Risk	40	1.77%
Others	164	7.26%

As said, the two remaining groups represented in Figure 6.8 show patients moving between two different BMI stages that supposed a big weight lost or gain in a considerable short period. Again, the results and the methodology showed their capabilities, as they allowed us to navigate from the model to the individuals, and check what exactly happened with these patients. Concretely, cluster 5 and cluster 7, shown in Figure 6.8a and 6.8b, included a population moving between obesity and underweight; and between underweight and overweight, respectively, always in less than three moths. These groups were left out of the re-definition of obesity until the validation by the clinical experts, as they show unusual patterns.

IPI2: Dynamic relationship between obesity and hypertension

The objective of the second posed question was to discover how obesity is related with hypertension, for this, we contemplated the results obtained in the previous IPI and the re-definition of obesity as a dynamic process through the groups included in Table 6.6. Consequently, these seven groups supposed the point of departure for the experimentation of this IPI, as we considered them as a new classification for obesity and therefore, we did not examine obesity under *Underweight, Normal, Overweight, and Obesity* categories, instead we utilised *Stable High Risk, Stable Increased Risk, Increasing Risk, Increasing High Risk, Decreasing Risk, Decreasing to Normal Risk, and Stable Normal Risk*. In the *Others* group we included the rest of the population, it is the outliers and the patients included in the unusual patterns –Cluster 5 and Cluster 7– that represent a total of 164 patients, the 7.26% of the considered population.

In order to do so, we incorporated the semantic definition of Dynamic Obesity Risk Model into our model as a new variable characterising the patients, and then discovered the hypertension flow for each group [127]. As in the previous IPI, we also applied processing algorithms to simplify the log, sequencing and fusing traces. At this point, PALIA discovery algorithm was applied to obtain the hypertension processes for the models included in Table 6.6.

Again, we used a colour map with a gradient scale for nodes and edges, using the median time of stay in the node, and the number of patients that proportionally followed the edge for each transition. The gradient colour goes from green (minimum) to red (maximum) as represented in Figure 6.3. The hypertension flows obtained for the seven Dynamic Obesity Risk Models after applying the previous experiment procedures are shown in Figure 6.9.

Analysing the different models and their hypertension flows, Figure 6.9a shows how population within the Stable High Risk Model has mainly high BP, as they are in elevated, and hypertension stage 1 and 2. In the case of hypertension behaviour for Stable Increased Risk Model, included in Figure 6.9b, the population spent most of the time between normal and high BP. The following model, the Increasing Risk (Figure 6.9c), illustrates BP flow for this population, highlighting that although the most common path is normal BP, some people included in this risk model were experimenting long episodes of hypertension stage 1, applying median to duration time spent per node. This situation endorses the fact of excess weight is translated into higher risk of hypertension. Contrary to our expectations, BP flow for the Increasing High Risk Model included in Figure 6.9d shows that population spent most of the time in normal BP with some high BP episodes, but far from a high risk situation expected from the BMI behaviour. Looking into the decreasing weight groups, in Figure 6.9e is presented the dynamic BP flow for the Decreasing Risk Model. The flow shows that patients spent most of the time in hypertension stage 1, followed by stage 2. Although patients are losing weight, the effects of this improvement on their health status have not yet been noticed on BP. Similarly, Figure 6.9f includes the BP flow for patients decreasing their risk to a normal situation thanks to a weight loss, however they consume a significant time in high BP states. Finally, population within a Stable Normal Risk Model followed a normal BP flow, with some episodes of elevated BP, without time significance, supporting the idea of a low risk thanks to a normal weight model.

Clinical validation

The last stage of the proposed methodology implies the validation of the corresponding IPIs with the clinical experts in the field. This validation was undertaken by means of sharing the results obtained with the clinical experts who formulated the posed questions and checking the results with the available literature. The main objectives were, on the one hand to evaluate the results from a clinical perspective. And on the other hand, to confirm the ability of the methodology to answer questions posed by clinical experts in the the field of chronic conditions.

Regarding the first objective, the first IPI permitted to redefine the classification of obesity from a dynamic perspective, obtaining a Dynamic Risk Models comprising seven groups plus two unusual behaviours due to the limitations and characteristics of the study. The two unusual patterns were also analysed during this stage. Clinical experts supported the idea of measurement errors for the group represented in Figure 6.8a, as the flow shows a population moving from obesity to underweight and moving another time to the initial situation in a very short period (always in less than three months). For the other pattern, unusual weight changes was the most plausible explanation for the cluster represented in Figure 6.8b, corresponding with special situations such as surgeries or pregnancy. Consequently, they were associated with the limitations and characteristics of the study and kept outside the risk model by the clinical experts.

The results show how Process Mining and trace clustering techniques were able to infer obesity generic behaviour of patients in a broad way, and to stratify them into well-defined groups based on their behaviour flow evolution. This supports the idea of considering obesity and related variables, such as blood pressure, as dynamic ones, showing how they evolve over time, and representing considered population behaviours, answering the first question posed by the clinicians, *Could obesity be re-defined as a dynamic process?* by *Stable High Risk, Stable Increased Risk, Increasing Risk, Increasing High Risk, Decreasing Risk, Decreasing to Normal Risk, and Stable Normal Risk* as the new classification for obesity. Moreover, using this novel definition we were able to deal with the second question, *How dynamic obesity is related with its comorbidities over time?*, and respond it for the concrete case of hypertension, not only in the identification of the relationship, but also in the discovery of the behavioural evolution of the hypertension itself for the different obesity models.

Accordingly, it was possible to confirm that the proposed methodology provides the necessary steps to answer the posed questions in the field of chronic conditions as Dynamic Risk Models.

6.3.1. Contributions

The application of the proposed methodology enabled to answer the two posed questions by the clinical experts in the field of chronic conditions, concretely about overweight and obesity, into two IPIs representing the Dynamic Risk Models associated with the re-definition of obesity as a dynamic process, and its dynamic relationship with hypertension. The two IPIs were presented as graphical models easily understandable by health professionals, one of the two pillars established as fundamentals for the methodology, as well as the interactive inputs, as it was specified in Figure 6.1. The utilisation of the question-based methodology proposed by [125] together with Interactive Process Mining paradigm permitted to formulate a novel interactive and question-based methodology for deploying Dynamic Risk Models in the field of chronic diseases.

The methodology includes six well-defined steps, with clear objectives and the appropriate tasks to achieve the desired outcomes. Its results, IPIs representing the Dynamic Risk Models, could be used for understanding and measuring chronic underlying processes as Interactive Process Mining has the potentiality of presenting findings over data in a comprehensive manner for health experts so they could find new medical evidence. Furthermore, the presented methodology and experimentation flow can be self-adapted to any target population and be automated over other data sources. The methodology allows clinical experts to propose problems and question related to chronic conditions, and analyse individual and groups behaviours. In this procedure, interactions between clinical and Process Mining experts are paramount to transform data from EHR to understandable and valuable information.

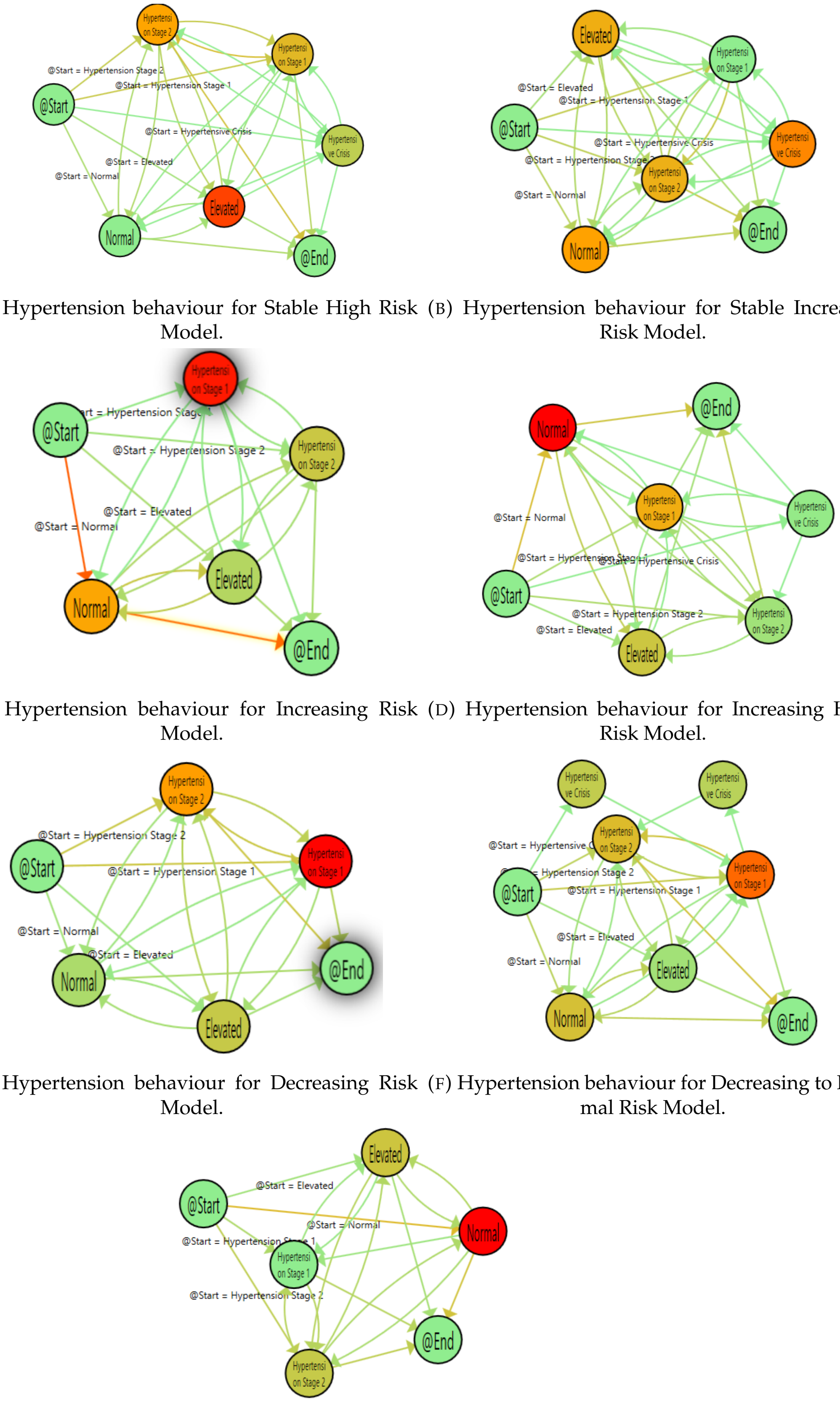
Putting into practise the proposed methodology, we obtained two different IPIs answering two posed questions, an IPI re-defining obesity as a dynamic process, resulting into seven well-characterised groups based on the associated risk. Three evolution patterns were discovered within the models, one pattern for patients with a stable weight evolution, but two other groups changing their weight, with increasing and decreasing evolution. This finding is very relevant, as we could stratify population based on this weight evolution, we were even capable to detect measurement errors. If we consider two patients from two different risk models, the first one from the *Increasing Risk Model* group (Figure 6.6a) and the second one from *Decreasing Risk Model* (Figure 6.7a) group; they have the same BMI at the end of the period, it is overweight; but their behaviours are clearly different. In a classic and static approach, the only insight is the BMI result, however this dynamic view let us consider other dimensions of the problem. The first patient is gaining weight, worsening her/his health status, consequently the patient is probably not well-engage with diet counselling, not properly motivated, or even not treated or diagnosed. On the other hand, the second patient is losing weight, she or he is doing things well and treatment is working. Therefore, interventions should not be the same for these two patients to improve their health status. In the first case, health professionals should focus on general lifestyle changes, patient literacy in these aspects and look for personal characteristics. When in the second case, they should reinforce the patient attitudes that resulted in the weightloss. In this way, the IPI allowed classifying population regarding weight behaviour. Moreover, aggregating another variable such as the BP in combination with the *Dynamic Risk Model for Obesity* we obtained a new IPI that related dynamic obesity with hypertension over time.

6.3.2. Limitations

To establish a methodology that allows answering questions posed by clinical experts we should consider two main issues. On the one hand, the nature of the diseases over the questions would be formulated, and on the other hand, the characteristics of the questions themselves, since we have to assume that we will not be able to answer or resolve any question.

One of the main chronic condition's characteristics to be considered is having an associated risk factor and thus, one of the main limitations to apply the methodology and obtain Dynamic Risk Models. As explained in Chapter 2, chronic diseases share common risk factors that can be modified and are common for women and men. They are the intermediate or metabolic risk factors of blood pressure, raised glucose levels, abnormal blood lipids, overweight, and obesity. The first two steps of the proposed methodology deal with the question formulation and the associated risk factors' which is crucial to answering it, and consequently one of the major limitations of the methodology.

Furthermore, the methodology exposes their potentiality working with chronic diseases that might benefit from a temporal or dynamic analysis. In these cases, the current snapshots of the gathered variables present gaps, not showing the variability or the evolution, as is the case of obesity disease in which the current state is as important as the evolution or the previous states to determine the associated risk. On the other hand, the questions to be represented by the Dynamic Risk Models through the methodology should represent what and how clinicians want to inquire about the underlying chronic conditions, however as said, not all questions would be answer thanks to the Dynamic Risk Models as a result of the corresponding methodology. The methodology might not work with too complex diseases or questions, meaning another important limitation of the proposed work. Although, a balance in the sampling guarantees a high quality data for producing high quality models. This supposes another way the models might be inaccurate, therefore it is needed to measure this balance using appropriate ways, such as the variance coefficient in the third step of the methodology.



(A) Hypertension behaviour for Stable High Risk Model.

(B) Hypertension behaviour for Stable Increased Risk Model.

(C) Hypertension behaviour for Increasing Risk Model.

(D) Hypertension behaviour for Increasing High Risk Model.

(E) Hypertension behaviour for Decreasing Risk Model.

(F) Hypertension behaviour for Decreasing to Normal Risk Model.

(G) Hypertension behaviour for Stable Normal Risk Model.

FIGURE 6.9: Dynamic relationship between obesity and hypertension.

Chapter 7

Conclusions and Future Work

Throughout this document, it has been stated how risk models in the health domain are valuable tools, they are being used to forecast the risk of populations, to determine the risk of an adverse event using a specific threshold for the parameter of interest, such as age combined with hospital admissions, or as a statistical model to provide patient's risk of an adverse episode in the future. However, they do also present some limitations as they are understood, as they do not usually consider the individual level, and the dynamic perspective on both the patients and the variables to be considered. To tackle this limitation, the work included in this document has presented a novel approach by way of the definition of the –*Dynamic Risk Models*– to deal with some of the lacks of the classical risk models' means, together with the implementation of a formal framework to obtain these Dynamic Risk Models for chronic conditions.

This chapter concludes the work and demonstrates how the specific research questions and objectives posed in Chapter 1 have been addressed, to confirm their proper achievement. For this, the main results obtained throughout this doctoral thesis are listed and analysed below.

7.1. Conclusions

The work presented during the whole document is based on the hypothesis, the research questions and the secondary objectives formulated at the beginning of the work (Chapter 1). Concretely, there were established three secondary objectives to answer the research questions. These secondary objectives were formulated over three main pillars, the analysis of the viability of the novel approach to medical conditions using Process Mining techniques, the particularisation for chronic conditions, and the proposal of a formal methodology based on the work carried out and the experience acquired. For this to end, the first objective was:

> **O1 - To evaluate the viability of approaching a medical condition using Process Mining techniques to find behavioural dynamics**

To reach this objective, Process Mining techniques were applied to a dataset from a bounded population in risk of malnutrition, concretely regarding their weight and malnutrition status through the MNA test results. This enabled the analysis of patients patterns, combining results from both BMI and MNA as included in Chapter 4. The obtained results incorporated the dynamic perspective for BMI, and showed an abstraction of the underlying process. Furthermore, the results even allowed to find new evidences that the classic approach did not detected [100]. The aggregation of the dynamic BMI model with the MNA scores stated that nutritional interventions worked for those patients underweight during all the study but generally did not work for those above a normal weight.

Moreover, Process Mining complemented with trace clustering techniques facilitated the generation of patients' groups according to similar behaviours. Using Process Mining techniques, it was possible to generate models representing the flow followed by patients regarding weight changes. Whereas the use of trace clustering techniques allowed differentiating certain behaviours from others. One of the main finding was that patients in nursing homes can be classified based on weight dynamic behaviours, showing if nutritional interventions would change these behaviours or not comparing cohorts of the same population in different moments [100].

These results allowed to confirm that approaching a medical condition using Process Mining techniques to find behavioural dynamic is not only possible but it also allows to discover new and interesting knowledge from the health professionals perspective.

Once the viability of the approach was validated, the work carried out deal with the particularisation for chronic conditions, as it was stated in the second objective defined at the beginning of the document:

> **O2 - To approach chronic health conditions using Process Mining techniques and information from EHR to obtain behavioural risk models**

To achieve this goal, Process Mining techniques were used in combination with retrospective data from EHR to discover and obtain behavioural risk models associated to chronic conditions. The first step was to model a medical conditions thanks to a bounded population as presented in Chapter 4, that allowed us to validate the approach and the needed steps, so they can be personalised to specific chronic conditions, and to investigate whether they could be used for representing Dynamic Risk Models that characterise them based on the evolution using Process Mining and trace clustering techniques. Using data from the EHR of a tertiary hospital we modelled two common and prevalent chronic conditions – hyperglycemia and hypertension– using the associated risk factors, blood sugar and blood pressure, respectively [29]. As a results, two valuable and innovative Dynamic Risk Models were obtained as presented in Chapter 5.

As results showed in [29], Process Mining techniques has permitted to characterise the population in a dynamic and personalised way for the two considered conditions. In the first case, the *Dynamic Risk Model for Hyperglycemia* used the FPG to evaluate the continuum and evolution of blood sugar management. The Dynamic Risk Model revealed that the time spent in diabetes and intermediate hyperglycemia stages was, on average, considerably high for the group considered. As the model took into account the behaviour over time, instead of a concrete situation or the transition between two concrete values, it was possible to see the patient's evolution.

In the second case, the *Dynamic Risk Model for Hypertension* showed thirteen different patterns with the continuum of BP and its evolution. By analysing the model groups, health professionals could infer if common BP behaviours have associated common population characteristics or patterns. Once again, health professionals could compare different groups' BP evolution, personalise interventions for the different groups and test their efficacy and effectiveness over time.

Going a step forward, Chapter 6 has proposed a third Dynamic Risk Model for obesity chronic diseases in combination with one of its co-morbidity, hypertension. Concretely, there were obtained two Dynamic Risk Models for the re-definition of obesity as a dynamic process, and its dynamic relationship with hypertension, using data from the EHR of a tertiary hospital [131], [127], as presented in Chapter 6.

Therefore, one medical condition –malnutrition– and three chronic conditions –hyperglycemia, hypertension and obesity– were approached using Process Mining techniques and data from EHR to obtained their dynamic behavioural models, as presented in Chapters 4, 5 and 6, respectively. These models supposed a supporting tool for advancing in the personalised medicine concept, incorporating evolution over time and patient's unique behaviour to the analysis.

The utilisation of Process Mining techniques and data from EHR in a concrete healthcare scenarios for concrete chronic diseases, permitted to obtain valuable and innovative Dynamic Risk Models in a formalised manner that could be used for understanding and measuring chronic underlying process abstractions, following concrete procedures and experiments. In this line, the third objective proposed at the beginning of the work was:

> **O3 - To propose a formal methodology to translate chronic conditions evolution into dynamic flows using Interactive Process Mining techniques**

To pursue this objective, Chapter 6 proposed a formal methodology for obtaining Dynamic Risk Models. The methodology defined as *interactive and question based methodology for deploying Dynamic Risk Models for chronic conditions using Process Mining* is an adaptation of the question-driven methodology specific for emergency rooms using Process Mining proposed by Rojas et al. in [125]. The methodology was designed to support medical and Process Mining experts in the understanding of chronic disease's processes, bringing them a straightforward guideline of how to apply Interactive Process Mining to the analysis of chronic disease's underlying processes. In this scenario, the medical expert contributes with her/his deeper knowledge and common sense of the chronic disease formulating questions, but also within the verification and correction of the results during the different stages of the methodology. On the other hand, the Process Mining expert adds her/his awareness on how to apply Interactive Process Mining techniques for obtaining the corresponding Dynamic Risk Models in the form of IPIs [131], [127]. This formal methodology was implemented through six well-defined steps covering the definition of the corresponding question by the health expert, the analysis of the risk factors and variables to be used to answer the posed question; the verification of the data quality and availability; the formalisation of the IPI associated with the posed question; the application of the stratification analysis; and the validation of the results with the health experts.

The attainment of the aforementioned objectives, was directly related with the achievement of the research questions proposed in Chapter 1, concretely the first research question identified was:

> **RQ1 - Can the evolution of a chronic condition be modelled using Process Mining techniques in an understandable manner for human?**

To validate the Process Mining techniques' capability for modelling the evolution of a chronic diseases and to answer this research question, it was studied several medical condition, such as malnutrition [100] in Chapter 4, hypertension and hyperglycemia [29] in Chapter 5, and obesity [131], [127] in Chapter 6. These chronic conditions were selected not only because they have a clear dynamic component but also because of their availability and prevalence among the population (hypertension and hyperglycemia), and their complexity in the diagnosis (obesity). The experiments performed and the results obtained have confirmed that the evolution of a chronic disease can be modelled using Process Mining techniques in an understandable manner for humans.

> **RQ 2 - Can we create a patients' behaviour based stratification (Dynamic Risk Models) using Process Mining techniques that allow us to build a perspective for better understanding the chronic conditions' evolution?**

To solve this research question we started from the fact that Process Mining techniques can construct individual and human behaviour models, allowing to include individual determinants and variability and evolution over time to the risk models. We have proposed an automatic stratification system based on trace clustering techniques that presents a set of groups separated depending on their behavioural differences. It enables healthcare experts in the understanding of evolutionary aspects of the considered condition in terms of the associated risk in an iterative and interactive manner. Combining Process mining with trace clustering techniques it was possible to group patients with similar behaviours, creating what we called Dynamic Risk Models. Therefore, corresponding experiments were conducted incorporating both techniques, Process Mining and trace clustering to obtain Dynamic Risk Models that incorporate the chronic condition's evolution. This was explained in Chapter 5 for hypertension and hyperglycemia, and in Chapter 6 for obesity [29], [131], [127].

> **RQ 3 - Can we define a formal methodology for approaching the dynamic perspective of chronic conditions using Interactive Process Mining techniques to obtain Dynamic Risk Models for chronic diseases?**

To answer this research question it was implemented an interactive and question based methodology using Interactive Process Mining techniques to obtain Dynamic Risk Models for chronic diseases that allows approaching chronic conditions [127]. This methodology was based on the application of the automatic based stratification system in an iterative and interactive way. This allowed the definition of a novel manner to characterise the chronic conditions' behaviour. Furthermore, the methodology was put into practise through a use case focused on obesity chronic disease, and effectively obtaining the Dynamic Risk Models for obesity and its relationship with hypertension, as presented in Chapter 6.

The proposed methodology, the experiments and the introduced definition for the Dynamic Risk Models have been implemented and developed with the main goal of achieving the central hypothesis of the present work:

> *The dynamic perspective of chronic conditions can be incorporated to risk models using historical data coming from Electronic Health Records and Process Mining techniques to analyse them, offering a novel way to better understand the diseases' behaviour with a process oriented view.*

The experiments and procedures carried out throughout this work have shown, according to this hypothesis, that it is possible to dynamically model a chronic condition using historical data and Process Mining techniques. Results have offered a novel approach to risk models with a process oriented view that have permitted a better understanding of the chronic diseases' behaviour, as included in Chapters 4, 5 and 6 where new evidence was discovered by the health professionals participating in the work thanks to the analysis of the results.

7.2. Discussion and Future Work

In this thesis, we have tackled the incorporation of the dynamic perspective to the risk models using Process Mining techniques and EHR data, moreover we have proposed an interactive and question based methodology that allows generating this models in an understandable manner for health experts in the field of chronic diseases. One of the main interest of these novel Dynamic Risk Models is that they are understood by the experts in the field, so they could be continuously updated, improved and adapted based on their needs, expectations, knowledge or population under study, thanks to the Interactive paradigm incorporated in the methodology throughout the use of the Interactive Process Mining techniques. Furthermore, the Dynamic Risk Models could be defined and personalised for concrete populations with particular needs or characteristics. Factors outside the healthcare system, such as social determinants including education, employment, housing, income, or social inclusion play an important role not only in the healthcare outcomes but also in the way care is perceived by the individuals. Including those determinants in the Dynamic Risk Models is possible thanks to the methodology and the different procedures implemented in it. Consequently, the use of the Dynamic Risk Models could suppose obtaining more personalised models that at the end will permit a better understanding of the diseases and clinical cases.

The classical risk models lack on this personalisation or in the consideration of other determinants outside the healthcare system variables. Thus, classical risk models could be misleading including people in a group that does not represent them in the best manner and in consequence they would not be treated accordingly, such as for example people trying called *miracle diets*. In miracle diets, people usually experience a fast weight loss, however it can not be maintain in the long-term and people finally gain weight back. Even more, people usually try several of those diets, with the associated health risk. However, only if we have the complete picture of the person, she/he would be correctly classified and treated, and this is only possible incorporating the dynamic and behavioural variables to the risk models. The patients' behaviour based stratification behind the Dynamic Risk Models allows stratifying patients with similar behaviours, maximising the differences in the risk evolution, and therefore classifying patients in the group that better represents their behaviour.

Other of the potentialities of the results presented in this work is that Dynamic Risk Models could be built upon all the information available in the EHR, and not only considering the last measure or variable or a set of them, this supposes they are mathematically more efficient that the classic risk models [48]. However, when considering data coming from EHR or other hospital systems the amount of data is not the only important aspect, but also the quality of the data. The proposed methodology supports the analysis of the quality of the data sampling as an important step. Fact that current models do not do in that clear and direct manner.

Furthermore, working with the proposed Dynamic Risk Models we are able to obtain richer models thanks to this personalisation through the incorporation of individual preferences, and social, mental and health determinants, but also the variability and evolution of a disease over time. These models could result in the discovery of new evidence in the medical domain that leads to improving the chronic diseases' knowledge, and their management from both healthcare experts and patients. This would be also supported by the understandability of the Dynamic Risk Models. Dynamic Risk Models are opposite to the *black-box* concept in which obtained models are too complex to be easily understood by the healthcare experts and in consequence not usually trusted. The Dynamic Risk Models are defined incorporating the expert thanks to the Interactive paradigm, therefore their results are understandable by them. It supposes they would trust on them as they comprehend the rationale behind the decisions of the technology as they are partially conducted by their knowledge and common sense.

The methodology and procedures presented in the different chapters of this work, suppose for the author the point of departure for a new promising framework that enables extracting knowledge from data in the field of chronic diseases. Concretely, the following lines are identified for future work:

- On the one hand, the author envisages the development of Dynamic Risk Models focused on other diseases or pathology, for which the time and evolutionary perspective could suppose an added value to the current knowledge. In this line, a new research work is being developed by the author in the field of cancer treatment and post-treatment period in the framework of the LifeChamps EU H2020 project[1]. In this context, the application of Process Mining techniques to sensor data could be applied to individual patient ill-health trajectory modelling, visual exploration of interacting cancer symptoms and comorbidities signs, patient stratification, and quality of clinical cancer care service, combining clinical events from EHR, patients' response to treatment through

[1] https://lifechamps.eu/

sensors and patient-reported outcome measures and patient-reported experiences measures, analysing patients' perceptions about their own health status and experiences whilst receiving care. Another promising action is taking place in the field of palliative care in the framework of the InAdvance EU H2020 project[2]. In this project, the author is applying the results of this thesis to obtain human-understandable graphical representations about palliative care pathways and interventions that could support healthcare professionals in comprehending current processes, taking into consideration patients' variability and nature, and clinical settings characteristics. This will be possible by applying the techniques that SABIEN research group has been using during the last years [78], [70].

- Although in this work we have worked with data collected from EHR, the procedures and experiments can be self-adapted to combine this information with other sources, such as smart sensors and personal devices. Adding such new sources could leverage the obtained models to a new level, as the Dynamic Risk Models resulted analysing these data sources could enrich them, showing more complete and accurate individual behaviours. In this field, and also in the framework of the LifeChamps EU H2020 project, data coming from EHR and reported by the patients is going to be combined with sensors monitoring different patients' variables at home and portable. Concretely, it incorporates wearable sensors, medical grade devices (blood pressure monitor, thermometer, and weight and body composition scale), and IoT smart plug for monitoring the activity of home appliances and location system. All these data will be analysed using the results established by this work for modelling breast, prostate and skin cancer patients.

- When considering the proposed methodology, one of the most important issues, largely explained in Chapter 6, is the health experts involvement in all the procedures, therefore it is paramount to comprise the involvement of clinicians not only to validate the clinical utility of these results, but also to measure the validity in concrete patients. During the realisation of this doctoral thesis we counted with the active involvement of the healthcare professionals collaborating in the different projects in which the research was done. Their participation was crucial posing the correct questions and validating the results, and thanks to the background acquired we realised how important is their involvement and their awareness of the methodology and the rationale behind. Future work in this line should include the development of best practices to involve healthcare experts in the use of the methodology, developing training materials so they could create their own Dynamic Risk Models after a training period, and doing dissemination efforts to engage them within this promising work.

- The dynamic modelling of chronic conditions is a complex task that could involve several dimensions and variables, for example, in the framework of the Dynamic Risk Model for hypertension, we have worked with a mutivariable measure, as the blood pressure stage is based on two variables, the systolic

[2] http://www.inadvanceproject.eu/

blood pressure and the diastolic blood pressure, nevertheless there is still a long path for research in this concrete aspect. This cloud include risk factors involving multivariable measures such as the metabolic syndrome that involves a group of risk factors including diabetes, high blood pressure (hypertension) and obesity; or the combination of multiple variables to gain the best understanding of a diseases, or even combining several Dynamic Risk Models, such as the combination Obesity Risk Model and type 2 diabetes or low back pain, as we did with obesity and hypertension in Chapter 6. The complexity of the analysis will increase, but this could also suppose to obtain richer and more personalised models.

- In the framework of this doctoral thesis, we have also dealt with the quality of the sampling data and how it could affect to the Dynamic Risk Models quality obtained, as we observed during the performance of the different experiments. For that purpose, in this work we included some tools for assessing the quality of the models, such as the variation coefficient, however we consider it could be not enough for a complete measure of the quality of the data or for all the considered situations. For that, future work in this regard is needed, looking for the creation of new systems, tools, or algorithms for measuring and assuring the quality of data regarding the sampling rate. Concretely, the development of new tools that also analyse the quality of the obtained models based on the sampling rate quality. In this context, another area for further research is the definition of standard data structures that enable to automatically obtain data from the hospital systems in an adequate manner and with the needed quality to support the automatic adaptation of a defined Dynamic Risk Model to other healthcare centres or scenarios.

Chapter 8

Main original contributions

In this doctoral thesis, solutions have been provided to problems not covered by the current risk models. In this way, a new formalised methodology has been created and published for discovering Dynamic Risk Models for chronic diseases, as well as the Dynamic Risk Models for three of the nowadays most prevalent chronic diseases, obesity, hypertension, and hyperglycemia to support the methodology. The main objective of this chapter is to collect the main and original contributions performed in each of the sections of this work. The achievement of the previous objectives has allowed the dissemination of the results in several scientific forums, such as book chapters, international congresses, and indexed journals. This research work has been carried out within the PM4Health Lab[1] at SABIEN research group[2] (Technological Innovation for Health and Well-Being) part of the Institute of Applied Information Technologies and Advanced Communication (ITACA), affiliated to Universitat Politècnica de València (UPV), and situated in the Polytechnic City of Innovation (CPI). In the framework of SABIEN is where the author of this doctoral thesis has performed most of the research participating in several national and European during the last 10 years.

Following, the original contributions done in each of the points discussed in this work are listed in the form of scientific publications associated with the work carried out on this doctoral thesis, as well as the research projects where the concepts studied have been applied.

8.1. Associated publications

During the performance of this work, a new formalised methodology has been created and published for discovering Dynamic Risk Models for chronic diseases, as main as the Dynamic Risk Models for three of the most prevalent non-communicable diseases, obesity, hypertension, and hyperglycemia to support the methodology. The following journal publications, books and conference contributions describe the results achieved, and reflect the work done in the context of the specific research projects that required a novel approach for well-known chronic conditions. Next, the original contributions made in each of the points discussed in this work are listed in the form of scientific publications associated with the work carried out on this document, as well as the research projects where the concepts studied have been applied.

[1] https://pm4health.com/

[2] http://www.sabien.upv.es/

8.1.1. Publications in journals

- **P1 -** Ibanez-Sanchez, G.; Fernandez-Llatas, C.; Martinez-Millana, A.; Celda, A.; Mandingorra, J.; Aparici-Tortajada, L.; **Valero-Ramon, Z.**; Munoz-Gama, J.; Sepúlveda, M.; Rojas, E.; Gálvez, V.; Capurro, D.; Traver, V. "Toward Value-Based Healthcare through Interactive Process Mining in Emergency Rooms: The Stroke Case". In: *International Journal of Environment Research and Public Health* 16.10 (2019), p. 1783. [55]. **Impact Factor 2.849, Q2**

  An analysis of how Process Mining techniques can support health professionals in the application of Value-Based Technologies to demonstrate the possibilities of Process Mining in the characterisation of health conditions processes. The results were published in the Special Issue Process-Oriented Data Science for Healthcare 2018 of the International Journal of Environment Research and Public Health with the title "Toward Value-Based Healthcare through Interactive Process Mining in Emergency Rooms: The Stroke Case".

- **P2 - Valero-Ramon, Zoe**, et al. "Dynamic Models Supporting Personalised Chronic Disease Management through Healthcare Sensors with Interactive Process Mining". In: *Sensors* 20.18 (2020), p.5330 [29]. **Impact Factor: 3.576; Q1**

  The results from the application of Process Mining techniques to obtain Dynamic Risk Models for chronic conditions based on patients' dynamic behaviour provided by health sensors were published in the distinguished journal Sensors with the title "Dynamic Models Supporting Personalised Chronic Disease Management through Healthcare Sensors with Interactive Process Mining". The outcomes validated the viability of the approach based on the dynamic behaviour of metabolic risk factors associated with particular chronic diseases.

- **P3 -** Gatta, R.; Vallati, M.; Fernandez-Llatas, C.; Martinez-Millana, A.; Orini, S.; Sacchi, L.; Lenkowicz, J.; Marcos, M.; Munoz-Gama, J.; Cuendet, M.A.; de Bari, B.; Marco-Ruiz, L.; Stefanini, A.; **Valero-Ramon, Z.;** Michielin, O.; Lapinskas, T.; Montvila, A.; Martin, N.; Tavazzi, E.; Castellano, M. "What Role Can Process Mining Play in Recurrent Clinical Guidelines Issues? A Position Paper". In: *International Journal of Environmental Research and Public Health* 17.18 (2020), p. 6616. [134]. **Impact Factor 3.390, Q1**

  The theoretical framework for the formalisation of the use of Process Mining to the healthcare domain, and concretely in the clinical Guidelines issues was published in the Special Issue Process-Oriented Data Science for Healthcare 2019 of the International Journal of Environment Research and Public Health, with the title "What Role Can Process Mining Play in Recurrent Clinical Guidelines issues? A Position Paper". The work explored the relevance to investigate concerns about the patients' individual differences in diagnostic/treatment care and how they cause great variances in the execution of the healthcare processes.

8.1.2. Book chapters

- **P4 - Valero-Ramon, Zoe** et al. *Towards Perceptual Spaces for Empowering Ergonomy in Workplaces by Using Interactive Process Mining.* 2019 [8].

  A proof of concept of how Interactive Process Mining technologies can be used for discovering flows regarding a concrete population was presented as a chapter in the IOS Press book Transforming Ergonomics with Personalised Health and Intelligent Workplaces, with the title "Towards Perceptual Spaces for Empowering Ergonomy in Workplaces by Using Interactive Process Mining". Concretely, the work explained how Interactive Process Mining technologies can be applied to employees workflows to support the ergonomy experts in the selection of more accurate interventions for improving occupational health.

- **P5 - Valero-Ramon, Zoe**, et al. "Interactive Process Indicators for Obesity Modelling Using Process Mining". In: *Advanced Computational Intelligence in Healthcare-7*. Springer, 2020, pp. 45-64 [131].

  The results of the application of the Interactive Process Mining methodology to the analysis of the BMI and the available data of the comorbidities associated with obesity, from a dynamic perspective thanks to the use of Process Mining tools, to obtain behaviour patterns of the patients was published as a book chapter by the prestigious editorial Springer, with the title "Interactive Process Indicators for Obesity Modelling using Process Mining". The work presented a set of human-readable and contextualised Interactive Process Indicators (IPI) in the field of obesity, and its related conditions that helped healthcare professionals to interact with the process. Modelling IPI as enhanced views would help professionals to better understand these processes. Moreover, professionals would monitor patient progress in an iterative manner and interact with the system to adjust interventions and treatments.

- **P6 - Valero-Ramon, Zoe** and Carlos Fernandez-Llatas. “Interactive Process Mining for Discovering Dynamic Risk Models in Chronic Diseases”. In: *Interactive Process Mining in Healthcare*. Springer, 2021, pp. 243–266 [127].

  The Interactive Process Mining methodology proposed in this doctoral thesis to obtain Dynamic Risk Models has been used as a new way to stratify the patient’s behaviour with chronic conditions. Concretely, there were created two different IPIs for two chronic conditions, obesity and hypertension, to understand the diseases’ dynamic processes. With the title "Interactive Process Mining for Discovering Dynamic Risk Models in Chronic Diseases", the work was published as a chapter in the Springer book Interactive Process Mining in Healthcare.

8.1.3. Conference contributions

- **P7 - Z. Valero-Ramon** et al. “Overweight and Obesity: review of medical conditions and risk factors for Process Mining approach”. In: *Workshop on innovation on Information and Communication Technologies (ITACA-WIICT 2018)*. Ed. by

C.Fernandez-Llatas and M. Guillen. 2018, pp. 95–104 [132].

The work presented a bibliographic review to determine the risk factors, their direct or indirect association, and the associated prevalence with overweight and obesity, including comorbidities and socio-economic risk factors. This work was presented in oral communication at the International Workshop on Innovation Information and Communication Technologies (ITACA-WIICT 2018), with the title "Overweight and Obesity: review of medical conditions and risk factors for Process Mining approach".

- **P8 - Zoe Valero-Ramon** et al. “A dynamic behavioral approach to nutritional assessment using process mining”. In: *Proceedings of the 32nd IEEE International Symposium on Computer-Based Medical Systems*. Vol. 2019. 2019, pp. 398–404 [100].

  The use of Process Mining techniques to dynamically approach a medical condition and to find behavioural models with a reduced population at risk of malnutrition was presented in oral communication at the 32nd IEEE International Symposium on Computer-Based Medical Systems (IEEE CBMS2019), with the title "A dynamic behavioural approach to nutritional assessment using Process Mining".

- **P9 -** A. Martinez-Millana, **Z. Valero-Ramon**, C. Fernandez-Llatas, P. Garcia-Segovia and V. Traver Salcedo, "Evaluation of an App Based Questionnaire for the Nutritional Assessment in Elderly Housing". In: *Proceeding of the 32nd IEEE International Symposium on Computer-Based Medical Systems*, Vol. 2019, pp. 245-248 [96].

  The implementation and evaluation of an Android-based app for the screening and intervention program for a population at risk of malnutrition, to collect the needed data for approaching the medical condition using Process Mining techniques was presented at the 32nd IEEE International Symposium on Computer-Based Medical Systems (IEEE CBMS2019), with the title "Evaluation of an app based questionnaire for the nutritional assessment in elderly housing".

- **P10 -** L. Montandon, D. Kyriazis, **Z. Valero-Ramon**, C. Fernandez-Llatas and V. Traver, "CrowdHEALTH - Collective Wisdom Driving Public Health Policies." In: *Proceedings of the 32nd International Symposium on Computer-Based Medical Systems*. Vol. 2019, pp. 1-3 [135].

  How diseases can be approached from an integrated manner considering a large amount of healthcare information, including diseases, risk factors, and patient data, was presented in oral communication at the 32nd IEEE International Symposium on Computer-Based Medical Systems (IEEE CBMS2019), with the title "CrowdHEALTH-Collective Wisdom Driving Public Health Policies".

- **P11 - Valero-Ramon, Z** et al. “Dynamic Risk Models supporting personalised Diabetes healthcare with Process Mining”. In: *Diabetes Technology & Therapeutics*. Vol. 22. Mary Ann Liebert Inc 140 Huguenot Street 3rd FL New Rochelle NY 10801 USA. 2020, A112–A112 [109].

  The results from the application of Process Mining techniques to a concrete chronic non-communicable disease as Diabetes to obtain a Risk Model was presented in oral communication at the International Conference on Advanced Technologies & Treatments for Diabetes (ATTD) with the title "Dynamic Risk Models supporting personalised Diabetes Healthcare with Process Mining". The work presented the result of using Process Mining to discover and identify HbA1c (Glycated hemoglobin A1c) changes during a period to explore a Dynamic Risk Model for Diabetes.

8.2. Associated projects

The author of this doctoral thesis has participated in several IT projects applied to the healthcare sector. This experience has served to learn about the problems and needs of the healthcare sector. Moreover, the work performed during this work has been tested in different scenarios coinciding with various research projects funded by the European Commission, listed below:

- **NutriPro project - 2015**[3] (Screening programme of the nutritional status of older people using ICT, agreement no. 2014/1/17037714-SI2.695476). NutriPro project is a European project funded by the European Commission – Directorate General Health and Food Safety. It used a cohort study to examine MNA such a tool to screening malnutrition in the general elderly population in a nursing home scenario, as a perfect environment to spread the implementation and support routine of standardised screening tools of older patients.

  It developed both a free Android application, NutriPro, based on MNA test, to be deployed in tablets simply and easily with multilingual support to assure further scalability to the whole of Europe; and a nutritional intervention programme to correctly intervene in malnutrition detected cases through a set of customised actions.

  In this project, we used both data generated by the NutriPro App and data stored in the nursing home system to approach malnutrition as a dynamic disease. We used Process Mining techniques and clustering algorithms to model

[3] https://www.projectnutripro.eu/project-description/

weight and MNA results dynamically and to present the results as a graphical representation to healthcare professionals involved in the project. We also compared the results from the two different cohorts, corresponding to the baseline and at the end of the study, obtaining the overall process and pathway.

- **CrowdHEALTH project - 2017**[4] (Collective wisdom driving public health policies, Programme under Grant Agreement no. 727560). CrowdHEALTH is an international research project partially funded by the Horizon 2020 Programme of the European Commission that intends to integrate high volumes of health-related heterogeneous data from multiple sources to support policy-making decisions.

  In the framework of the present and regarding this project, it was offered a solution for the systematic detection of obesity and overweight, and the use of risk comorbidities due to obesity in detection. The obtained Dynamic Risk Models for obesity and related risk comorbidities were used to sensitise professionals of the National Health System and to promote the systematic detection of both, obesity and overweight, within the population.

- **AD-AUTONOMY project, 2018**[5] (Development of a training program for enhancing the autonomy of persons with Alzheimer, Erasmus + KA2: Strategic Partnerships - 2020-1-ES01-KA204-083089). AD-AUTONOMY project is an international project co-funded by the Erasmus+ Programme of the European Union that was launched with the main objective of increasing the competencies (attitudes, skills, knowledge) of Persons with Initial/Mild Alzheimer, Families and Caregivers, about how to improve Quality of Life of Persons with Initial/Mild Alzheimer through Autonomy through an innovative training program.

  In the framework of this work, Process Mining techniques were used to evaluate the impact of the training program thanks to the collection of users' data about their daily activities. The results also permitted the validation of the use of Process Mining techniques for behaviour modelling.

- **INFINITy project, 2019**[6] (INtelligent system to empower Functional Independence of people with mild cogNItive impairmenTs, activity 19342). INFINITy is an international innovation project funded by the EIT Health supported by the EIT, a body of the European Commission. It enables people with Mild Cognitive Impairments to preserve their functional abilities by empowering and extending their autonomy in indoor and outdoor surroundings.

  In the scheme of this document, Process Mining techniques were applied to analyse the behaviour of the project users, to extract behaviour patterns and to determine the impact of the use of the service for the autonomy and independence of the users. The results allowed to validate Process Mining techniques for behaviour and risk modelling in other contexts.

[4] https://www.crowdhealth.eu/
[5] http://www.adautonomy.eu/
[6] https://infinityeit.eu/

8.3. Contributions to the objectives

- **C1 - Study of clinical variables and risk factors associated with chronic conditions and included in the EHR that can be used to model underlying chronic diseases.**

  It contributes to objectives O1 and O2, and research questions RQ1 and RQ2.

  A review of the associated comorbidities and risk factors with obesity chronic disease was performed. The review considered a variety of medical conditions, and other socioeconomic factors and their association with overweight and obesity, as main as their availability in EHR, so they can be included and used for dynamic modelling of obesity, and as predictors for the evolution of the disease. The review confirmed the strong evidence of the following risk factors with obesity, hypertension, insulin resistance, dyslipidemia, coronary heart disease, gallbladder disease, respiratory disease, and knee osteoarthritis. The review also found other common risk factors associated with obesity such as poverty, and built environment, gender, heavy alcohol consumption and medication use. However, in most cases, comorbidities were considered as static, instead of dynamic factors that evolve with the patient. In consequence, the point of departure of this work was the idea that those comorbidities and risk factors could be used as inputs for Process Mining techniques and algorithms to obtain the evolution models for obesity or other chronic conditions, and consequently meaningful knowledge and information about overweight and obesity dynamic processes.

  This contribution was published in [132].

- **C2 - Analysis of the capability of Process Mining techniques to model a concrete medical condition.**

  Contributes to objective O1, and research question RQ1.

  EHR include a lot of valuable information about patients, collected during consultations, emergency episodes, or even laboratory results that might be used to enrich risk models from a dynamic perspective. In this regard, Process Mining techniques can be used to extract knowledge from EHR information to understand underlying healthcare processes. Process Mining techniques have been used in the healthcare domain mainly from a management perspective with promising results, but fewer works are modelling concrete diseases with this technique. Therefore, the proposed analysis was to use data coming from EHR or other health system and Process Mining techniques to model a concrete medical condition. This capability was firstly analysed with a well-characterised population combining the results for the nutritional status of older adults. Then, it was analysed chronic conditions associated risk factors to be used in the Process Mining approach.

  The results of this analysis can be found in Chapters 4 and 5. They have also been published in several of the contributions listed in the previous section, including the publications [100] and [109].

- **C3 - Develop a experiment strategy to incorporate the dynamic perspective of a chronic condition into a graphical and understandable representations.**

Contributes to objective O3, and research questions RQ2 and RQ3.

The business sector gives an interesting approach on how to formalise questions through the concept of Key Performance Indicators that provide a measurable value on how effectively a company is achieving key objectives, and therefore are used for the analysis and evaluation of processes. This philosophy has been applied in the context of chronic conditions to obtain human-understandable and contextualised Key Performance Indicators, named Interactive Process Indicators, in the form of enhanced views which help health professionals to perceive processes behind the disease. Health professionals can inquire what and how they want to know about the processes in the form of posed questions that are translated into dynamic perspectives of chronic conditions using Interactive Process Mining techniques to obtain Dynamic Risk Models.

The development of this strategy in the form of a formal methodology has been elaborated in Chapter 6. It has been also published in [29], and [134].

- **C4 - Development of Dynamic Risk Models for chronic conditions with real data through human-understandable graphical representations.**

  Contributes to objectives O1 and O2, and research questions RQ1 and RQ2, and RQ3.

  To illustrate the problem and solution to deal within this document, it was necessary to personalise the experiments to some specific chronic conditions to continue with the approach of the Dynamic Risk Models using Process Mining techniques through a human-understandable graphical representation. Based on the information available in the EHR and the risk factors associated with chronic conditions, there were discovered three risk models for three of the most prevalent chronic conditions, obesity, hypertension, and hyperglycemia, using the results from the BMI, fasting plasma glucose, and blood pressure respectively. As a result, there were produced three understandable representations of the real flows for the three chronic diseases incorporating the dynamic and behavioural perspectives.

  The resultant models are presented and described in the Chapters 5 and 6. They have also been published in several contributions, highlighting [29], [131], and [127].

- **C5 - Formalisation of a methodology to dynamically model chronic conditions using Interactive Process Mining techniques.**

  Contributes to objective O3, and research question RQ3.

  A formal methodology has been implemented for discovering Dynamic Risk Models correlated with chronic conditions. This methodology integrates not only the steps followed to discover the different risk models developed during the different experiments performed in this work but also the needed methods, techniques, and tools to perform each step. It is an interactive and question based methodology for deploying Dynamic Risk Models for chronic conditions using Process Mining. The methodology introduces a six-step procedure designed to support medical and Process Mining experts in the understanding of chronic disease's processes, bringing them a straightforward guideline of how to apply Interactive Process Mining to the analysis of chronic disease's underlying processes. The complete methodology has been applied for the obesity chronic disease risk model and the dynamic relationship between obesity and hypertension.

The methodology was published in [127]. The results from the application of the methodology are presented in Chapter 6.

Table 8.1 presents the relationship between the different contents presented in this document, and the research questions, objectives, contributions and publications they cover.

TABLE 8.1: Relationship among the research questions, objectives, publications and contributions.

Chapter #	**Contents**	**Research Questions**	**Objectives**	**Contributions**	**Publications**
1	Introduction: motivation, hypothesis, and objectives				
2	Background: chronic diseases, risk factors, and risk models			C1, C2	P4, P7, P10
3	Materials and methods				P9
4	Modelling a medical condition with Process Mining	RQ1	O1, O2	C2	P1, P3, P8
5	Dynamic Risk Models with PM applied to chronic conditions	RQ1, RQ2	O2	C1, C2, C4, C5	P2, P11
6	Towards a formal methodology for obtaining Dynamic Risk Models	RQ2, RQ3	O2, O3	C3, C4, C6	P5, P6
7	Conclusions				P2, P9
8	Main original contributions				

Bibliography

[1] World Health Organization. *Noncommunicable diseases*. May 2020. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.

[2] Verena Struckmann et al. “Caring for people with multiple Chronic conditions in Europe”. In: *EuroHealth* 20.3 (2014), pp. 35–40.

[3] Paul Brennan et al. “Chronic disease research in Europe and the need for integrated population cohorts”. In: *European journal of epidemiology* 32.9 (2017), pp. 741–749.

[4] Kathleen Strong et al. “Preventing chronic disease: a priority for global health”. In: *International Journal of Epidemiology* 35.2 (2006), pp. 492–494.

[5] Harry Campbell et al. “Integrated care pathways”. In: *Bmj* 316.7125 (1998), pp. 133–137.

[6] Jennifer Dixon and Martin Bardsley. *Predictive risk modelling using routine data: underexploited potential to benefit patients*. 2012.

[7] Paulo Cortez and Mark J Embrechts. “Using sensitivity analysis and visualization techniques to open black box data mining models”. In: *Information Sciences* 225 (2013), pp. 1–17.

[8] **Valero-Ramon, Zoe** et al. *Towards Perceptual Spaces for Empowering Ergonomy in Workplaces by Using Interactive Process Mining*. 2019.

[9] Carlos Fernandez-Llatas. *Interactive Process Mining in Healthcare*. 2020.

[10] Maikel L Van Eck et al. “PM 2: a process mining project methodology”. In: *International Conference on Advanced Information Systems Engineering*. Springer. 2015, pp. 297–313.

[11] Amy B Bernstein. *Health care in America: Trends in utilization*. Center for Disease Control and Prevention, National Center for Health Statistics, 2004.

[12] Wullianallur Raghupathi and Viju Raghupathi. “An empirical study of chronic diseases in the United States: a visual analytics approach to public health”. In: *International journal of environmental research and public health* 15.3 (2018), p. 431.

[13] GBD 2015 Risk Factors Collaborators et al. “Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015”. In: *Lancet (London, England)* 388.10053 (2016), p. 1659.

[14] Jorge Gómez, Byron Oviedo, and Emilio Zhuma. "Patient monitoring system based on internet of things". In: *Procedia Computer Science* 83 (2016), pp. 90–97.

[15] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[16] Alison Harvey et al. "The future of technologies for personalised medicine". In: *New biotechnology* 29.6 (2012), pp. 625–633.

[17] J Larry Jameson and Dan L Longo. "Precision medicine—personalized, problematic, and promising". In: *Obstetrical & gynecological survey* 70.10 (2015), pp. 612–614.

[18] Alice S Whittemore. "Evaluating health risk models". In: *Statistics in medicine* 29.23 (2010), pp. 2438–2452.

[19] Laura E Panattoni et al. "Predictive risk modelling in health: options for New Zealand and Australia". In: *Australian Health Review* 35.1 (2011), pp. 45–51.

[20] Juan F Orueta et al. "Predictive risk modelling in the Spanish population: a cross-sectional study". In: *BMC health services research* 13.1 (2013), pp. 1–9.

[21] Vincenzo Lagani et al. "A systematic review of predictive risk models for diabetes complications based on large scale clinical studies". In: *Journal of Diabetes and its Complications* 27.4 (2013), pp. 407–413. ISSN: 1056-8727.

[22] Hadi Kharrazi et al. "Assessing the Impact of Body Mass Index Information on the Performance of Risk Adjustment Models in Predicting Healthcare Costs and Utilization". In: *Medical care* 56.12 (2018), p. 1042.

[23] Jennifer Anne Cooper et al. "The use of electronic healthcare records for colorectal cancer screening referral decisions and risk prediction model development". In: *BMC gastroenterology* 20.1 (2020), pp. 1–16.

[24] Anja Schienkiewitz, Gert BM Mensink, and Christa Scheidt-Nave. "Comorbidity of overweight and obesity in a nationally representative sample of German adults aged 18-79 years". In: *BMC Public Health* 12.1 (2012), p. 658.

[25] Aviva Must et al. "The disease burden associated with overweight and obesity". In: *Jama* 282.16 (1999), pp. 1523–1529.

[26] Etienne Audureau, Jacques Pouchot, and Joël Coste. "Gender-related differential effects of obesity on health-related quality of life via obesity-related Comorbidities: a mediation analysis of a French Nationwide survey". In: *Circulation: Cardiovascular Quality and Outcomes* 9.3 (2016), pp. 246–256.

[27] James E Everhart et al. "Duration of obesity increases the incidence of NIDDM". In: *Diabetes* 41.2 (1992), pp. 235–240.

[28] S Goya Wannamethee, A Gerald Shaper, and Mary Walker. "Overweight and obesity and weight change in middle aged men: impact on cardiovascular disease and diabetes". In: *Journal of Epidemiology & Community Health* 59.2 (2005), pp. 134–139.

[29] **Valero-Ramon, Zoe** et al. "Dynamic Models Supporting Personalised Chronic Disease Management through Healthcare Sensors with Interactive Process Mining". In: *Sensors* 20.18 (2020), p. 5330.

[30] Ben Christopher Reynolds et al. “Association of Time-Varying Blood Pressure With Chronic Kidney Disease Progression in Children”. In: *JAMA Network Open* 3.2 (2020), e1921213–e1921213. ISSN: 2574-3805.

[31] Yuval Shahar. “A framework for knowledge-based temporal abstraction”. In: *Artificial intelligence* 90.1-2 (1997), pp. 79–133.

[32] Kalia Orphanou, Athena Stassopoulou, and Elpida Keravnou. “DBN-extended: a dynamic Bayesian network model extended with temporal abstractions for coronary heart disease prognosis”. In: *IEEE journal of biomedical and health informatics* 20.3 (2016), pp. 944–952.

[33] Stefano Concaro et al. “Temporal data mining for the assessment of the costs related to diabetes mellitus pharmacological treatment”. In: *AMIA Annual Symposium Proceedings*. Vol. 2009. American Medical Informatics Association. 2009, p. 119.

[34] Arianna Dagliati et al. “Temporal data mining and process mining techniques to identify cardiovascular risk-associated clinical pathways in Type 2 diabetes patients”. In: *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2014, pp. 240–243.

[35] Mira Balaban, David Boaz, and Yuval Shahar. “Applying temporal abstraction in medical information systems”. In: *Ann Math Comput Teleinform* 1.1 (2003), pp. 56–64.

[36] Carolyn McGregor, Christina Catley, and Andrew James. “A process mining driven framework for clinical guideline improvement in critical care”. In: *Proceedings of the Learning from Medical Data Streams Workshop. Bled, Slovenia (July 2011)*. 2011.

[37] Carlos Fernandez-Llatas et al. “Using Process Mining for Automatic Support of Clinical Pathways Design”. en. In: *Data Mining in Clinical Medicine*. Ed. by Carlos Fernández-Llatas and Juan Miguel García-Gómez. Methods in Molecular Biology 1246. Springer New York, Jan. 2015, pp. 79–88. ISBN: 978-1-4939-1984-0 978-1-4939-1985-7.

[38] Donna Spruijt-Metz et al. “Building new computational models to support health behavior change and maintenance: new opportunities in behavioral research”. In: *Translational behavioral medicine* 5.3 (2015), pp. 335–346.

[39] Francis S. Collins and Harold Varmus. “A New Initiative on Precision Medicine”. In: *New England Journal of Medicine* 372.9 (Feb. 2015), pp. 793–795. ISSN: 0028-4793.

[40] David A Chambers, W Gregory Feero, and Muin J Khoury. “Convergence of implementation science, precision medicine, and the learning health care system: a new model for biomedical research”. In: *Jama* 315.18 (2016), pp. 1941–1942.

[41] David L Sackett et al. *Evidence based medicine: what it is and what it isn't*. 1996.

[42] Margaret A Hamburg and Francis S Collins. “The path to personalized medicine”. In: *New England Journal of Medicine* 363.4 (2010), pp. 301–304.

[43] Volker Amelung et al. *Handbook integrated care*. Springer, 2017.

[44] Sabyasachi Dash et al. "Big data in healthcare: management, analysis and future prospects". In: *Journal of Big Data* 6.1 (2019), pp. 1–25.

[45] Phillip C-Y Sheu et al. *Semantic Computing*. John Wiley & Sons, 2011.

[46] Dharmendra S Modha et al. "Cognitive computing". In: *Communications of the ACM* 54.8 (2011), pp. 62–71.

[47] Amit Sheth. "Internet of things to smart iot through semantic, cognitive, and perceptual computing". In: *IEEE Intelligent Systems* 31.2 (2016), pp. 108–112.

[48] Carlos Fernández-Llatas et al. "Applying evidence-based medicine in telehealth: An interactive pattern recognition approximation". In: *International journal of environmental research and public health* 10.11 (2013), pp. 5671–5682.

[49] Mark W Craven and Jude W Shavlik. "Using neural networks for data mining". In: *Future generation computer systems* 13.2-3 (1997), pp. 211–229.

[50] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[51] Lawrence Rabiner and Biinghwang Juang. "An introduction to hidden Markov models". In: *ieee assp magazine* 3.1 (1986), pp. 4–16.

[52] Wil Van Der Aalst. *Process Mining.Data science in action*. Springer, 2016.

[53] Peter E Hart, David G Stork, and Richard O Duda. *Pattern classification*. Wiley Hoboken, 2000.

[54] Danton S Char, Nigam H Shah, and David Magnus. "Implementing machine learning in health care—addressing ethical challenges". In: *The New England journal of medicine* 378.11 (2018), p. 981.

[55] Gema Ibanez-Sanchez et al. "Toward Value-Based Healthcare through Interactive Process Mining in Emergency Rooms: The Stroke Case". In: *International Journal of Environmental Research and Public Health* 16.10 (2019), p. 1783.

[56] Octavio Loyola-Gonzalez. "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view". In: *IEEE Access* 7 (2019), pp. 154096–154113.

[57] Pradeep Chowriappa, Sumeet Dua, and Yavor Todorov. "Introduction to machine learning in healthcare informatics". In: *Machine Learning in Healthcare Informatics*. Springer, 2014, pp. 1–23.

[58] Meherwar Fatima, Maruf Pasha, et al. "Survey of machine learning algorithms for disease diagnostic". In: *Journal of Intelligent Learning Systems and Applications* 9.01 (2017), p. 1.

[59] Carlos Fernández-Llatas and Juan Miguel García-Gómez. *Data mining in clinical medicine*. Springer, 2015.

[60] Igor Kononenko. "Machine learning for medical diagnosis: history, state of the art and perspective". In: *Artificial Intelligence in medicine* 23.1 (2001), pp. 89–109.

[61] Phil Gooch and Abdul Roudsari. "Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems". In: *Journal of the American Medical Informatics Association* 18.6 (2011), pp. 738–748.

[62] Wil Van Der Aalst. "Process mining: Overview and opportunities". In: *ACM Transactions on Management Information Systems (TMIS)* 3.2 (2012), pp. 1–17.

[63] Boudewijn F Van Dongen, AK Alves De Medeiros, and Lijie Wen. "Process mining: Overview and outlook of petri net discovery algorithms". In: *transactions on petri nets and other models of concurrency II* (2009), pp. 225–242.

[64] Tugba Gurgen Erdogan and Ayca Tarhan. "Systematic mapping of process mining studies in healthcare". In: *IEEE Access* 6 (2018), pp. 24543–24567.

[65] Eric Rojas et al. "Process mining in healthcare: A literature review". In: *Journal of biomedical informatics* 61 (2016), pp. 224–236.

[66] Josep Carmona et al. *Conformance checking*. Springer, 2018.

[67] Minseok Song, Christian W Günther, and Wil MP Van der Aalst. "Trace clustering in process mining". In: *International Conference on Business Process Management*. Springer. 2008, pp. 109–120.

[68] Carlos Fernandez-Llatas et al. "Process Mining in Healthcare". In: *Interactive Process Mining in Healthcare*. Springer, 2021, pp. 41–52.

[69] Juan J Lull et al. "Interactive Process Mining in IoT and Human Behaviour Modelling". In: *Interactive Process Mining in Healthcare*. Springer, 2021, pp. 217–231.

[70] Onur Dogan et al. "Individual behavior modeling with sensors using process mining". In: *Electronics* 8.7 (2019), p. 766.

[71] A Hoerbst and E Ammenwerth. "Electronic health records". In: *Methods Inf Med* 49.4 (2010), pp. 320–336.

[72] AJMM Weijters and Wil MP van der Aalst. "Process mining: discovering workflow models from event-based data". In: *Belgium-Netherlands Conf. on Artificial Intelligence*. Citeseer. 2001.

[73] Cleiton dos Santos Garcia et al. "Process mining techniques and applications–A systematic mapping study". In: *Expert Systems with Applications* 133 (2019), pp. 260–295.

[74] Álvaro Rebuge and Diogo R Ferreira. "Business process analysis in healthcare environments: A methodology based on process mining". In: *Information systems* 37.2 (2012), pp. 99–116.

[75] Sooyoung Yoo et al. "Assessment of hospital processes using a process mining technique: Outpatient process analysis at a tertiary hospital". In: *International journal of medical informatics* 88 (2016), pp. 34–43.

[76] Ronny S Mans et al. "Application of process mining in healthcare–a case study in a dutch hospital". In: *International joint conference on biomedical engineering systems and technologies*. Springer. 2008, pp. 425–438.

[77] Ronny Mans et al. "Process mining techniques: an application to stroke care". In: *MIE*. Vol. 136. 2008, pp. 573–578.

[78] Carlos Fernández-Llatas et al. "Process mining for individualized behavior modeling using wireless tracking in nursing homes". In: *Sensors* 13.11 (2013), pp. 15434–15451.

[79] Marco Cameranesi et al. "Extraction of User Daily Behavior from Home Sensors through Process Discovery". In: *IEEE Internet of Things Journal* 7.9 (2020), pp. 8440–8450.

[80] Berndt Müller, Joachim Reinhardt, and Michael T Strickland. *Neural networks: an introduction*. Springer Science & Business Media, 1995.

[81] Sean R Eddy. "Hidden markov models". In: *Current opinion in structural biology* 6.3 (1996), pp. 361–365.

[82] Tania Conca et al. "Multidisciplinary Collaboration in the Treatment of Patients With Type 2 Diabetes in Primary Care: Analysis Using Process Mining". In: *Journal of medical Internet research* 20.4 (2018).

[83] Carlos Fernández-Llatas et al. "Activity-based process mining for clinical pathways computer aided design". In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE. 2010, pp. 6178–6181.

[84] Carlos Fernandez-Llatas et al. "Timed parallel automaton: A mathematical tool for defining highly expressive formal workflows". In: *2011 Fifth Asia Modelling Symposium*. IEEE. 2011, pp. 56–61.

[85] Michael E Porter and Elizabeth Olmsted Teisberg. *Redefining health care: creating value-based competition on results*. Harvard Business Press, 2006.

[86] Daniela Luengo and Marcos Sepúlveda. "Applying clustering in process mining to find different versions of a business process that changes over time". In: *International Conference on Business Process Management*. Springer. 2011, pp. 153–158.

[87] Massimiliano De Leoni, Wil MP van der Aalst, and Marcus Dees. "A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs". In: *Information Systems* 56 (2016), pp. 235–257.

[88] Gonzalo Navarro. "A guided tour to approximate string matching". In: *ACM computing surveys (CSUR)* 33.1 (2001), pp. 31–88.

[89] Richard W Hamming. "Error detecting and error correcting codes". In: *The Bell system technical journal* 29.2 (1950), pp. 147–160.

[90] RP Jagadeesh Chandra Bose and Wil MP Van der Aalst. "Context aware trace clustering: Towards improving process mining results". In: *Proceedings of the 2009 SIAM International Conference on Data Mining*. SIAM. 2009, pp. 401–412.

[91] RP Jagadeesh Chandra Bose and Wil MP van der Aalst. "Process diagnostics using trace alignment: opportunities, issues, and challenges". In: *Information Systems* 37.2 (2012), pp. 117–141.

[92] Carol Evans. "Malnutrition in the elderly: a multifactorial failure to thrive". In: *The Permanente Journal* 9.3 (2005), p. 38.

[93] Nestlé Nutrition Institute. *MNA®-SF - MNA® Elderly*. Feb. 2021. URL: https://www.mna-elderly.com/forms/mna_guide_spanish_sf.pdf.

[94] Laurence Z Rubenstein et al. "Screening for undernutrition in geriatric practice: developing the short-form mini-nutritional assessment (MNA-SF)". In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 56.6 (2001), pp. M366–M372.

[95] M_J Kaiser et al. "Validation of the Mini Nutritional Assessment Short-Form (MNA®-SF): A practical tool for identification of nutritional status". In: *JNHA-The Journal of Nutrition, Health and Aging* 13.9 (2009), p. 782.

[96] Antonio Martinez-Millana et al. "Evaluation of an App Based Questionnaire for the Nutritional Assessment in Elderly Housing". In: *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2019, pp. 245–248.

[97] World Health Organization. *Obesity and overweight*. Mar. 2020. URL: https://www.who.int/en/news-room/fact-sheets/detail/obesity-and-overweight.

[98] Robert J Kuczmarski and Katherine M Flegal. "Criteria for definition of overweight in transition: background and recommendations for the United States". In: *The American journal of clinical nutrition* 72.5 (2000), pp. 1074–1081.

[99] Gulistan Bahat et al. "Which body mass index (BMI) is better in the elderly for functional status?" In: *Archives of gerontology and geriatrics* 54.1 (2012), pp. 78–81.

[100] **Zoe Valero-Ramon** et al. "A dynamic behavioral approach to nutritional assessment using process mining". In: *Proceedings of the 32nd IEEE International Symposium on Computer-Based Medical Systems*. Vol. 2019. 2019, pp. 398–404.

[101] World Health Organization. *Hypertension*. Jan. 2020. URL: https://www.who.int/news-room/fact-sheets/detail/diabetes.

[102] World Health Organization et al. "Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation". In: *World Health Organization* (2006).

[103] Lydia E Makaroff. "The need for international consensus on prediabetes". In: *The lancet Diabetes & endocrinology* 5.1 (2017), pp. 5–7.

[104] Chee W Chia, Josephine M Egan, and Luigi Ferrucci. "Age-related changes in glucose metabolism, hyperglycemia, and cardiovascular risk". In: *Circulation research* 123.7 (2018), pp. 886–904.

[105] Ivy Shiue, Peter McMeekin, and Christopher Price. "Retrospective observational study of emergency admission, readmission and the 'weekend effect'". In: *BMJ open* 7.3 (2017), e012493.

[106] Ralph A DeFronzo. "Glucose intolerance and aging". In: *Diabetes care* 4.4 (1981), pp. 493–501.

[107] Hiroshi Shimokata et al. "Age as independent determinant of glucose tolerance". In: *Diabetes* 40.1 (1991), pp. 44–51.

[108] DC Muller et al. "Insulin response during the oral glucose tolerance test: the role of age, sex, body fat and the pattern of fat distribution". In: *Aging Clinical and Experimental Research* 8.1 (1996), pp. 13–21.

[109] **Valero-Ramon, Z** et al. "Dynamic Risk Models supporting personalised Diabetes healthcare with Process Mining". In: *Diabetes Technology & Therapeutics*. Vol. 22. Mary Ann Liebert Inc 140 Huguenot Street 3rd FL New Rochelle NY 10801 USA. 2020, A112–A112.

[110] World Health Organization. *Hypertension*. Mar. 2020. URL: https://www.who.int/news-room/fact-sheets/detail/hypertension.

[111] Joint National Committee on Detection, Treatment of High Blood Pressure, and National High Blood Pressure Education Program. Coordinating Committee. *Report of the joint national committee on detection, evaluation, and treatment of high blood pressure*. National Heart, Lung, and Blood Institute, National High Blood Pressure, 1995.

[112] American Heart Association. *Understanding Blood Pressure Readings*. May 2020. URL: https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.

[113] Gianfranco Parati, Andrea Faini, and Mariaconsuelo Valentini. "Blood pressure variability: its measurement and significance in hypertension". In: *Current hypertension reports* 8.3 (2006), pp. 199–204.

[114] Giuseppe Mancia et al. "Blood pressure variability". In: *Handbook of hypertension* 17 (1997).

[115] Giuseppe Mancia. "Short-and long-term blood pressure variability: present and future". In: *Hypertension* 60.2 (2012), pp. 512–517.

[116] Gianfranco Parati et al. "Assessment and management of blood-pressure variability". In: *Nature Reviews Cardiology* 10.3 (2013), p. 143.

[117] Roberto Nuño et al. "Integrated care for chronic conditions: the contribution of the ICCC Framework". In: *Health Policy* 105.1 (2012), pp. 55–64.

[118] Ronny S Mans, Wil MP Van der Aalst, and Rob JB Vanwersch. *Process mining in healthcare: evaluating and exploiting operational healthcare processes*. Springer, 2015.

[119] Angelo Croatti et al. "On the integration of agents and digital twins in healthcare". In: *Journal of Medical Systems* 44.9 (2020), pp. 1–8.

[120] Enrique Vidal et al. "Interactive pattern recognition". In: *International Workshop on Machine Learning for Multimodal Interaction*. Springer. 2007, pp. 60–71.

[121] Carlos Fernandez-Llatas. "Interactive Process Mining in Practice: Interactive Process Indicators". In: *Interactive Process Mining in Healthcare*. Springer, 2021, pp. 141–162.

[122] Mohamed A Ghazal, Osman Ibrahim, and Mostafa A Salama. "Educational process mining: A systematic literature review". In: *2017 European Conference on Electrical Engineering and Computer Science (EECS)*. IEEE. 2017, pp. 198–203.

[123] Melike Bozkaya, Joost Gabriels, and Jan Martijn van der Werf. "Process diagnostics: a method based on process mining". In: *2009 International Conference on Information, Process, and Knowledge Management*. IEEE. 2009, pp. 22–27.

[124] Wil Van Der Aalst et al. "Process mining manifesto". In: *International conference on business process management*. Springer. 2011, pp. 169–194.

[125] Eric Rojas et al. "Question-Driven Methodology for Analyzing Emergency Room Processes Using Process Mining". In: *Applied Sciences* 7.3 (2017), p. 302.

[126] Carlos Fernandez-Llatas. "Bringing Interactive Process Mining to Health Professionals: Interactive Data Rodeos". In: *Interactive Process Mining in Healthcare*. Springer, 2021, pp. 119–140.

[127] **Valero-Ramon, Zoe** and Carlos Fernandez-Llatas. "Interactive Process Mining for Discovering Dynamic Risk Models in Chronic Diseases". In: *Interactive Process Mining in Healthcare*. Springer, 2021, pp. 243–266.

[128] Luke M Funk, Sally A Jolles, and Corrine I Voils. "Obesity as a disease: has the AMA resolution had an impact on how physicians view obesity?" In: *Surgery for Obesity and Related Diseases* 12.7 (2016), pp. 1431–1435.

[129] Carlotta Pozza and Andrea M Isidori. "What's behind the obesity epidemic". In: *Imaging in bariatric surgery*. Springer, 2018, pp. 1–8.

[130] World Health Organization et al. *Global health risks: mortality and burden of disease attributable to selected major risks*. World Health Organization, 2009.

[131] **Valero-Ramon, Zoe** et al. "Interactive Process Indicators for Obesity Modelling Using Process Mining". In: *Advanced Computational Intelligence in Healthcare-7*. Springer, 2020, pp. 45–64.

[132] **Z. Valero-Ramon** et al. "Overweight and Obesity: review of medical conditions and risk factors for Process Mining approach". In: *Workshop on innovation on Information and Communication Technologies (ITACA-WIICT 2018)*. Ed. by C.Fernandez-Llatas and M. Guillen. 2018, pp. 95–104.

[133] Carlos Fernandez-Llatas et al. "Process mining methodology for health process tracking using real-time indoor location systems". In: *Sensors* 15.12 (2015), pp. 29821–29840.

[134] Roberto Gatta et al. "What Role Can Process Mining Play in Recurrent Clinical Guidelines Issues? A Position Paper". In: *International Journal of Environmental Research and Public Health* 17.18 (2020), p. 6616.

[135] Lydia Montandon et al. "CrowdHEALTH-Collective wisdom driving public health policies". In: *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2019, pp. 1–3.