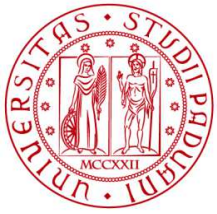
SCUOLA DI DOTTORATO
IN INGEGNERIA
DELL'INFORMAZIONE



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento: Ingegneria dell'Informazione

Scuola di dottorato di ricerca in: Ingegneria dell'Informazione

Indirizzo: Bioingegneria

Ciclo: XXVI

MULTI-LEVEL MODELING AND COMPUTATIONAL APPROACHES TO INVESTIGATE LONG-TERM DIABETES COMPLICATIONS

Direttore della Scuola: Ch.mo Prof. Matteo Bertocco

Supervisore: Ing. Barbara Di Camillo

Dottorando: Emanuele Trifoglio

Sommario

Il diabete mellito rappresenta una delle patologie più diffuse nel mondo e si stima che la sua incidenza aumenterà del 50 % nell'arco di 15 anni, passando da 250 milioni a quasi 400 milioni di malati nel 2025. La patologia comporta l'insorgenza di devastanti complicanze croniche, tra cui disturbi legati al danneggiamento dei vasi sanguigni sia a livello macro-vascolare – come coronopatia, infarto, insufficienza cardiaca, angina pectoris, ictus – che micro-vascolare, con conseguente danno a carico dei reni (nefropatia) e degli occhi (retinopatia). La patologia diabetica ha un'enorme impatto sia in termini di qualità di vita dei pazienti, sia a livello economico, in quanto si stima che più del 10 % dei costi dell'assistenza sanitaria di tutta l'Europa siano imputabili alla cura del diabete. Per questo motivo, nuovi mezzi che permettano di prevenire l'insorgere e il progredire della malattia e delle sue complicanze sono assolutamente necessari.

L'obiettivo del seguente lavoro di tesi è quello di proporre nuovi metodi computazionali per lo studio delle complicanze del diabete in un ambito di modellistica multi-livello.

Il diabete mellito è una malattia fortemente multifattoriale, nella quale molteplici fattori di rischio di diversa natura (genetica e ambientale) concorrono a provocarne l'insorgenza e lo sviluppo. I meccanismi fisiologici che sottendono allo scatenarsi e al progredire della patologia sono ancora per la maggior parte sconosciuti.

Data la natura multifattoriale del diabete, lo studio delle complicanze si presta ad essere affrontato con un approccio multi-livello. Lo schema generale di una malattia multifattoriale, come il diabete, prevede l'azione combinata di 3 elementi chiave sullo stato patologico (l'outcome) del paziente: *i)* il fenotipo, ovvero l'insieme di tutte le variabili metaboliche, antropometriche e ambientali caratteristiche del paziente, *ii)* il genotipo, ovvero la sequenza DNA del paziente, *iii)* il trattamento, ovvero l'insieme di interventi esterni effettuati sul paziente, come terapie ed utilizzo di farmaci. Queste 3 variabili sono interconnesse tramite interazioni e concorrono tutte insieme a determinare l'outcome del paziente.

L’approccio multi-livello consente di scomporre il problema completo in sottoproblemi, focalizzando l’attenzione di volta in volta solo su un sottoinsieme di variabili e di interazioni, a seconda del livello di informazione contenuto nei dati a disposizione.
Nel seguente lavoro, vengono considerati 3 principali livelli di studio delle complicanze diabetiche, e, per ognuno dei 3 ambiti, vengono proposti nuovi metodi sviluppati durante il periodo di dottorato.
I 3 livelli di studio trattati sono: *i)* modellizzazione dell’effetto del genotipo sull’outcome, *ii)* modellizzazione dell’effetto combinato di fenotipo e trattamento sulla progressione dell’outcome, *iii)* modellizzazione dell’azione del trattamento sul fenotipo.
Il primo livello di studio si propone di studiare le complicanze diabetiche da un punto di vista statico, ovvero senza considerare l’evolversi e il progredire di tali complicanze nel tempo, ed ha come obiettivo quello di identificare i principali biomarcatori genetici che consentano di predire lo stato di malattia dei pazienti, e di stratificare i pazienti in base al rischio di sviluppare o meno la malattia. I Genome Wide Association Studies (GWAS), sono studi di associazione volti a identificare gli SNPs che, da soli o in combinazioni con altri SNPs, consentono di spiegare le differenze che si osservano in un determinato outcome (a presenza o meno di una patologia) tra casi (soggetti malati) e controlli (soggetti sani) in una popolazione di studio. Diversi metodi di selezione univariata e multivariata sono presenti in letteratura per l’identificazione di marcatori genetici da studi GWAS. In questo ambito, è stato sviluppato un nuovo metodo per la selezione multivariata di biomarcatori genetici e per la classificazione di soggetti a partire da dati di SNPs di studi GWAS, basato sui classificatori di Bayes e arricchito da 3 principali componenti: *i)* una predizione ottenuta da un insieme di classificatori di Bayes, utilizzando una strategia basata sul bootstrap, *ii)* un nuovo metodo per ordinare e selezionare gli attributi selezionati da ogni classificatore, *iii)* una procedura, bastata sulle permutazioni, per selezionare i biomarcatori significativi, sulla base della loro utilità marginale nel processo di classificazione. Il metodo è stato validato sui dati genome-wide del Wellcome Trust Case-Control Consortium, (WTCCC) relativi a diabetici di tipo 1 e le sue performance confrontate con gli algoritmi rappresentanti lo stato dell’arte in letteratura per studi di associazione genetica, in particolare un classificatore di Bayes e un algoritmo di regressione logistica penalizzata (HyperLASSO).
Il secondo livello di studio riguarda l’analisi dinamica delle complicanze, nella quale interviene anche la variabile tempo come fattore chiave. In quest’ottica, si vuole

modellizzare l'insorgere e la progressione temporale delle principali complicanze legate al diabete utilizzando l'informazione fenotipica e terapeutica, con l'obiettivo di stimare la probabilità che il paziente diabetico possa o meno sviluppare una certa complicanza, ottimizzando quindi i trial clinici ed evitando esami costosi e invasivi. In letteratura, sono presenti diversi modelli delle complicanze di diabete, ma nessuno è in grado di integrare in maniera flessibile le diverse conoscenze –omiche (proteomica, metabolomica, genomica) ad un livello clinico macroscopico. I principali modelli presenti in letteratura sono infatti basati sui modelli di Markov (detti anche modelli si transizione di stato) e utilizzano l'informazione fenotipica senza la possibilità di integrare facilmente informazioni aggiuntive. In questo ambito di studio, viene proposto un nuovo modello *in-silico* delle complicanze cardiovascolari e renali del diabete, che propone come aspetto innovativo l'utilizzo delle reti dinamiche bayesiane (Dynamic Bayesian Networks, DBNs) per modellizzare le interazioni tra le variabili. Rispetto ai modelli di Markov, che richiedono tanti nodi quante sono le possibili combinazioni degli stati delle variabili, le DBN hanno il vantaggio di rappresentare ogni variabile tramite un singolo nodo e permettono quindi una maggiore facilità nella gestione della struttura e nell'integrazione di eventuale informazione aggiuntiva. Il modello è stato costruito utilizzando i dati del Diabetes Control and Complications Trial (DCCT), un trial clinico randomizzato condotto con lo scopo di confrontare gli effetti della terapia intensiva rispetto a quelli della terapia convenzionale sullo sviluppo delle complicanze vascolari e neurologiche a lungo termine. Il modello sviluppato, è in grado di predire la progressione delle complicanze diabetiche trattate con un'accuratezza superiore al 95% a livello di popolazione. Il modello si presta quindi ad essere utilizzato come tool di supporto nel processo di decisione terapeutica da parte dei clinici e, in quest'ottica, sta portando alla realizzazione di un'interfaccia web. La struttura flessibile del modello inoltre consentirà di integrare facilmente l'informazione genotipica, con l'obiettivo futuro di migliorare le prestazioni a livello di predizione.

Il terzo ed ultimo livello di studio considerato è lo studio dell'azione di uno specifico farmaco su un particolare fenotipo, con l'obiettivo finale di sviluppare metodologie che consentano di personalizzare i farmaci, adattandoli alla specifica risposta dell'individuo. Nell'ambito specifico delle complicanze cardiovascolari del diabete, una delle terapie più diffuse è quella del trattamento con aspirina per la prevenzione di eventi avversi nei pazienti ad alto rischio. L'aspirina deve la sua azione preventiva alla capacità di inibire

un enzima chiave (la prostaglandina-endoperossido sintase PTGS-1, conosciuta anche come cicloossigenasi COX-1) nella cascata che porta alla formazione di trombossano B2 (TxB_2), il principale responsabile dell'aggregazione piastrinica nel sangue e della conseguente formazioni di trombi. È noto, da letteratura, come i pazienti diabetici rispondano in maniera differente alla terapia con aspirina rispetto ai soggetti sani, evidenziando una risposta ridotta al farmaco, tanto da portare in ambito clinico alla coniazione del termine 'aspirino-resistenza'. Data la mancanza di una trattazione matematica del fenomeno in letteratura, si è deciso di studiare il problema utilizzando un approccio modellistico di farmacodinamica, con un intento. Utilizzando informazioni biologiche ricavate da letteratura, si è sviluppato un modello, in parte compartimentale e in parte distribuito, che descrive: *i)* la cinetica dell'enzima COX-1 a partire dalla sua produzione all'interno dei megacariociti del midollo osseo fino a giungere nelle piastrine del sangue, *ii)* la farmacocinetica e la farmacodinamica dell'aspirina, ovvero la distribuzione del farmaco nel corpo e la sua interazione con l'enzima COX-1. Il modello è stato testato su dati sperimentali relativi al recupero di trombossano B2 sierico dopo la sospensione di aspirina in pazienti sani. Sono stati infine discussi meccanismi potenzialmente candidati a spiegare il fenomeno dell'aspirino-resistenza in pazienti diabetici.

Summary

Diabetes mellitus is a lifelong, incapacitating disease affecting multiple organs. Worldwide prevalence figures estimate that there are 250 million diabetic patients today and that this number will increase by 50% by 2025. The disease is associated with devastating chronic complications including coronary heart disease, stroke and peripheral vascular disease (macrovascular disease) as well as microvascular disorders, leading to damage of kidneys (nephropathy) and eyes (retinopathy). These complications impose an immense burden on the quality of life of the patients and account for more than 10% of health care costs in Europe. Therefore, novel means to prevent the onset and the progression of these devastating diabetic complications are needed.

The aim of the work presented in this thesis is to propose novel computational methods to study diabetes complications with a multi-level approach.

Diabetes mellitus is a strongly multifactorial disease, and several risks factors (such as genetic, and environmental factors) are combined together in a complex trait, leading to the onset of the disease.

Physiological mechanisms that underlie the disease and the onset and progression of the different complications are still mostly unknown.

Given the complex nature of diabetes, the study of the complications can be faced with a multi-level modeling approach. In the general scheme for complex disease, such as diabetes, 3 key elements act together to determine the disease status (outcome) of a patient: *i)* the phenotype, i.e. the set of all metabolic, anthropometric and clinical variables characterizing the patient, *ii)* the genotype, i.e. the DNA sequence of the patient, *iii)* the set of interventions on the patient, i.e. therapies and treatments with drugs. All these 3 variables are connected each other through interactions and have a joint effect on the final outcome of the patient.

The multi-level approach allows to disjoint the full problem into sub-problems, focusing only on a set of variables and interaction (reflecting a specific level of information) according to available data.

In the present work, 3 main levels of study of diabetes complications are considered, and, for each approach, novel methodologies developed during my PhD are proposed.

The 3 levels of study considered in the present work are: *i)* modeling the effect of genotype on the outcome, *ii)* modeling the effect of phenotype and treatment on the progression of the outcome, *iii)* modeling the effect of treatment on the phenotype.

In the first level of study, diabetes complications are studied from a static point of view, i.e. without considering their progression over time, and the main objective is to identify the genetic biomarkers that allow to predict the disease state of the patients with the final goal to stratify patients according to the risk of developing the disease. Genome Wide Associations Studies (GWAs) are statistical studies aiming at identify those SNPs able to explain the differences observed for a certain outcome (the disease status) between cases (diseased subjects) and controls (healthy subjects) in a study population. Several methods performing univariate and/or multivariate selection have been used in literature for the identification of genetic markers from GWAs data. In this thesis, a novel algorithm for genetic biomarker selection and subjects classification from genome-wide SNP data has been developed. The algorithm is based on the Naïve Bayes classification framework, enriched by three main features: *i)* bootstrap aggregating of an ensemble of Naïve Bayes classifiers, *ii)* a novel strategy for ranking and selecting the attributes used by each classifier in the ensemble, *iii)* a permutation-based procedure for selecting significant biomarkers, based on their marginal utility in the classification process. The algorithm has been validated on the Wellcome Trust Case-Control Consortium on Type 1 Diabetes and its performance compared with the ones of both a standard Naïve Bayes algorithm and HyperLASSO, a penalized logistic regression algorithm from the state-of-the-art in simultaneous genome-wide data analysis.

The second level of study is represented by the dynamic analysis of diabetes complications, where the variable “time” plays a major role. In particular, the objective is to model the onset and the progression of diabetes complications over time, using phenotypic and therapeutic information, with the final goal to estimate a probability for the diabetic patient to develop a certain complication, thus optimizing clinical trials and avoiding invasive and expensive tests. So far, several models of diabetes complications are present in literature, but none is able to flexibly integrate accumulating –omics knowledge (i.e. proteomics, metabolomics, genomics) into a clinical macro-level. The most interesting complication models, in fact, are based on Markov Models (also called

state transition model) and use phenotypic information to describe the cohort of interest without the possibility to easily integrate additional information. A new in-silico model for simulating the progression of cardiovascular and kidney complications in diabetic patients is presented. The model proposes, as innovative feature, the use of Dynamic Bayesian Networks (DBNs) for modeling the interactions between variables. Compared to Markov Models, which require as many nodes as the number of combinations of variables' values, DBNs are more advantageous in handling both the structure and possible additional information, since each variable is simply represented by a node in the network. The model was built relying on data from the Diabetes Control and Complications Trial, a multicenter randomized clinical trial designed to compare intensive with conventional therapy with regard to their effects on the development and progression of the early vascular and neurologic. The developed model is able to predict the progression of the main diabetes complications with an accuracy greater than 95% at a population level. The model is suitable to be used as a decision support tool to help clinicians in the therapy design through cost-effectiveness analysis: exploiting the simulations generated through the model, it is possible, for example, to choose the best strategy between two different therapies for treating a specific cohort of patients. To this aim, a user-interface based on the present model is currently under development. The flexible structure of the model will allow to easily add genotypic information in the next feature as a potential mean to improve predictions.

The last level of study focuses on the action of a specific drug on a target phenotype, with the final aim to develop rational means to personalize drug therapy and to ensure maximum efficacy with minimal adverse effects. Focusing on cardiovascular diseases as a direct complication of diabetes, aspirin therapy is an important component of cardiovascular prevention for high risk patients. Aspirin performs its preventive action by inhibiting a key enzyme (the prostaglandin-endoperoxide synthase PTGS-1, also known as cyclooxygenase COX-1) in the cascade leading to the production of thromboxane B_2 (TxB_2), the major factor involved in the platelets aggregation with consequent formation of thrombi. It is known, from literature, that diabetic patients exhibit a different response to aspirin therapy in comparison to healthy subjects, showing a reduced effectiveness of the drug, which is often referred to as 'aspirin resistance'. Given the lack of a mathematical characterization of these phenomena, the problem was faced using a pharmacodynamics modeling approach, with an explorative intent. Relaying

on biological knowledge retrieved from literature, a partially lumped and partially distributed compartmental model was developed, able to describe: *i)* the kinetics of COX-1 enzyme, from its production within megakaryocytes in bone-marrow to circulating platelets in blood, *ii)* the pharmacokinetics and pharmacodynamics of aspirin, i.e. its distribution in the body tissues and its interaction with COX-1. The model was tested using data of serum thromboxane TxB_2 recovery levels after aspirin withdrawal in healthy subjects. Possible mechanisms to explain the so-called 'aspirin resistance' have been finally discussed.

Contents

Chapter 1

Introduction

Diabetes mellitus is a metabolic diseases in which a person's blood sugar is too high, either because the pancreas does not produce enough insulin, or because cells do not respond to the insulin that is produced, or a combination of the above mechanisms. Different variables, such as genetic, metabolic and environmental factors, play together in the onset and the progression of the disease, thus classifying diabetes as a complex-trait disease.

Diabetes is associated with severe long-term complications, mainly caused by the damage of blood vessels, both at micro and macro-level, because of the high glucose concentration in blood. As a result, the main organs involved are the heart and the cardiovascular system (diabetic cardiovascular complications), the kidney (diabetic nephropathy), the retina (diabetic retinopathy) and the nervous system (diabetic neuropathy). These complications heavily affect the quality of life of the patients and impose an immense impact on health care costs.

Therefore, novel means to prevent and/or treat these devastating diabetic complications are needed. Since long-term clinical trials are costly, time-consuming, and difficult to conduct, the use of computer-simulated disease models has increased considerably in recent years to facilitate the simultaneous evaluation of long-term clinical and economic effects of treatment. It is now widely accepted that models can provide valuable information for clinical practice and are important tools in medical, regulatory, governmental, and public health decision-making. A requirement for diabetes simulation models has been identified in the medical and healthcare policy community, and, as a result, a number of models have been developed and reported in the literature.

Given the complex nature of diabetes, the problem of investigating its long-term complications can be faced with a multi-level modeling approach.

In the present chapter, after a brief introduction on diabetes and its complications, the complex-trait nature of the disease will be described. Finally, the multi-level modeling approach, representing the general framework of this thesis, will be introduced.

1.1 *Diabetes and its complications*

Diabetes mellitus is a lifelong, incapacitating disease affecting multiple organs, that causes a person's blood sugar level to become too high. There are two main types of diabetes, referred to as type 1 (T1D) and type 2 (T2D).

T1D is often referred to as insulin-dependent diabetes. It is also sometimes known as juvenile diabetes or early-onset diabetes because it often develops before the age of 40, usually during the teenage years. T1D is an autoimmune condition, where immune system attacks and destroy pancreatic cells, responsible for insulin secretion. Thus, in T1D, the pancreas does not produce any insulin. Insulin is a hormone that regulates blood glucose levels. If the amount of glucose in the blood is too high, it can seriously damage the body's organs [1].

T2D occurs when the body doesn't produce enough insulin to function properly, or the body’s cells don't react to insulin. This is known as insulin resistance. T2D is far more common than T1D and it usually affects people over the age of 40, although increasingly younger people are also being affected. It is more common in people of South Asian, African-Caribbean or Middle Eastern descent [1].

It is important that diabetes is diagnosed as early as possible so that treatment can be started.

Diabetes cannot be cured, but treatment aims to keep blood glucose levels as normal as possible, and control symptoms to prevent health complications developing later. The therapy usually consists in a mixture of insulin infusions, diet and physical exercise [1].

Worldwide prevalence figures estimate that there are 250 million diabetic patients today and that this number will increase by 50% by 2025 [56]. The disease is associated with devastating chronic complications including coronary heart disease, stroke and peripheral vascular disease (macrovascular disease) as well as microvascular disorders leading to damage of kidneys (nephropathy) and eyes (retinopathy). These complications impose an immense burden on the quality of life of the patients and account for more than 10% of health care costs in Europe [27].

In the following, we will focus on the main vascular complications in T1D and T2D, i.e. diabetic nephropathy and retinopathy in both T1D and T2D and cardiovascular disease in T2D.

1.1.1 Diabetic nephropathy (DN)

Around 30% of patients with T1D and T2D develop DN [14]. Once manifest, DN is characterized by a progressive decline in kidney function, leading to end-stage renal disease (ESRD). DN represents the most common cause of ESRD (and hence the major precipitant of dialysis and transplantation therapy) in the Western world [23]. Metabolic control and elevated blood pressure are important risk factors, but these act in concert with genetic and other factors [23].

1.1.2 Diabetic retinopathy (DR)

Most patients with diabetes will develop some degree of DR and 2% will become blind. There is a strong correlation between duration of diabetes, glycemic control and development of DR. The prevalence of proliferative DR increases from 0% in those with less than 5 year duration to 26% after 15 years and to 56% after 20 years duration [36]. Both hypertension and dyslipidemia accelerate progression of DR. However, genetic factors clearly contribute to individual differences in the rate of progression and extent of DR.

1.1.3 Cardiovascular disease (CVD)

Up to 75% of all deaths in T2D are due to CVD. Also men and women with T1D have a fourfold and sevenfold risk of major CVD [48]. In addition to established risk factors such as smoking, dyslipidemia, hypertension and glycemic control [34], genetic factors are likely to be playing a substantial role in determining individual risk. Although there are no reliable heritability estimates for CVD in diabetic families, siblings of diabetic patients suffering from an early myocardial infarction have a 7-fold increased risk of CVD. The risk for development of a first myocardial infarction is increased 2-5 fold in subjects with diabetes, which makes the risk equivalent to that of a non-diabetic person with a previous myocardial infarction [25]. Moreover, the risk for recurrent acute cardiac events is more than 2-fold higher in diabetics than in non-diabetics. Patients with diabetes also have a 2 to 4-fold increased risk for development of stroke and peripheral arterial disease. Diabetes affects stroke outcome as well, with increased risk for subsequent

development of dementia, recurrence of a new stroke and death. It is also important to note that the relative protection from cardiovascular disease that characterizes pre-menopausal women is diminished by diabetes.

1.2 *Diabetes as a complex-trait disease*

Although hyperglycemia represents one of the most important risk factors for development of diabetic vascular complications, not all hyperglycemic patients seem to be at equal risk: other factors clearly modify an individual's susceptibility to develop complications, as reported in the previous section. It is thus clear that diabetes can be classified as a complex-trait disease, in which different factors such as genetic profile, metabolic and anthropometric phenotype and environmental risk factors, as well as individual response to treatments, concur to cause the onset of the disease and the development of different complications [51]. This complex nature is common both to T1D and T2D, even if a preponderant genetic cause characterizes T1D [42].

A general scheme of the main variables (and their interactions) involved in a complex disease such as diabetes is reported in Figure 1.1.

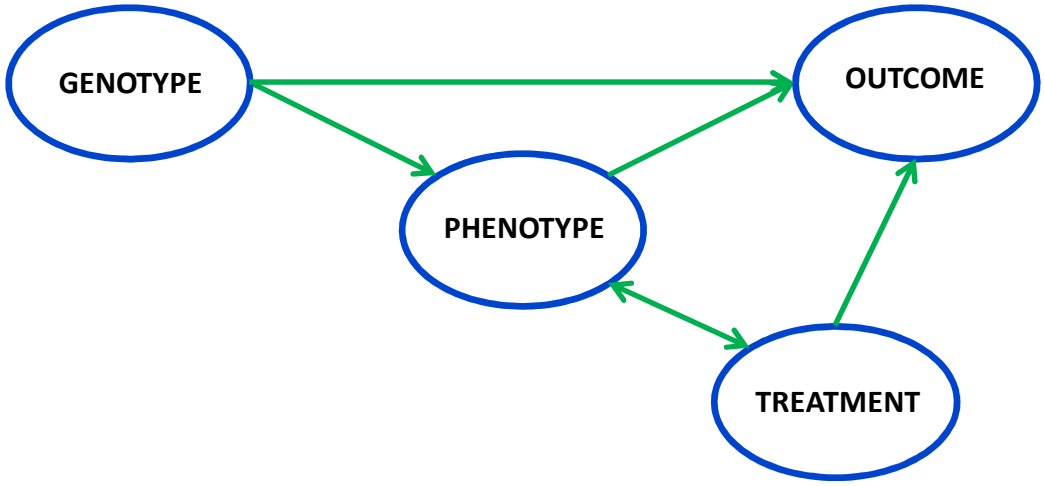


Figure 1.1: *General scheme of the multi-level model for diabetic complications.*

where,

- ***Genotype*** represents the genotypic profile of the individual, i.e. the Single Nucleotide Polymorphisms (genotypic biomarkers).
- ***Phenotype*** represents phenotypic biomarkers such as:

    - metabolic and physiological biomarkers (lipids, glycosylated hemoglobin, blood, pressure, heart rate, etc.);
    - anthropometric measures (weight, body mass index, etc.);
    - environmental factors (smoke status, physical activity, etc.);
- ***Outcome*** is the target of the study, i.e. diabetes or diabetic complications (Cardio Vascular Diseases, Nephropathy, Retinopathy, etc.);
- ***Treatment*** is the intervention variable, representing the specific therapy for the individual.

According to this general scheme, the *genotype* acts both on *phenotype* and on *outcome*, while the *phenotype* acts only on *outcome*. Treatment acts both on *phenotype* and *outcome*, but not on *genotype*.

1.3 *Multi-level approach*

Since diabetes is a complex-trait disease, the problem of investigating its long-term complications can be faced with a multi-level modeling approach: the scheme of Figure 1.1, in fact, represents different kind of variables and interaction between them. According to the level of detail which available data allows to reach, the interconnected structure of a multi-level model can be decomposed in sub-schemes, each one potentially analyzable independently on the others. Of course, the final ambitious aim of such an approach is to integrate all the possible sub-models (or at least most of them) in order to finally obtain a macro-model able to offer a complete characterization of the studied phenomena. Unfortunately, this objective is very difficult to achieve for most of the biological problems, given both the paucity of available data and the intrinsic limitations in the modeling process.

In this work, three main levels of study of diabetic complications will be considered, according to data availability: *i)* modeling the effect of genotype on the outcome (Figure 1.2.A), *ii)* modeling the combined effect of phenotype and treatment on the progression of the outcome (Figure 1.2.B), *iii)* modeling the effect of treatment on the phenotype (Figure 1.2.C).

For each approach, novel investigation methodologies are proposed.

The research presented in this thesis has been supported by the European Union's Seventh Framework Program (FP7/2007-2013) for the Innovative Medicine Initiative under grant

agreement n° IMI/ (the SUMMIT consortium), whose objective is to identify markers that predict the risks of developing diabetes chronic micro- and macro-vascular complications with focus on Diabetic Nephropathy, Diabetic Retinopathy and Cardiovascular disease. For a more detailed presentation of the concept and organization of the SUMMIT consortium, see public available information at *http://www.imi-summit.eu/*.

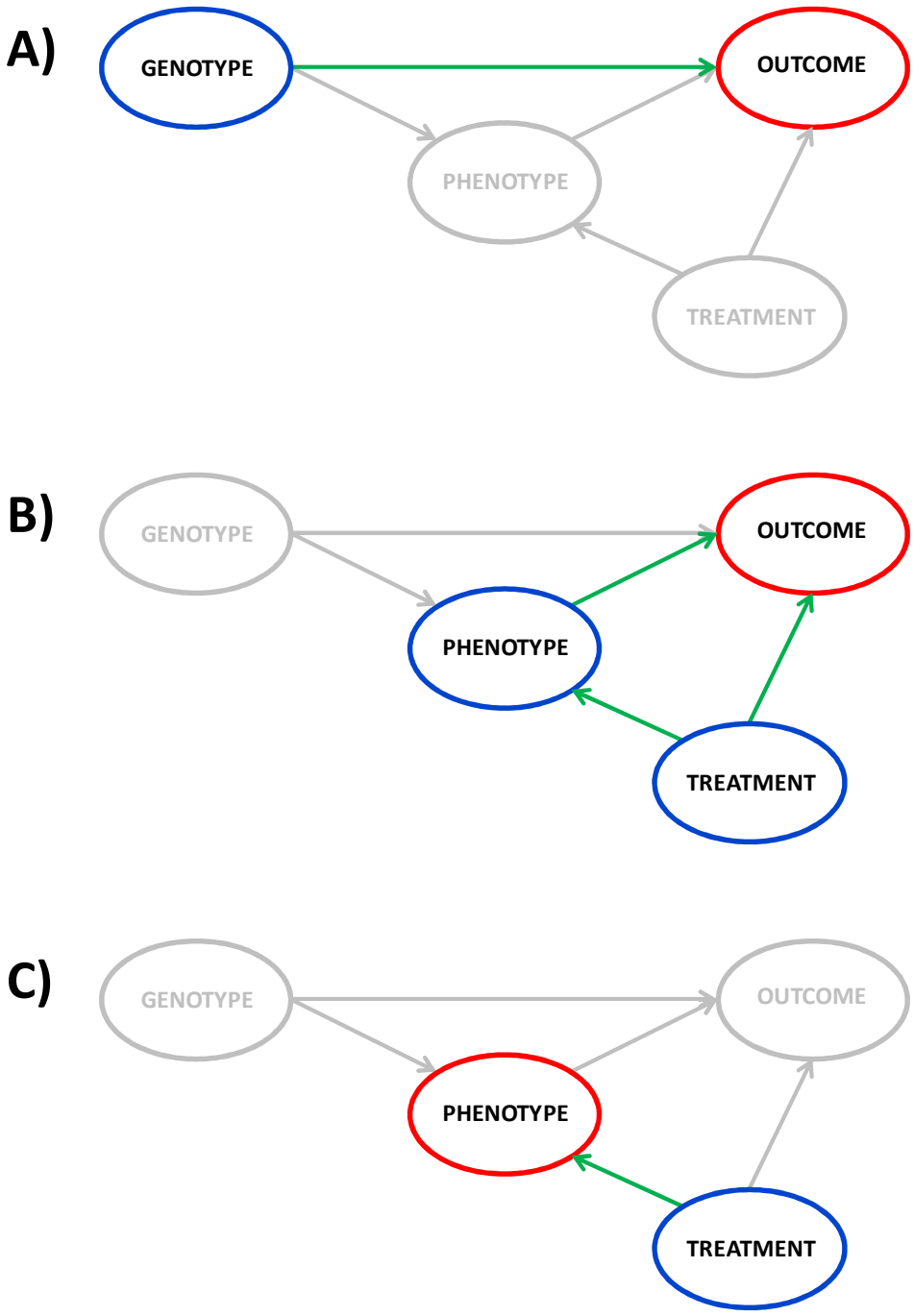


Figure 1.2: *Possible decompositions of the overall model for a multifactorial diseases: A) focus on the effect of the genotype on the outcome; B) focus on the combined effect of genotype, phenotype and treatment on the outcome; C) focus on the combined effect of treatment and genotype on the phenotype.*

1.4 *Outline*

Chapter 2 will deal with the first level of study – modeling the effect of genotype on the outcome. Data exploited to this aim are SNPs data from Genome Wide Association Studies, whose objective is to detect correlation between one or more genetic polymorphisms and a discrete trait (the presence or absence of a disease condition).

Chapter 3 will treat the second level of study – modeling the combined effect of phenotype and treatment on the progression of the outcome. To this aim, longitudinal data (coming from intervention clinical trials) regarding the main diabetes complications, as well as information on clinical variables (the phenotype) and on the treatment will be exploited.

Chapter 4 will focus on the last level of study – modeling the effect of treatment on the phenotype – in which data regarding the effect of a drug on a specific target phenotype will be exploited.

Chapter 2

Modeling the effect of genotype on diabetes: biomarker selection ad subject classification

Referring to the multi-level scheme presented in Figure 1.1, this chapter will focus on the effect of the genetic variables on the outcome, as shown in Figure 2.1.

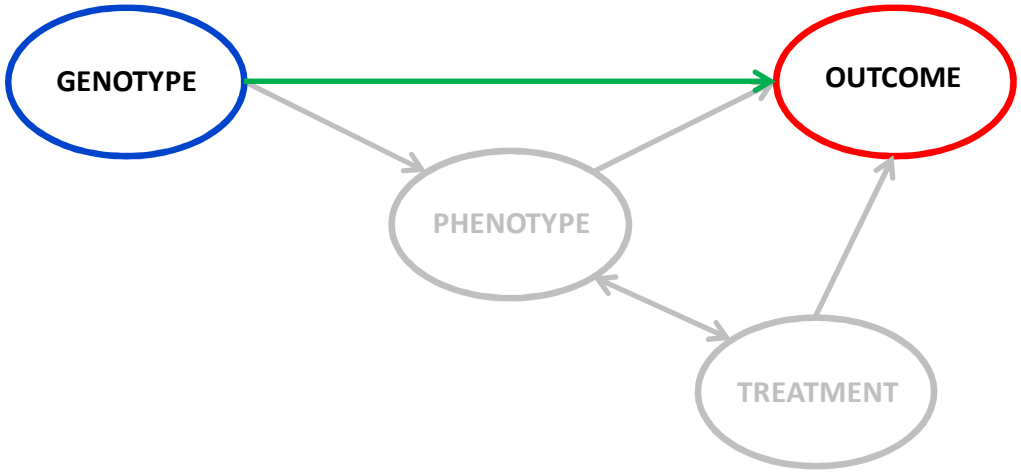


Figure 2.1: *Modeling the effect of genotype on the outcome.*

Genetic Association Studies and the objective of their study, i.e. SNPs, will be first introduced, with particular regard to Genome Wide Association Studies. Then, the most widely used approaches to analyze results coming from these studies will be briefly described to provide an exhaustive overview of the *state of the art*. Finally, a new algorithm for biomarker selection and subject classification from Genome Wide SNP data will be presented and its performance assessed by a comparison with a penalized logistic regression algorithm from the *state of the art* in simultaneous Genome Wide data analysis.

2.1 *Single Nucleotide Polymorphisms (SNPs)*

The DeoxyriboNucleic Acid (DNA) represents the hereditary material in humans and in most of all the other organisms.

The biological information is stored into the DNA as a string composed by four chemical bases: adenine (A), cytosine (C), guanine (G) and thymine (T). The sequence of these bases determines the information available for building and maintaining the organisms. DNA bases pair with each other (A-T; C-G) and form structures called "base pairs" (bp). Each base is linked to a sugar molecule and a phosphate molecule, forming a *nucleotide*.

Within the nucleus of each cell, the DNA macromolecules are packed into structures called chromosomes. In humans, each somatic cell contains 23 pairs of homologous chromosomes (46 chromosomes in total). In each pair of homologous chromosomes, one is inherited from the father, and one from the mother.

The *genes* represent the basic unit of heredity. A gene represents a segment of DNA (which physical location on the chromosome is called genic locus), and it contains the knowledge for coding *proteins*, macromolecules with either structural or functional biological roles [52].

In humans, as in other species, the length of the DNA sequence constituting a gene is extremely variable (from few hundreds of bp up to more than 2 millions bp). The *Human Genome Project* [http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml] estimated that human DNA sequence contains about 20,000 – 25,000 genes (~ 3% of the whole human DNA sequence, which is long ~ 3 billion base pairs).

More than 99.9 % of DNA sequence is identical between any two individuals. Even though most of the DNA sequence is identical, since the human genome sequence is so long, there are still many genetic variations.

Alleles are different versions of the same gene, carrying variations in terms of DNA sequence, which determinate the physical characteristics that differentiate individuals belonging to the same specie.

The *Single Nucleotide Polymorphism* (SNP) is defined as a DNA segment showing 2 or more alleles in a population and represents the simplest and most common source of variability among individuals [52]. For example, a SNP may replace the nucleotide pair G-C on a chromosome with the nucleotide pair A-T on the homologous chromosome, in a certain position of the DNA (Figure 2.2).

The combination of *allele*s in the same position on the two homologous chromosomes is called *genotype*, while a set of *alleles* on the same chromosome represents an *haplotype* [52].

For an SNP showing 2 alleles in a population, since chromosomes are in pairs, we can have 3 possible *genotype* for the single individual, each corresponding to a possible combination of the 2 *alleles*. Denoting with *A* the common form of the allele (i.e. most frequent in the population), and with *a* the rare form of the allele (i.e. the less frequent in the population), the 3 possible genotypes are: *AA*, *Aa* and *aa*, which are referred to as, respectively, common homozygous, heterozygous and rare homozygous [52].

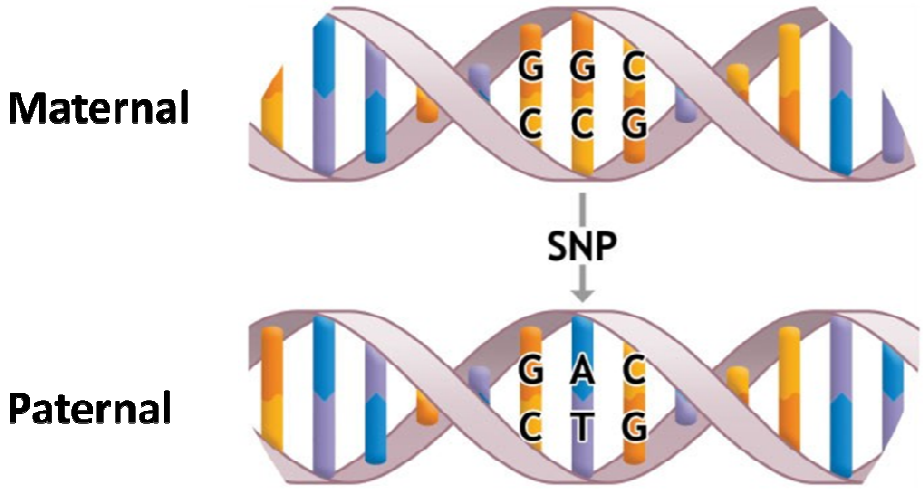


Figure 2.2: *Example of Single Nucleotide Polymorphism at a given locus.*

SNPs are not directly involved in causing a disease condition, but they modulate the probability of its occurrence, by interacting with other non-genetic predisposing factors (e.g. smoke or alcohol intake for some classes of cancer) and modulating the effect of the external interventions, such as treatments and drug assumption [5].

2.2 *Genome Wide Association Studies*

The study of complex diseases, such as diabetes, requires adequate tools. Genetic Association studies (*GAS*) are variants of the clinical and epidemiological case-control association studies applied to the field of population genetics [5]. The objective of a *GAS*

is to detect correlation between one or more genetic polymorphisms and a discrete trait (the presence or absence of a disease condition), comparing the frequencies of SNP alleles in two well-defined groups of individuals: cases who have been diagnosed with the disease under study, and controls, who are either known to be unaffected, or who have been randomly selected from the population. An increased frequency of a SNP allele or genotype in the cases class compared with the controls class indicates that presence of the SNP allele may increase risk of disease.

According to the a priori knowledge defined by the study design, *GAS* can be classified as follows, as the number of analyzed SNPs increases:

- *Candidate SNPs association studies.* This kind of studies focus on a single SNP which is suspected to have a causal role in the disease of interest.

- *Candidate genes association studies*. The object of the study is not a single SNP, but a set of markers (typically 5 - 10) located within the same potentially causative gene.

- *Fine mapping.* This kind of studies involve typically up to hundreds of nucleotides; the aim is to have a better definition (coverage) of a genome region potentially involved in physiological/pathological processes and previously identified by linkage studies or genome-wide association studies.

- *Genome Wide Association Studies (GWAS).* This approach consists in scanning markers across the complete sets of SNPs of many people to identify genetic variations associated with a particular disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses. *GWAS* are essentially “hypotheses free approaches”, i.e. this kind of studies do not require a prior knowledge about the right gene to be analyzed, but represent *hypotheses generating* instruments.

In the following sub-sections, a focus on *GWAS* will be presented.

2.2.1 Study design

When planning to perform a *GWAS* on unrelated individuals based on a case-control design, cases are selected on the basis of the trait of interest (i.e. type 2 diabetes), while control individuals must be clinically proven to be free not only from the condition of interest, but also from other traits that are not common to cases, otherwise a second sub-phenotype may be introduced in the analysis. The choice of the reference group may introduce confounding effects, and therefore an appropriate case-controls matching based on some phenotypic, exposure or environmental factors (gender, smoking history, ancestry) is required, in order to avoid spurious associations. Limiting factors when planning a *GWAS* are often represented by the difficulty of enrolling a sufficient number of cases and matched controls, and by the genotyping costs [5].

2.2.2 Genotyping

The process of examining DNA sequence in order to determine individual's polymorphisms is called *genotyping*. For the past decade, microarrays have grown in popularity as the primary tool for genotype analysis. Recently, however, *next-generation sequencing* (*NGS*) technologies has been introduced as a promising, new platform for genetic analysis, since they parallelize the sequencing process (i.e. s the process of determining the precise order of nucleotides within a DNA molecule), producing thousands or millions of sequences concurrently [26], thereby allowing to measure a huge amount of polymorphisms for each individual at the same time (up to > 1,000,000 SNPs). Nevertheless, for genotyping studies, microarrays are still widely adopted as they are substantially less expensive than NGS and much more conducive to processing thousands of samples required for typical genome-wide associations studies [26]. Illumina [Illumina, San Diego, CA] and Affymetrix [Affymetrix, Santa Clara, CA] represent the reference technologies for cost genotyping large amount of samples with high coverage in a cost effectively way (~ 200 $ for a 370 K SNPs chip).

Genotyping workflow according to Illumina protocols is represented in Figure 2.3.

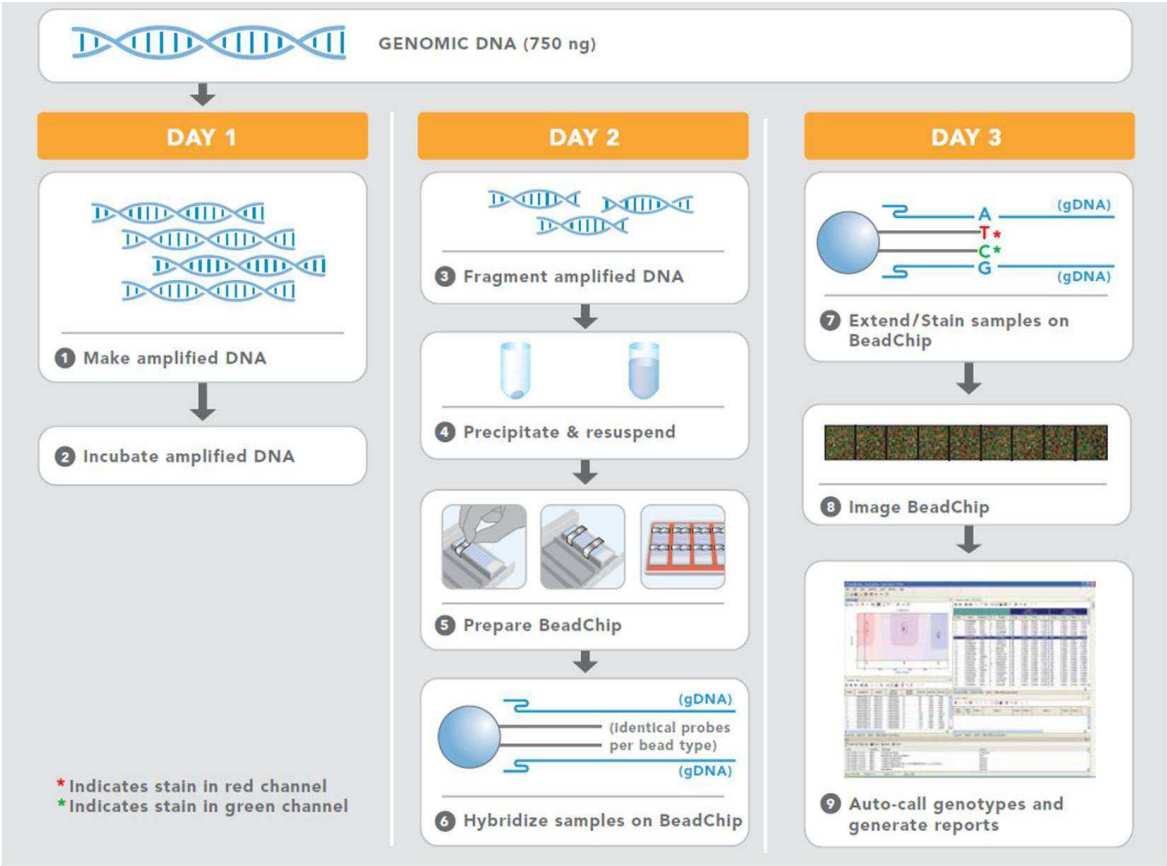


Figure 2.3: *Genotyping process according to Illumina protocol. The schema represent the main steps of the Illumina Infinium genotyping workflow. Figure from [http://res.illumina.com/documents/products/workflows/workflow_infinium_ii.pdf].*

Even if genotyping protocols differ between manufacturers, the main steps are common both to Illumina and Affymetrix protocols. A brief description of the genotyping processes is reported below.

1. *DNA extraction, purification and amplification*. Peripheral blood is drawn from each enrolled individual and successively DNA is extracted, amplified and purified.

2. *Hybridization on the chip*. DNA is labeled using fluorochromes and hybridized to the chips, each containing a redundant set of probes for each analyzed SNP. Mismatched and crosshybridization problems are avoided by different strategies, according to the manufacturer.

3. *Fluorescence intensities acquisition*. Fluorescence intensity for each SNP is captured for each analyzed sample by a scanner or "arrayer" at a fixed wave-length specific for different fluorochrome.

4. *Genotypes determination*. Once fluorescence captures have been extracted, “ad-hoc” programs allow for the quantization of the fluorescence intensities and for the genotypes determination. These softwares have been specifically developed by Illumina (Bead Studio and Genome Studio) and by Affymetrix (BRLMM) and implement multivariate clustering strategies for genotypes assignment on the basis of fluorescence intensity signals corresponding to each of the two alleles.

2.2.3 Quality Control

A preliminary analysis step is represented by data Quality Control, which is necessary to filter out low quality data in order to reduce the probability of false positive findings. Experimental systems involving biological material are typically prone to errors, often non-randomly distributed [5]. This lack of randomness is both due to the very nature of the available experimental technologies and to the presence of several concurrent factors such as DNA quality and preparation, specific experimental conditions or different skills of the experimenters, errors during the phase of genotypes assignment. Non random distribution of errors can affect results and reduce the power of the study [5]. Since most *GWAS* aim to identify very slight variations in allele frequencies between cases and controls, even the presence of small experimental errors could dramatically affect the outcome [6], [16]. Therefore it is necessary to apply filtering procedures in order to identify specific SNPs yielding errors in multiple individuals (markers-affecting errors), or individuals in the sample with errors across multiple SNPs (problems with the DNA sample), and simply exclude them from the analysis.

The basic Quality Control parameters that could help in identifying and removing low-quality samples and markers are the following:

- *Samples genotyping rate*: fraction of determined genotypes for each sample. This measure ranges from 0 (no genotype has been assigned) to 1 (all genotypes have been assigned).

- *SNPs genotyping rate*: fraction of determined genotypes for each SNPs. This measure ranges from 0 (no genotype has been assigned) to 1 (all genotypes have been assigned).

- *MAF value*: SNPs showing an extremely low value for the Minor Allele Frequency (*MAF* <0.01/0.05), could represent low genotyping-quality markers or too rare polymorphisms [5].

- *HWE p-value*: neutral genetic variants in a large random-mating population are expected to display Hardy Weinberg Equilibrium (*HWE*), under which assumption expected genotype frequencies satisfy the rules: E[*AA*]=p^2, E[*Aa*]=2*pq*, E[*aa*]=q^2, where *p* and *q* are the frequency of *A* and *a* alleles in the population, respectively. Genotyping errors can shift the SNPs observed frequencies from the expected proportions, and therefore testing for deviations from the *HWE* in the control population defining a cut-off on the result of the test (the *HWE p-value*) represents a standard approach to detect genotyping errors [53]. Such test can be performed using a Pearson goodness-of-fit statistic with one degree of freedom (d.f), under the null hypothesis of *HWE*.

2.3 *Univariate Analysis*

Once the preliminary phase of data QC has been performed, the next step usually consists in analyzing the whole set of markers, one SNP at a time, by univariate association tests under the null hypothesis of no association in order to identify SNPs statistically associated with the outcome, once a significance threshold *P* has been fixed.

The strength of the association between each single variable (SNP) and the outcome (disease/trait) is expressed in terms of *p-value*, which represents the probability of detecting an association that is stronger than that derived from data "by chance", when there is no evidence of association (i.e. a false positive): a very low *p-value* indicates that the observed result would be highly unlikely under the null hypothesis, which will be then rejected when the *p-value* turns out to be less than the significance threshold *P* [46].

The common way to represent the results of such a test is the so-called Manhattan Plot, in which, for each SNP, the *–log(p-value)* is reported, thus placing the most significant SNPs in the top part of the plot. Figure 2.4 represents, as an example, the association results coming from univariate association tests on a *GWAS* dataset on myopia [31],

where the –log(thresholds) are very high (i.e. the significance thresholds are very low) since a correction for multiple test has been performed (see section 2.3.3).

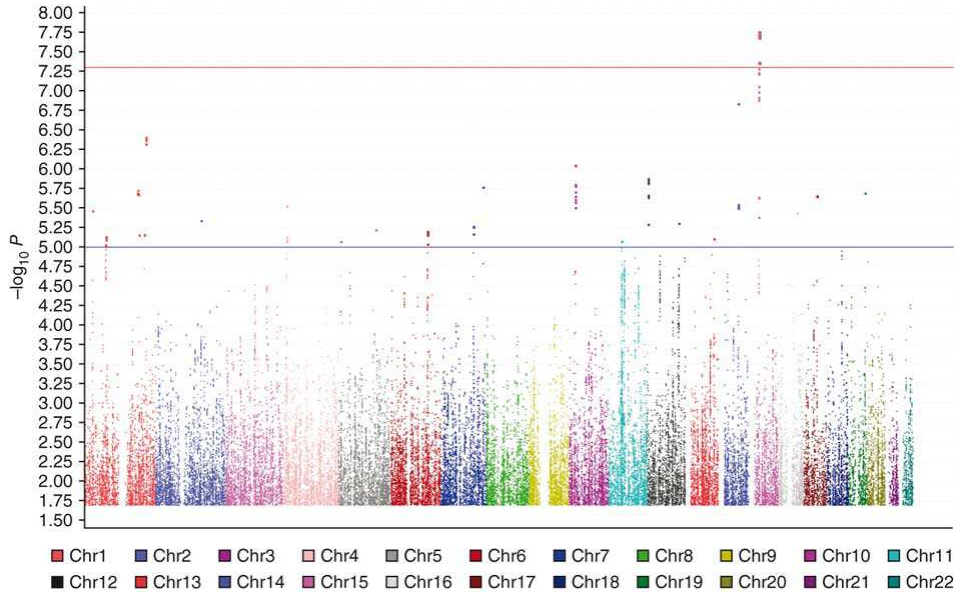


Figure 2.4: *The statistical significance values across the 22 autosomes of each SNP's association with refractive error (measured as spherical equivalent) are plotted as $-\log_{10}P$ values. X-axis represent the location of each SNP on the chromosomes, ordered according to their physical position. Y-axis represents the strength of association (-log10(p-value)) corresponding to each SNP. The blue horizontal line indicates $P = 10^{-5}$ and the red line indicates $P = 5 \times 10^{-8}$.*

The way of testing for association depends on the genetic model assumed for the SNP [39].

By defining the minor allele as *a*, the common allele as *A*, and the risk of developing the disease given a certain allele or genotype configuration as *R*, the most commonly used genetic models can be defined as follows:

- *Genotypic model* (*aa* vs *aA* vs *AA*). No a priori assumption is made about the association between genotype and phenotype, therefore the risk *R* is assumed to be equal for each genotype.

- *Dominant Model* (*aa*/*aA* vs *AA*). The underlying assumption of this model is that having one (*aA*) or two copies (*aa*) of the risk allele *a* induces the same risk *R* of being affected with respect to *AA* genotypes. The genotypes *aA* and *aa* are therefore pooled into the same group (*aa*/*aA*) and their frequency compared with the estimated frequency of the *AA* genotyes.

- *Recessive Model* (*aa* vs *aA*/*AA*). The assumption is that the risk *R* linked to a certain allele *a* is manifest only if it occurs in double copy *aa*. The frequency of *aa* genotype is therefore compared with *aA*/*AA* genotypes frequency.

- *Allelic or multiplicative model* (*a* vs *A*): the unity of the analysis is represented by alleles instead of genotypes. It assumes a multiplicative effect of the allele dosage (i.e. if heterozygote individuals *aA* risk *R* of developing the disease with respect to AA individuals, homozygote *aa* individuals risk *R***R* with respect to *AA* individuals).

- *Additive or trend model* (*aa* > *aA* > *AA*). The unity of risk is represented by alleles and it assumes that the risk linked to a certain allele has an additive effect on the case/control outcome (i.e. if heterozygote individuals *aA* risk *R* of developing the disease with respect to AA individuals, homozygote *aa* individuals risk *R*+*R* with respect to *AA* individuals).

The most widely employed association tests are based on the Pearson's χ^2 test and Cochran-Armitage test for trend.

2.3.1 Pearson χ^2 test.

Considering a pool of n_{case} unrelated cases, affected by the disease of interest, and unaffected n_{cont} controls for which a certain marker with alleles *A* and *a* has been genotyped, the sample genotype data can be represented by a 2 x 3 contingency table, as represented in Table 2.1.a, where the total number of subject is $n = n_{case} + n_{cont}$. The contingency table can be analyzed directly using an observed-expected test statistic, which has a χ^2 distribution on two degrees of freedom.

The χ^2 statistic tests for departure from the expected values across cells in the table. Thus the observed value for AA genotype in cases ($O_1 = N_{11}$) is compared with its expected value (E_1) given the total number of cases and the total number of AA genotypes, so $E_1 = n_{AA} \cdot n_{case}/n$. The full test statistic is given by equation (2.1):

$$X = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2 \qquad (2.1)$$

where the summation is over all six cells in the table, and O_i are the observed values N_{11}, N_{12}, N_{13}, N_{21}, N_{22}, N_{23} in each cell. Notice that this test statistic compares the observed number of AA genotypes in cases with that expected assuming both cases and controls have the same frequency of AA genotypes.

The Pearson's χ^2 test can be used for each genetic model, simply modifying the contingency table, as shown in Table 2.1.

(a) Full genotype table for a generic genetic model			
	AA	**Aa**	**aa**
Cases	N_{11}	N_{12}	N_{13}
Controls	N_{21}	N_{22}	N_{23}
(b) Dominant model: allele B increases risk			
	AA	**Aa + aa**	
Cases	N_{11}	$N_{12} + N_{13}$	
Controls	N_{21}	$N_{22} + N_{23}$	
(c) Recessive model: two copies of allele B required for increased risk			
	AA + Aa	**aa**	
Cases	$N_{11} + N_{12}$	N_{13}	
Controls	$N_{21} + N_{22}$	N_{23}	
(d) Multiplicative model: r-fold increased risk for AB, r^2 increased risk for BB. Analyzed by allele, not by genotype			
	A	**a**	
Cases	$2N_{11} + N_{12}$	$N_{12} + 2N_{13}$	
Controls	$2N_{21} + N_{22}$	$N_{22} + 2N_{23}$	
(e) Additive model: r-fold increased risk for AB, 2r increased risk for BB. Genotypes analyzed by Armitage's test for trend			
	AA	**Aa**	**aa**
Cases	N_{11}	N_{12}	N_{13}
Controls	N_{21}	N_{22}	N_{23}

Table 2.1: *Contingency tables for case control analyses, by genetic model. Test 1 is a baseline analysis, and any further analysis should be driven by prior hypothesis. a, b, c, d, e, f are genotype counts observed in cases and controls. Figure from [39]*.

2.3.2 Cochran-Armitage test for trend

For complex traits, it is widely thought that contributions to disease risk from individual SNPs will be often roughly additive – that is, the heterozygote risk will be intermediate between the two homozygote risk. The Pearson's χ^2 test have reasonable power regardless of the underlying risks, but if the genotype risks are additive they will be not as powerful

as tests that are tailored to this scenario. In fact, when data consist of a series of proportions occurring in groups which fall into some natural order, the question asked is then not so much whether the proportions differ significantly, but whether they show a significant trend, upwards or downwards, with the ordering of the groups. In this case, the Cochran-Armitage test can be applied [3]. It modifies the Pearson's χ^2 test to incorporate a suspected ordering in the effects of the three categories of the SNP. The idea is to test the hypothesis of zero slope for a line that fits the three genotypic risk estimates best (see example in Figure 2.5).

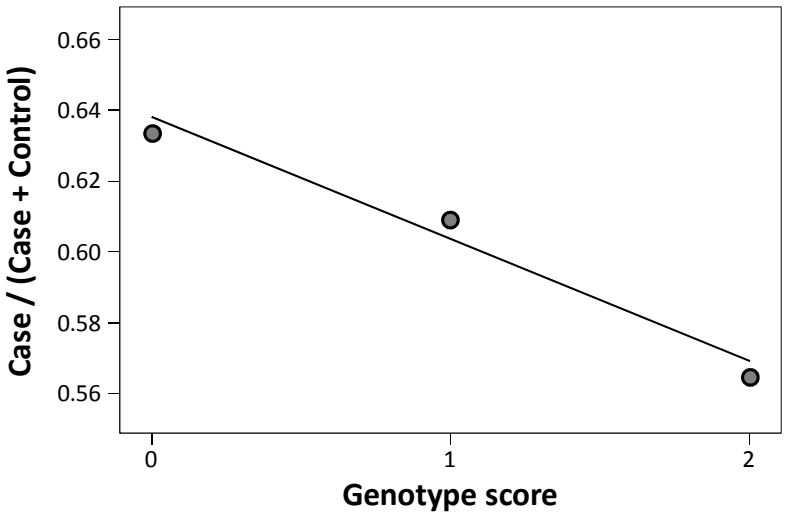


Figure 2.5: *Example of Armitage test of single-SNP association with case–control outcome. The dots indicate the proportion of cases, among cases and controls combined, at each of three SNP genotypes (coded as 0, 1 and 2), together with their least-squares line.*

Referring to Table 2.1.e, the trend statistic is:

$$T = \sum_{i=1}^{3} t_i (N_{1i} R_2 - N_{2i} R_1) \qquad (2.2)$$

where $R_1 = N_{11} + N_{12} + N_{13}$ is the number of cases, $R_2 = N_{21} + N_{22} + N_{23}$ is the number of controls, and the t_i are the weights selected according to the suspected mode of inheritance. For example, in order to test whether allele *A* is dominant over allele *a*, the choice $t = (1, 1, 0)$ is locally optimal. To test whether allele *A* is recessive to allele *a*, the optimal choice is $t = (0, 1, 1)$. To test whether alleles *A* and *a* are codominant, the choice $t = (0, 1, 2)$ is locally optimal. For complex diseases, the underlying genetic model is often unknown. In GWAS, the additive (or codominant) version of the test is often used. The test has good power in this case but power is reduced by deviations from

additivity. In an extreme scenario, if the two homozygotes have the same risk but the heterozygote risk is different (overdominance), then the Armitage test will have no power for any sample size even though there is true association.

2.3.3 Correction for multiple tests

Once a set of association tests have been performed, a *significance threshold* (usually a cut-off value on *p-value*) must be fixed in order to discriminate between statistically associated and null SNPs. The question of what strength of evidence should be considered significant has yet to be fully resolved in genetic association analysis [19]. On the one hand, multiple testing issues arise in most studies, whether based on candidate genes or genome wide scans, with attendant issues of how to quantify the multiplicity, what error rate to control and which method to use [40].

The following most commonly used approaches for dealing with the multiple testing issue are the following:

- *Bonferroni Correction*. Bonferroni correction [9] is the simplest procedure for assessing the significance threshold when multiple tests have been performed. This approach consists in rescaling the significance threshold α by the number of tests that have been performed N, in order to obtain a new multiple testing-adjusted significance threshold α':

$$\alpha = 1 - (1 - \alpha')^n \rightarrow \alpha' \approx \frac{\alpha}{n} \qquad (2.3)$$

- *False Discovery Rate* (*FDR*). Bonferroni correction is often considered an overconservative correction, with the deriving risk of losing biologically relevant associations [40]. A less conservative approach for facing the issue of multiple testing is represented by the calculation of the *False Discovery Rate* (*FDR*) as described by *Benjamini and Hochberg* [8]. For a family of hypothesis tests, let R denote the number of rejected null hypotheses, and V the number of falsely rejected null hypotheses. The *FDR* is then computed as follows:

$$FDR = E\left(\frac{V}{R}\middle| R > 0\right) \cdot P(R > 0) \qquad (2.4)$$

Benjamini and Hochberg introduced a step-up procedure for the control of FDR: given m null hypotheses to test $H_1, \ldots, H_m$, and $p_1, \ldots, p_m$ their correspondent p-values, chosen a significance level α, the control procedure works as follows:

a. order the p-values in increasing order and denote them by $p_{(1)}, \ldots, p_{(m)}$
b. find $k = \max_i \left\{ p_{(i)} \leq \frac{i}{m} \alpha \right\}$
c. reject all $H_{(i)}$ for $i = 1, \ldots, k$

This procedure is valid when all the m hypotheses are independent, otherwise step *b* changes into:

b. find $k = \max_i \left\{ p_{(i)} \leq \frac{i}{m} \frac{\alpha}{\sum_{i=1}^{m} \left(\frac{1}{i}\right)} \right\}$

2.3.4 Univariate Analysis: drawbacks

The described approaches examine one SNP at the time in relation to a defined trait. This over-simplistic strategy is not able to capture the multi-factorial nature of complex diseases, leading to the identification of a large set of associated SNPs (correlated by *Linkage Disequilibrium*, i.e. the association between two alleles located near each other on a chromosome, such that they are inherited together more frequently than would be expected by chance) but missing potential informative interactions [44].

Hoh and *Ott* [30] described the case in which the simultaneous presence of three genotypes at different loci induces a disease. By analyzing them though univariate models they would not result associated with the trait, since they share a low penetrance (i.e. poor association). This example is known as the Simpson's paradox and it explains also how the marginal independence of two variables (i.e. the evidence that knowledge of the first variable's value doesn't affect the belief in the second variable's value) does not necessary require their independence when other variables are taken into account [44].

2.4 *Multivariate Analysis*

The extremely large numbers involved in a GWAS ($O(10^6)$ SNPs in $O(10^3)$ individuals) have led the vast majority of studies to rely upon single SNP association tests, as already

described. Complex diseases, however, have an heterogeneous nature, arising from complex patterns of interaction between a set of genetic traits and the environment: to fully capture the optimal set of genetic biomarkers, thus, all SNPs in a GWAS should be analyzed simultaneously in a multivariate framework [29].
Multivariate models aim to do this, thus overcoming the described limitations that characterize the standard approaches. Moreover, they allow also to learn a rule for classifying unknown subjects as cases or controls, given their genetic profile and, possibly, other environmental covariates.

The most widely employed multivariate tests are based on Penalized Logistic Regression models.

2.4.1 Penalized Logistic Regression

Although the more usual way of modeling case and control data is in terms of probability distribution of genotype conditional upon disease status, reflecting the manner in which data are generated, *Prentice and Pyke* [43] demonstrated that comparable results could be obtained by applying a likelihood based approach in which case-control condition is considered a random outcome.

Given the random binary outcome Y (assuming only 2 possible states: 0 = control, or 1 = case) and one or more independent variables $\boldsymbol{X} = X_1, \ldots, X_p$ (SNPs in this case) the relation between Y and $\boldsymbol{X}$ can be modeled as the probability $P(Y|\boldsymbol{X})$.

Denoting $P(Y = 1| \boldsymbol{X} = x_1, \ldots, x_p)$ with π, that is the probability that an individual randomly drawn from the population is a case, equation (2.5) reports the logistic regression model

$$\ln \frac{\pi}{1-\pi} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \qquad (2.5)$$

Where $\beta_0 + \sum_{i=1}^{p} \beta_i x_i$ represents the linear relation between a function of π (named *logit*) and the independent variables x_i, in this case the SNPs of interest for the individual.

From equation (2.5) the probability π can be computed as:

$$\pi = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}} \tag{2.6}$$

Equation (2.6) states that the probability for the individual to be a case is a non linear function of his SNPs' values x_1, …, x_p and it ranges between 0 and 1. Figure 2.6 reports an example of equation (2.6) with only one independent variable x and parameters value $\beta_0 = 0$ and $\beta_1 = 1$.

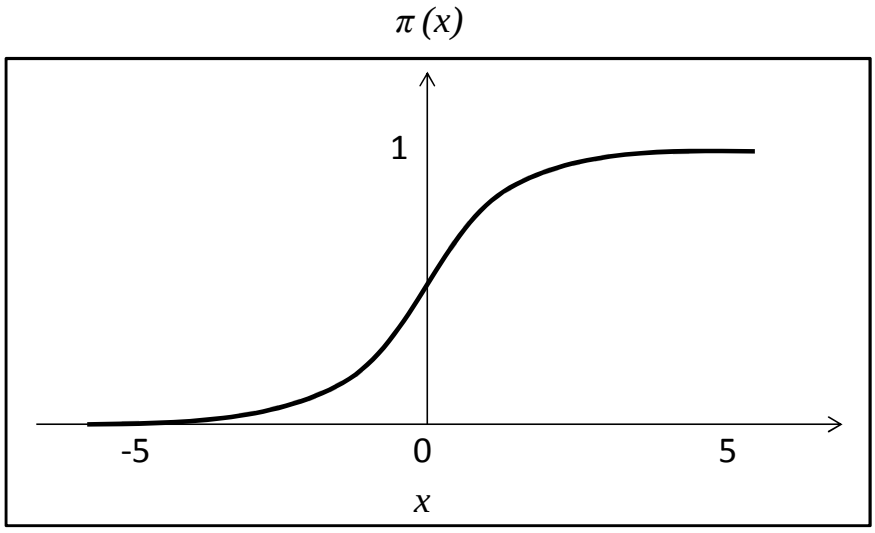


Figure 2.6: *Example of logistic regression model in the case of a single independent variable x.*

The probability for the observation $Y = y_i$ is given by:

$$P(Y_i = y_i | x_i) = (\pi)^{y_i} \cdot (1 - \pi)^{1 - y_i} \tag{2.7}$$

Given N independent observation (i.e. N different patients) the *maximum likelihood estimation* maximizes the log-likelihood for the N observations:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} \{y_i \cdot \ln[\pi_i^*] + (1 - y_i) \cdot \ln[1 - \pi_i^*]\} \tag{2.8}$$

$$\widehat{\boldsymbol{\beta}} = arg\max_{\boldsymbol{\beta}} \{l(\boldsymbol{\beta})\} \tag{2.9}$$

In the case of GWAS, when the number of markers is larger than the number of test subjects, since only a very small set of SNPs (compared to the total number of SNPs) is

likely to have an effect on the outcome, a *penalization* term is introduced in the logistic regression model, in order to obtain *sparse* solution, i.e. select few predictive variable without information loss [54]. These methods operate by shrinking the size of the coefficients of markers with little or no apparent effect on a trait down to zero.

The main penalization strategies are the following:

1. *LASSO regression* (*L1 penalization regression*). The *Least Absolute Shrinkage and Selection Operator* (*LASSO*) penalized regression, proposed by Tibshirani [50], estimates the parameters of the logistic model adding to the likelihood the penalization term given by:

$$L_1 = |\boldsymbol{\beta}| = \sum_{i=1}^{p} |\beta_i| \qquad (2.10)$$

The model parameters vector $\boldsymbol{\beta}$ is then estimated by:

$$\widehat{\boldsymbol{\beta}} = arg\max_{\boldsymbol{\beta}}\{l(\boldsymbol{\beta}) - \lambda \cdot |\boldsymbol{\beta}|\} \qquad (2.11)$$

Where λ is the weight of the penalization term.
In a Bayesian interpretation, Lasso Regression can be derived as Bayes posterior mode under independent double-exponential priors for the β_i [50]. Lasso Regression shrinks coefficients $\beta_1, \beta_2, \ldots, \beta_p$ setting most of them to 0 and thus selecting the most significant variables. However, for a problem with N observation, it can select no more than N variables.

2. *Ridge regression* (*L2 penalization regression*). The L2 penalized regression, proposed by Hoerl and Kennad [28], estimates the parameters of the logistic model adding to the likelihood the penalization term given by:

$$L_2 = |\boldsymbol{\beta}|^2 = \sum_{i=1}^{p} {\beta_i}^2 \qquad (2.12)$$

The model parameters vector $\boldsymbol{\beta}$ is then estimated by:

$$\widehat{\boldsymbol{\beta}} = arg\max_{\boldsymbol{\beta}}\{l(\boldsymbol{\beta}) - \lambda \cdot |\boldsymbol{\beta}|^2\} \quad (2.13)$$

Where λ is the weight of the penalization term.

In a Bayesian interpretation, Ridge Regression can be derived as Bayes posterior mode under independent Gaussian priors for the β_i [28]**.**

Ridge regression shrinks coefficients β_1, β_2, …, β_p but does not set any one to 0, thus makes no real variable selection.

3. *Elastic Net regression* (*L1L2 penalization regression*). The L1L2 penalized regression, proposed by Zou and Hastie [57], estimates the parameters of the logistic model adding to the likelihood the penalization term given by a convex combination of L1 and L2 penalties. The model parameters vector $\boldsymbol{\beta}$ is then estimated by:

$$\widehat{\boldsymbol{\beta}} = arg\max_{\boldsymbol{\beta}}\{l(\boldsymbol{\beta}) - \lambda \cdot [(1-\alpha) \cdot |\boldsymbol{\beta}| + \alpha \cdot |\boldsymbol{\beta}|^2]\} \quad (2.14)$$

Where λ is the weight of the global penalization and α determines the relative weight of L1 and L2 penalties.

Elastic Net produces sparse models encouraging a grouping effect, where strongly correlated predictors tend to be in or out the model together.

4. *Minimax Concav Penalty* (*MCP regression*). Zhang et al. [55] propose a penalization term give by:

$$f_{\lambda,a}(\beta) = \begin{cases} \lambda\beta - \dfrac{\beta^2}{2a} & if\ \beta \le a\lambda \\ \dfrac{1}{2}a\lambda^2 & if\ \beta > a\lambda \end{cases} \quad (2.15)$$

The effect of the penalty is determined by the gradient of equation (2.15):

$$\frac{\partial f_{\lambda,a}}{\partial \beta}(\beta) = \begin{cases} \lambda - \dfrac{\beta}{a} & if\ \beta \le a\lambda \\ 0 & if\ \beta > a\lambda \end{cases} \quad (2.16)$$

Where λ is determines the magnitude of the penalization and a the range over which the penalization is applied. MCP regression causes unimportant variables to be eliminated, leaving the important ones unpenalized.

5. *HyperLASSO regression* (*NEG regression*). Hoggart et al. [29] propose a variant of the Bayesian interpretation of the LASSO regression, using a Normal Exponential Gamma (*NEG*) distribution as a sparseness-inducing prior on the weights $\beta_1, \beta_2, \ldots, \beta_p$. *NEG* is a generalization of double-exponential distribution with 2 extra parameters regulating scale and sharpe of the curve. The sharper peak induces sparse solutions, while heavy tails result in variables being minimally shrunk once included in the model [29].

In Figure 2.7, plots of the negative of the penalty functions $-\lambda f(\beta)$ are shown. Considering that the effect of the penalty is determined by the derivative of the penalty function, one can observe the different behavior of the different methods. The Lasso encourages sparsity, setting most small coefficients to zero, due to the penalty function's sharp peak at zero. However, given a sufficiently large penalty parameter, the Lasso also imposes heavy shrinkage on large coefficients due to the absence of tails (constant rate of penalization), leading to biased coefficient estimates. A similar issue of bias can be seen also for other penalty functions such as the Ridge penalty and the Elastic Net. MPC and HyperLasso strive to relieve some of this bias introducing penalties with flatter tails, so that large coefficients are only minimally shrunk.

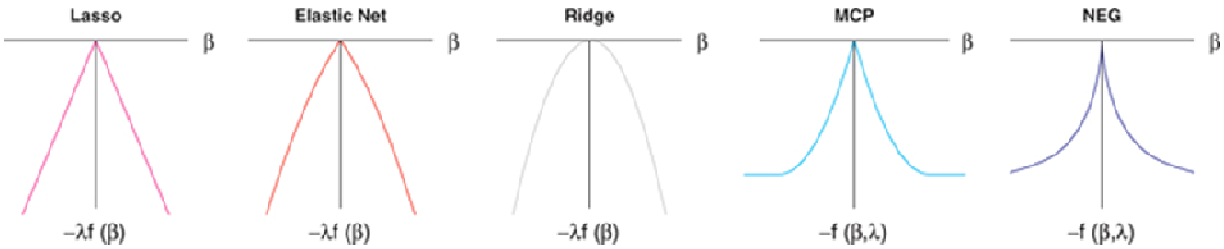


Figure 2.7: *Plots of the negative of the penalty functions – λf(β). The penalty (y-axis) is plotted against β (x-axis) for the Lasso, Elastic Net, Ridge, MCP and NEG. The peaks of each function are at β=0. Figure from [4].*

Ayers and Cordell [4] used computer simulation to compared the five penalties described above (*L1*, *L2*, *L1L2*, *MCP* and *NEG*) to standard single locus analysis (Armitage test for

trend *ATT*) and simple forward stepwise regression (*FSTEP*). The authors explored the performances of penalization in selecting SNPs as predictors in two simulated genetic association studies. In particular, the methods were first compared with respect to *detection* of effect, in which detection of an allele in linkage disequilibrium ($r^2 > 0.05$) with a true causal variant counted as a success (and any other detection counted as a false positive), and second with respect to *localization*/*differentiation*, in which only counted detection of the true causal locus itself as a success. In the first simulation study, a GWAS was simulated by generating 500 replicate datasets, each composed of 1000 cases and 1000 controls, 4000 SNPs and 6 causal loci. In the second simulation study, a Fine Mapping study was simulated by generating 500 replicate datasets, each composed of 1000 cases and 1000 controls, 3 given gene regions (*CYP2D6*, *CFTR* and *CTLA4*, containing 110, 190 and 228 SNPs respectively) and 5 causal loci within each region.

Since all penalized regression methods required input of one or more values for the penalization parameter/s (to which we can refer as λ), rather than finding the best value for λ, results were analyzed in terms of AUC (Area Under Curve) in the ROC (Receiver Operating Characteristic) space, as the penalty parameter λ was varied. In particular, with respect to the first simulation study, Figure 2.8 shows the relationship between true and false-positive detection for each of the methods as λ is varied, while, with respect to the second simulation study, Figure 2.9 shows he relationship between true and false-positive detection for each of the methods as λ is varied in the three gene regions of interest.

Figure 2.8 and Figure 2.9 show similar performances between the different multivariate methods, with NEG giving the overall best and ATT the overall worst performance. Although larger parameter estimates are always more heavily penalized, methods that apply larger relative penalties on small parameters estimates and relatively lower penalties to larger estimates performed better and more accurately estimate the effect size of the selected SNPs. The superior performance of the HyperLasso regression with respect to detection as well as with respect to differentiation/localization of effects makes it a gold standard for GWAS SNPs analysis [4].

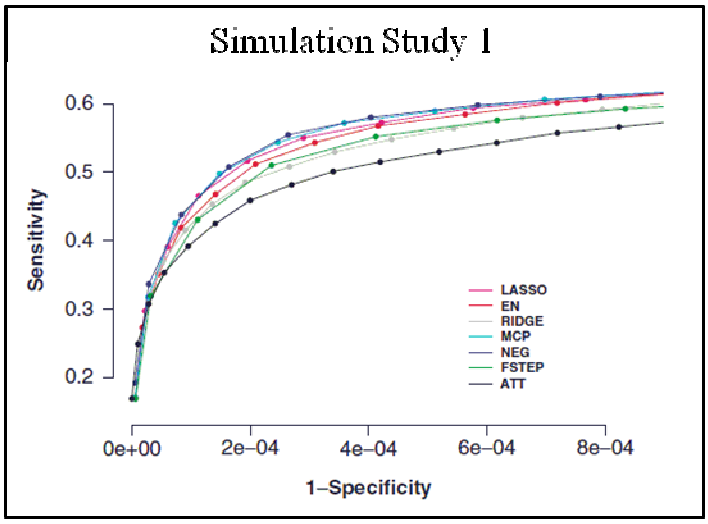


Figure 2.8: *Sensitivity (detection rates) versus 1-specificity (false positive rates) as the penalty parameter λ is varied, for simulation study 1. Figure from [4].*

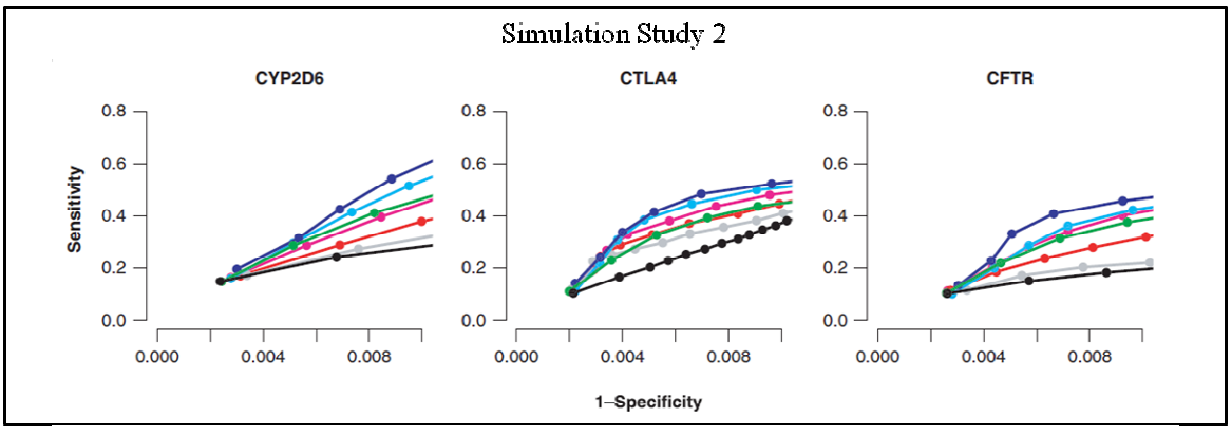


Figure 2.9: *Sensitivity (detection rates) versus 1-specificity (false positive rates) as the penalty parameter λ is varied, for the three gene regions of simulation study 2. Figure from [4].*

2.4.2 The problem of robustness for multivariate approaches

The identification of robust lists of biomarkers represents a fundamental issue as it may greatly influence subsequent steps, such as the definition of targets for clinical and pharmaceutical applications, as well as early diagnosis and treatment of diseases. All the methods described in the previous section for the discovery of biomarkers of complex diseases from high-throughput data often provide results with limited overlap or reduced statistical significance [21], [10]. As summarized in [18], these difference in results are mainly imputable to:

1. heterogeneity of both experimental protocols and computational pipelines carried out for the analysis;
2. dataset size, which often include few subjects (some hundreds) with respect to the number of features (up to $O(10^6)$ SNPs);
3. heterogeneity of the complex disease, responsible for high correlation in the features, some of these being real causal loci, other being correlated by linkage disequilibrium.

As a result, different features may thus be selected under different settings, even when good classification accuracy is reached (it is in principle possible to have a lack of stability due to the presence of many highly correlated features, even with accuracy equal to one).

The stability issue in feature selection has received much attention recently, as well as the precision of biomarker identification, i.e. the ability to select true biomarkers, defined as features biologically related to the physiological or clinical condition under study as cause or effect of it [17], [58], [2].

In the next section, a new algorithm for biomarker selection and subject classification from genome-wide SNPs, developed to effectively handle the problem of robustness in the biomarker discovery will be presented.

2.5 *Bag of Naïve Bayes*

As described in the previous sections, the analysis of genome-wide SNP data for complex diseases mainly suffers from two, intertwined problems: on the one hand, multifactorial diseases are caused by complex patterns of interaction between multiple genetic traits and the environment, on the other hand, linkage disequilibrium confounds the search for genetic biomarkers, because of the non-random association between the true genetic causes and the SNPs in genomic regions close to them, thus resulting in a lack of precision and stability of the lists of biomarkers selected by different methods, as reported in section 2.4.2.

In this section a new algorithm, Bag of Naïve Bayes (BoNB), developed to effectively tackle both of these problems, is presented.

As reported in section 2.4, the most widely used methods for the simultaneous SNP analysis on a genome wide scale relay on the penalized logistic regression framework, where SNPs are modeled as discrete variables from the domain {0,1,2} and a log-additive model of genetic effect on the disease is assumed.

BoNB is based on Naïve Bayes (NB) classification [41], thus it relies on contingency table analysis without having to assume a pre-specified model of genetic effect and, differently from logistic regression methods, it can easily handle missing values in the data, without having to perform imputation. Three main strategies are exploited in BoNB to tailor the Naïve Bayes framework to Genome Wide SNP data analysis: (a) a bagging of Naïve Bayes classifiers, to improve the robustness of the predictions, (b) a novel strategy for ranking and selecting the attributes used by each bagged classifier, to enforce attribute independence, and (c) a permutation-based procedure for selecting significant biomarkers, based on their marginal utility in the classification process.

Before describing the algorithm, a brief introduction on the Naïve Bayes Classifier is reported in the following.

2.5.1 Naïve Bayes Classifier

The Naïve Bayes classifier (*NB*) is one of the most efficient classification algorithms for machine learning and data mining [41]. *NB* has been widely used for classification purposes in the biomedical fields and, more recently, in the context of *GWAS* [44].
The reasons of its diffusion are essentially its good classification performance and computational efficiency. *NB* is the simplest form of Bayesian classifier, in which all the variables are assumed to be independent given the value of the outcome [41].

Given a dataset $X = \{X_1, \ldots, X_n\}$, consisting of n observations (subjects) of p attributes (SNPs), and a set Y of class labels, one for each observation (case/control), a Naïve Bayes classifier estimates, from the dataset D, a classification rule in the form:

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) \cdot \prod_{i=1}^{n} P(X_i | Y = y_k)}{\sum_{j=1}^{p} P(Y = y_k) \cdot \prod_{i=1}^{n} P\left(X_i \middle| Y = y_j\right)} \qquad (2.17)$$

The classification rule of equation (2.17) states that the probability of a subject being in class y_k, given a combination of values for the attributes $X_1, \ldots, X_p$, is equal to the a priori probability of class y_k, $P(Y = y_k)$, times the probability of each attribute given class y_k, $P(X_i|Y = y_k)$: the implicit assumption below this classification rule is that attributes $X_1, \ldots, X_p$ are all conditionally independent given Y.

Given discrete-valued inputs and binary outcome, the algorithm learns the probability distribution of equation (2.17) estimating two sets of parameters. The first is

$$\theta_{ijk} = P\left(X_i = x_{ij} \middle| Y = y_k\right) \cong \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\} + l}{\#D\{Y = y_k\} + lJ}, \qquad (2.18)$$
$$j = 1, \ldots, J$$

For each input attribute X_i, each of its possible values x_{ij} ($J = 3$ in case of SNPs) and each of the two possible values y_k of Y. The $\#D\{x\}$ operator returns the number of elements in the set D that satisfy property x. The second is the *prior* probability over Y:

$$\pi_k = P(Y = y_k) = \frac{\#D\{Y = y_k\} + l}{|D| + 2l} \qquad (2.19)$$

Where |D| denotes the number of elements in the set D.

The l term is the only tunable parameter of the Naïve Bayes algorithm and it is known in the Bayesian literature as Equivalent Sample Size or Dirichlet Weight [41], and represents a prior probability which prevents the class-conditional probabilities from becoming zero when training attributes are sparsely populated.

2.5.2 Methods

BoNB consists in an ensemble of Naïve Bayes Classifiers, trained on GWAS data with the procedure known as Bootstrap Aggregating or Bagging [12].

Given a training dataset X, the Bagging procedure starts by computing a set of Bootstrap replicates of X, i.e. a set $\{X^{(1)}, \ldots, X^{(B)}\}$ of datasets, each one obtained by sampling n observations with replacement from the training set X [20]. A Naïve Bayes Classifier $NBC^{(b)}$ is then trained on each Bootstrap sample $X^{(b)}$. Classification of unseen subjects, drawn from an independent test set, is then obtained by majority vote or weighted average of the output class probabilities computed by each $NBC^{(b)}$ (Figure 2.10). Such an approach is known not only to increase the robustness of the predictions in terms of

classification accuracy [12], but also to improve the precision and stability in the step of feature selection [18].

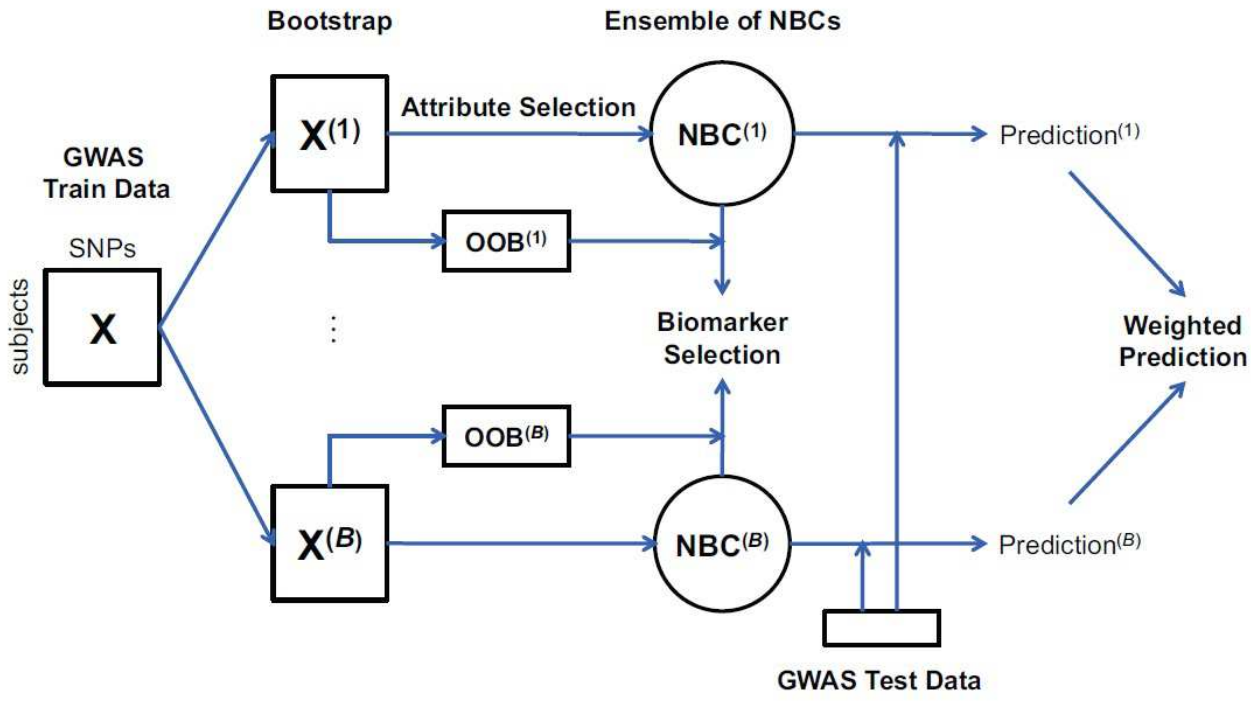


Figure 2.10: *Schematics of the BoNB algorithm: B Bootstrap samples {X(1),…, X(B)} are drawn from a GWAS training dataset X; B Naïve Bayes Classifiers (NBC) are trained on the Bootstrap samples, with the novel procedure for attribute ranking and selection; predictions of unseen subjects from a GWAS test dataset are carried out independently by each NBC and class probabilities are then averaged; biomarker selection is carried out with the novel permutation-based procedure, exploiting Out-of-Bag (OOB) samples.*

Given the binary nature of the case/control classification problem and the frequent unbalance between the number of cases and controls in a GWAS, classification performances are evaluated by the *Matthews Correlation Coefficient* (*MCC*, [12]). The *MCC* is defined as:

$$MCC = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fp) \cdot (tp + fn) \cdot (tn + fp) \cdot (tn + fn)}} \qquad (2.20)$$

where *tp*, *tn*, *fp* and *fn* stand for true positives, true negatives, false positives and false negatives, respectively.

The *MCC* is often preferred to standard classification accuracy, i.e. to the proportion of correctly classified examples, because it is not sensitive to class unbalance: the *MCC*, in

fact, ranges from -1 (all examples incorrectly classified) to 1 (all correctly classified) and equals 0 in case of majority classification, i.e. when all labels are assigned to the most represented class.

For estimating probabilities as in equation (2.17), the Naïve Bayes Classifier makes the assumption that the attributes $\{X_1, \ldots, X_n\}$ are all conditionally independent of one another, given Y . Such an assumption is unlikely to hold if all the SNPs of a GWAS are exploited as attributes, because of genetic linkage. Moreover, computing equation (2.17) for the whole SNP set can be computationally heavy and can lead to numerical and overfitting problems.

A procedure for selecting a good set of independent SNPs for each $NBC^{(b)}$ was thus developed.

The procedure first ranks each SNP according to the classification performance obtained on the training set $X^{(b)}$ itself by using the SNP as a single attribute of the $NBC^{(b)}$. To account for possible class unbalance, classification performance is assessed with the *MCC* (equation (2.20)). The obtained *MCC* represents the score of the SNP. SNPs are then ranked in decreasing order of score, obtaining a ranked list for each $NBC^{(b)}$.

In the second step, SNPs are iteratively added, in decreasing order of score, as attributes of each $NBC^{(b)}$ from its corresponding ranked list, computed on $X^{(b)}$. Each time a SNP is included as an attribute, all the SNPs in the ranked list that are both close to the SNP on the genome (distance < 1Mb) and correlated with it ($r^2 > \theta$, where r^2 is the squared correlation coefficient and θ is a user defined threshold) are removed from the list: such an approach enforces attribute independence, thus coping with the problems arising from genetic linkage. Rather than including one SNP at a time, uncorrelated SNPs are added in groups of exponentially increasing size, starting from one SNP and doubling the size at each new addition. New SNPs are added as long as the generalization ability of $NBC^{(b)}$ increases: to estimate the generalization ability, each $NBC^{(b)}$ is tested on the corresponding Out-of-Bag sample $OOB^{(b)}$, consisting of all the observations left out from X when sampling $X^{(b)}$, and the *MCC* of the prediction is measured. The exponential increase in the number of added attributes allows BoNB to reach the adequate size for the attribute set of each NBC in a logarithmic number of steps.

Such an attribute selection procedure, iterated for the B bootstrap samples, results in an ensemble of B Naïve Bayes Classifiers, each with a possibly different set of features. Classification of new subjects, the first objective of GWASs, is then obtained by having each *NBC* estimate output class probabilities and by averaging the probabilities across all the B *NBCs*. Classification performance of the ensemble of *NBCs* can then be assessed on an independent GWAS test set, by measuring the *MCC* of the predictions.

For the second objective of GWASs, biomarker selection, a procedure originally designed for the Random Forests bagged classifier [11] was adapted for BoNB: for each of the SNPs included as attributes by at least one *NBC*, the genotype of the SNP is randomly permuted in the *OOB* sets, each $NBC^{(b)}$ is tested on its corresponding $OOB^{(b)}$ and the relative decrease in *MCC* due to the permutation is recorded. Such a measure, which can be used as an indicator of the importance of each selected attribute given all other selected attributes, is defined *marginal utility* (*MU*).
For each SNP, the permutation procedure returns a list of values of *MU*, one value for each *NBC* that included the SNP: *MUs* significantly greater than zero are tested with a one-tailed Wilcoxon signed rank test on the list of values, selecting as biomarkers the SNPs for which the p-value of the test is lower than 0.05.

2.5.3 Results

BoNB was tested on the WTCCC case-control study on Type 1 Diabetes [13], where approximately 2000 T1D cases and 3000 healthy controls were examined. Each subject was genotyped on the Affymetrix GeneChip 500K Mapping Array Set. A small number of subjects was excluded according to the sample exclusion lists provided by the WTCCC. In addition, a SNP was excluded if (i) it is on the SNP exclusion list provided by the WTCCC; (ii) it has a poor cluster plot as defined by the WTCCC. The resulting dataset consists of 458376 SNPs, measured for 1963 cases and 2938 controls.

The BoNB algorithm exposes two parameters to the user: the number of Bootstrap replicates and Naïve Bayes Classifiers, *B*, and the threshold on the squared correlation coefficient above which two SNPs are considered correlated, θ. *B* and θ were set to200 and 0.1, respectively (see section 2.5.4 for an analysis of how performance is affected by variations of the parameters *B* and θ).

Classification performance was estimated on independent train-test set pairs obtained by repeatedly sub-sampling at random 90% of the dataset for training and 10% for testing. The procedure was iterated 10 times and classification performance was assessed with the MCC of the predictions on the test sets. The list of selected biomarkers, on the other hand, was computed on the whole dataset.

Since BonB is based on the Naïve Bayes classification framework and has been developed as a valid alternative to penalized logistic regression methods, classification performance was compared with the ones obtained by a standard Naïve Bayes Classifier, trained on all the SNPs that reached the significance threshold of 5×10^{-7} (as in [13]) in a single $2df\ \chi^2$ test of association with a general genetic model, and by HyperLASSO, a logistic regression method representing the gold standard for the simultaneous analysis of all SNPs in a GWAS, described in section 2.4.1. The former algorithm was chosen to assess the improvement of BoNB both in terms of biomarker selection, with respect to a standard univariate test, and in terms of classification performance, with respect to the algorithm on which BoNB is based. The latter algorithm was chosen because of its best performance among classification and biomarker selection methods for genome-wide data, as reported in [54] and [4], and because of the complete availability of the source code. On the experimental dataset, BoNB reached an MCC of 0.55 ± 0.03 (mean ± standard deviation), significantly higher than the ones reached by both the standard Naïve Bayes Classifier (0.31 ± 0.05, Wilcoxon signedrank p-value 0.002) and by HyperLASSO (0.45 ± 0.03, p-value 0.002). Figure 2.11 (left panel) shows the boxplots of the MCC obtained by the three algorithms on the ten iterations of the sub-sampling procedure. For the sake of completeness, Figure 2.11 (right panel) shows also the boxplots of classification accuracy.

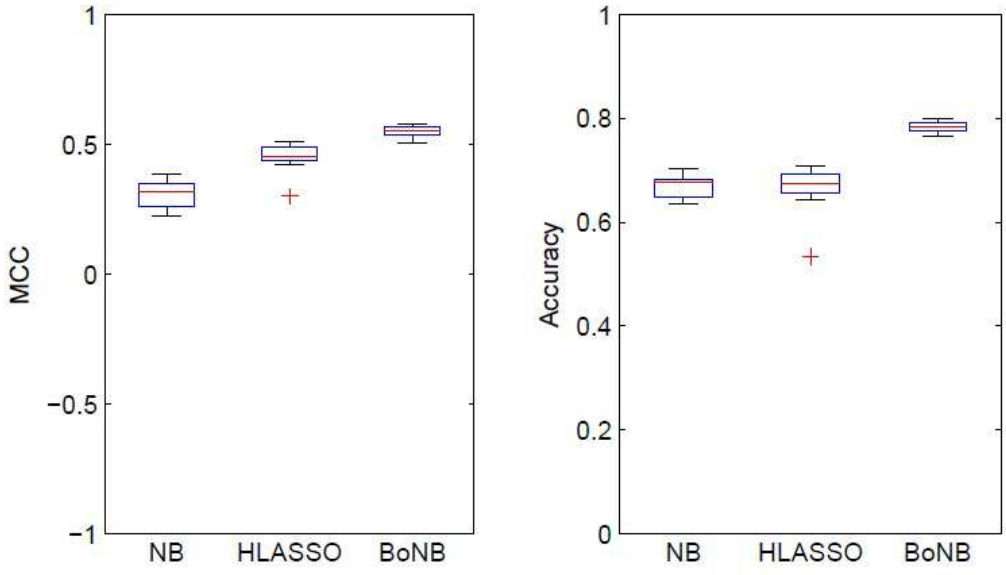


Figure 2.11: *Boxplots of MCC (left panel) and classification accuracy (right panel) of the simple Naïve Bayes classifier, HyperLASSO and BoNB on ten random sub-samplings of the WTCCC T1D dataset.*

To further analyze the behavior of the three methods at different levels of the output function (i.e. of the output class probability for BoNB and the standard Naïve Bayes classifier and of the logistic regression value for HyperLASSO) in Figure 2.12 the Precision vs Recall curve and the Receiver Operating Characteristic, or True Positive Rate vs True Negative Rate curve, of the three algorithms on one of the ten random sub-samplings are reported (the behavior on the other sub-samplings is similar). As it is clear from the figure, the performance of the standard Naïve Bayes classifier is completely dominated by the performance of both BoNB and HyperLASSO. Concerning the two latter algorithms, one can observe that HyperLASSO has a better performance at the two extremities of the curves, i.e. for subjects whose logistic regression value is closer to the maximum or the minimum; moving from the extremities to the middle scores, BoNB outperforms HyperLASSO, being indeed able to reach overall higher MCC and classification accuracy.

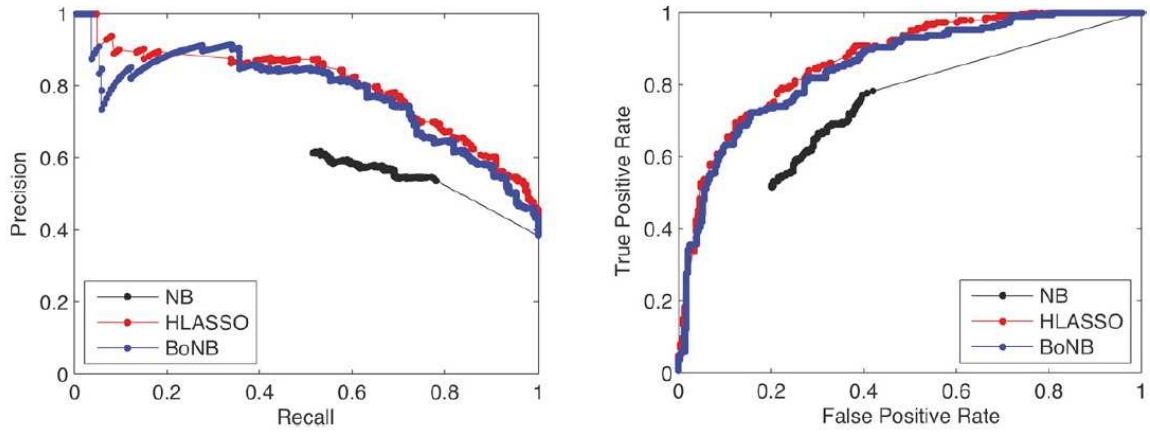


Figure 2.12: *Precision vs Recall curve (left panel) and Receiver Operating Characteristic (right panel) of the standard Naïve Bayes classifier, HyperLASSO and BoNB on a random sub-sampling of the WTCCC T1D dataset.*

For biomarker selection, BoNB was run on the whole dataset and its results compred with the biomarkers identified by HyperLASSO and by the general 2*df* test (see section 2.3.1). The average number of attributes included by BoNB in each NBC was 3.24, 75 SNPs were included by at least one NBC and 9 SNPs by at least 5% of the NBCs (see Table 2.2). Among the 9 SNPs, only 7 SNPs reached the significance level on the permutation test and were chosen as genetic biomarkers (marked in bold in Table 2.2). All the 7 selected SNPs fall into regions of interest for Type 1 Diabetes according to the on-line database T1DBase [http://www.t1dbase.org] (cytobands p13.2 on chromosome 1 and p21.32 on chromosome 6, also known as the MHC region) and their association with the disease was confirmed in a larger meta-analysis, subsequent to the WTCCC study [7]. The squared correlation coefficients between all pairs of selected SNPs are all lower than 0.155, indicating low redundancy in the information coded by the set of 7 SNPs. Compared to the 394 SNPs that reached the significance level on the 2*df* general test, both the list of 75 SNPs used for classification and the list of 7 biomarkers selected by BoNB are more compact, but this does not prevent BoNB to reach significantly higher classification performance.

HyperLASSO selected 8 SNPs, all in the MHC region of chromosome 6: 4 of the SNPs are in the list of biomarkers selected by BoNB, thus suggesting a certain coherence between the two algorithms and providing further confidence on the identified biomarkers.

SNP	Chr	Gene	Relation	% NBCs	MU (median)
rs6679677	**1**	**RSBN1**	**downstream**	**7**	**0.033**
rs9266774	6	MICA	upstream	5.5	0.011
rs805301	**6**	**BAT3**	**intron**	**17.5**	**0.043**
rs492899	**6**	**SKIV2L**	**intron**	**8.5**	**0.025**
rs9273363	**6**	**HLA-DQB1**	**downstream**	**100**	**0.835**
rs9275418	**6**	**HLA-DQB1**	**upstream**	**80**	**0.160**
rs6936863	**6**	**HLA-DQA2**	**upstream**	**8**	**0.08**
rs9784858	6	TAP2	intron	5	0.008
rs3101942	**6**	**LOC100294145**	**exon**	**21.5**	**0.045**

Table 2.2: *SNPs selected as attributes for at least 5% of the Naïve Bayes Classifiers by BoNB on the WTCCC T1D dataset, with B = 200 Bootstrap samples and classifiers. First column: dbSNP RS ID. Second column: SNP chromosome. Third and fourth column: annotated gene and relation with the SNP. Fifth column: percentage of Naïve Bayes Classifiers that included the SNP as attribute. Sixth column: median of the marginal utility of the SNP. SNPs selected as genetic biomarkers by the permutation procedure are marked in bold.*

2.5.4 Sensitivity analysis

As already pointed out in the results section, the BoNB algorithm exposes two parameters to the user: the number of Bootstrap replicates and Naïve Bayes Classifiers, B, and the threshold on the squared correlation coefficient above which two SNPs are considered correlated, θ. In this section a brief analysis to describe how performance is affected by variations of the parameters B and θ was carried out.

Figure 2.13, left panel, represents the *MCC* obtained by BoNB on ten random sub-samplings of the WTCCC T1D dataset, for $B = 200$ and θ ranging from 0.02 to 0.5. As it is clear from the figure, $\theta = 0.1$ is optimal and results in a significantly higher classification performance (Kruskal-Wallis test *p-value* 3.7×10^{-4}).

Concerning the number of Bootstrap replicates B, on the other hand, one can observe from Figure 2.13, right panel, that classification performance is not much sensitive to variations of B (Kruskal-Wallis test *p-value* 0.98), though it is slightly higher for $B = 50$ and 200. Analyzing the list of selected biomarkers, BoNB returns the same seven biomarkers reported in Table 2.2 for $B = 200$ and 500, adds SNP rs2856688 to the list for $B = 100$ and misses SNPs rs6679677 and rs492899 for $B = 50$. Given the consistency among the results for higher values of B, suggested values for BoNB parameters are thus $\theta = 0.1$ and $B = 200$.

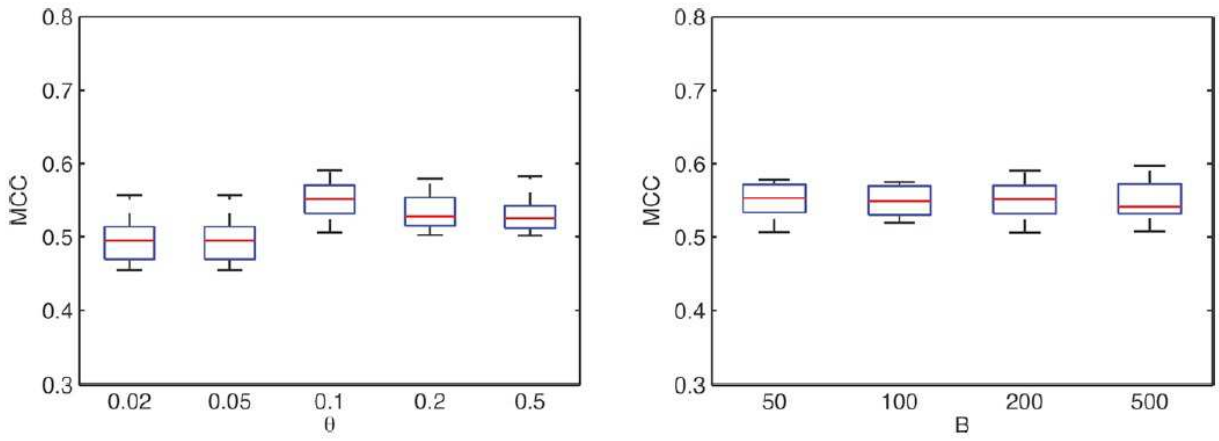


Figure 2.13: *Box plots of the MCC obtained by BoNB on ten random sub-samplings of the WTCCC T1D dataset, for B = 200 and ϑ ranging from 0.02 to 0.5 (left panel) and for ϑ = 0.1 and B ranging from 50 to 500 (right panel).*

2.5.5 Computational complexity

For analyzing the computational complexity of BoNB, the pseudocode summarizing the training phase and the biomarker selection phase of the BoNB algorithm is reported in the following:

// Training

for b = 1 **to** B
  [$X^{(b)}$, $OOB^{(b)}$] = bootstrap replicate from X
  for s = 1 **to** p
    Compute the contingency table for SNP s from $X^{(b)}$
    Compute the Naïve Bayes attribute score of s
  $L^{(b)}$ = list of SNPs in decreasing order of score
  Initialize $NBC^{(b)}$ as a Naïve Bayes Classifier with no attributes
  Extract M = 1 new attributes for $NBC^{(b)}$ from the top of $L^{(b)}$, excluding from future additions all SNPs at distance > 1 Mb and with $r^2 < \theta$
  while MCC of $NBC^{(b)}$, tested on $OOB^{(b)}$ with the new attributes, increases
    Add the new attributes to $NBC^{(b)}$
    Update M = 2 * M
    Extract M new attributes from the top of L, excluding each time from future additions all SNPs at distance > 1 Mb and with $r^2 < \theta$

// Biomarker selection

for s **in** all SNPs selected by at least 5% of the NBCs
  for b **in** all NBCs that selected s
    Permute the genotype of s in $OOB^{(b)}$
    Record the Marginal Utility (MU) of s
    Select as biomarkers the SNPs with MU significantly larger than zero.

For each b in B, the attribute ranking step (lines 3-6) takes $O(np)$ for computing the contingency tables and the scores (where n is the number of subjects and p is the number of SNPs in the dataset) plus $O(p\ log\ p)$ for sorting the score list, thus has a total complexity of $O[Bpn + Bp\ log\ p]$. The attribute selection step (lines 7-12), executed for each b in B, has a computational complexity dominated by two operations: computation of the squared correlation coefficient r^2 between SNPs and test of $NBC^{(b)}$ on $OOB^{(b)}$. Defining with M^* the average number of attributes included by each NBC (which is problem dependent) and p^*_{1Mb} the average number of SNPs in a 1 Mb section of the DNA (which is dataset dependent, but is a roughly linear function of p), the first operation costs $O(n)$ for each SNP pair and is executed $M^* \cdot p^*_{1Mb}$ times, having thus a total computational complexity of $O(BnM^*p^*_{1Mb})$. The second operation, on the other hand, is executed $\log(M^* + 2)$ times, each time with a doubling number of features for $NBC^{(b)}$, and its computational complexity is thus expressed by the following summation:

$$\sum_{i=0}^{log(M^*+1)} n^*_{OOB} \cdot 2^i = n^*_{OOB} \cdot \left(2^{log(M^*+2)} - 1\right) \cong O(nM^*) \qquad (2.21)$$

where n^*_{OOB} is the average number of subjects in an OOB set, tending to $(1 – 1/e) \cdot n$ for large n [20]; the total complexity of the second operation is thus $O(BnM^*)$, asymptotically negligible with respect to the cost of computing the squared correlation coefficients. The total computational complexity of the training phase of the BoNB algorithm is thus $O[B(pn + p\ \log\ p + nM^*p^*_{1Mb})]$. For the complexity of the biomarker selection phase of BoNB, the number of SNPs selected by at least 5% of NBCs (which is problem dependent) is defined as $p^*_{5\%}$. The inner loop of lines 15-16 is executed at most $O(B\ p^*_{5\%})$ times; since the cost of the two operations in the loop is linear in n, the biomarker selection phase has a total computational complexity of $O(Bn\ p^*_{5\%})$.

2.5.6 Implementation

BoNB is implemented in C++ and relies only on standard libraries, thus being fully portable across operating systems. On the WTCCC case-control study on Type 1 Diabetes, BoNB takes approximately 50 minutes for training 200 NBCs and selecting the biomarkers on a 3.00 GHz Intel Xeon Processor E5450. A careful allocation strategy makes BoNB occupy around 600 MB of RAM for the WTCCC dataset, allowing it to be easily run on a desktop computer.

2.6 *Discussion*

In this chapter the problem of modeling the effect of genotype on the outcome was discussed. In the context of Genome Wide Association Studies, the objective of such a modeling is twofold: on the one hand, *GWAS* aim to perform biomarker selection detecting correlation between one or more SNPs and a discrete trait (the presence or absence of a disease condition or a complication), on the other the modeling process allows also to learn a rule for classifying unknown subjects as cases or controls.

For complex diseases this is not trivial, since such pathologies have an heterogeneous nature, and to fully capture the optimal set of genetic biomarkers, all the SNPs in a *GWAS* should be analyzed simultaneously in a multivariate framework. Moreover, linkage disequilibrium confounds the search for genetic biomarkers, because of the non-random association between the true genetic causes and the SNPs in genomic regions close to them, thus resulting in a lack of precision and stability of the lists of biomarkers selected.
The standard approaches generally analyze one SNP at time, thus losing information on biomarkers interaction and suffering for statistical significance of the selected features. Multivariate approaches try to overcome these limitations, but the most widely used methods in the literature still suffer for the problem of robustness of the list of selected biomarkers. In fact, it is in principle possible to have a lack of stability due to the presence of many highly correlated features, even with classification accuracy equal to one.
The presented algorithm, Bag of Naïve Bayes, was developed to effectively tackle this problem.
BoNB is based on Naïve Bayes classification enriched by three main features to tailor the Naïve Bayes framework to Genome Wide SNP data analysis: (a) a bagging of Naïve Bayes classifiers, to improve the robustness of the predictions, (b) a novel strategy for ranking and selecting the attributes used by each bagged classifier, to enforce attribute independence, and (c) a permutation-based procedure for selecting significant biomarkers, based on their marginal utility in the classification process.
Learning an ensemble of classifiers from a bootstrap sample of the original dataset guarantees a higher generalization ability by increasing the stability of the learning process [12]-[18] and, simultaneously, it allows to define a measure of the marginal

utility of each SNP, given all the other SNPs exploited for classification, and to select significant biomarkers among these SNPs in a statistically robust way.

Two features of the Naïve Bayes Classifier, chosen as building block of the BoNB algorithm, make it more appealing for genome-wide data analysis than logistic regression approaches: on the one hand, conditional probability table analysis does not assume a pre-specified model of genetic effect, on the other hand, missing values are seamlessly handled by both the learning and the classification procedure.

BonB approach to attribute selection, consisting in a univariate ranking step followed by a multivariate selection step, has the advantage of favoring informative attributes, but without the need of pre-selecting fixed sets of attributes or of defining cut-offs on the strength of the association with the disease: attributes, in fact, are added to the classifiers as long as their combined effect on the generalization ability increases.

The effectiveness of BoNB was demonstrated by applying it to the WTCCC case-control study on Type 1 Diabetes: BoNB indeed outperforms two algorithms from the state of the art, namely a Naïve Bayes Classifier and HyperLASSO, in terms of classification performance and all the genetic biomarkers identified by BoNB are meaningful for Type 1 Diabetes, thus confirming the good performance also in terms of precision of the selected biomarkers.

Chapter 3

Modeling the combined effect of phenotype and treatment on the progression of diabetes complications

Referring to the multi-level scheme of Figure 1.1, this chapter will focus on the combined effect of phenotype and treatment on the outcome, as shown in Figure 3.1.

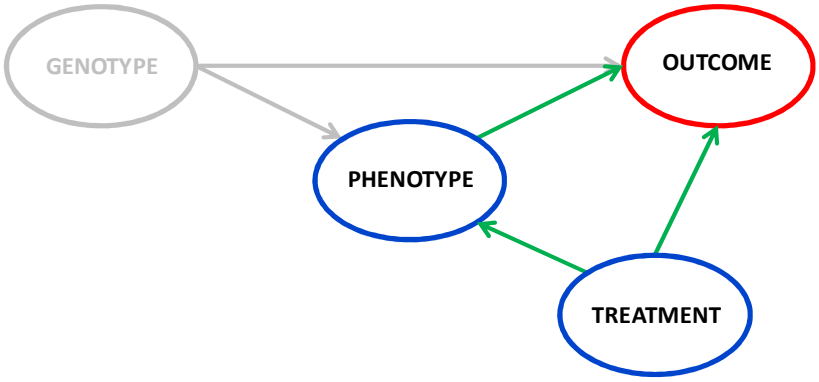


Figure 3.1: *Modeling the effect of genotype, phenotype and treatment on the outcome.*

After a brief overview describing the most interesting models already developed in the literature to model the progression of diabetes complications, a new *in-silico* model, based on Dynamic Bayesian Networks and accounting for phenotypic information as well as information on treatment, will be presented. Validation of the model on the Diabetes Control and Complications Trial dataset will be then reported and discussed. Finally, the under development web interface as a decision support tool for clinicians will be presented.

3.1 *Introduction*

In complex disease, such as diabetes mellitus, the development of complications and their impact on costs are difficult to assess through short-term studies. Since long-term clinical trials are costly, time-consuming, and difficult to conduct, the use of computer-simulate disease models has increased considerably in recent years to facilitate the simultaneous evaluation of long-term clinical end economic effects of treatment [87],[71]. It is now widely accepted that models can provide valuable information for clinical practice and are important tools in medical, regulatory, governmental and public health decision-making [82],[90]].

For example an *in-silico* model of chronic disease can be used as a tool to simulate a clinical trial based on the available medical literature and publicly available data sources. Even in situations where a clinical trial does exist, models are often used to incorporate the benefits and costs beyond the time horizon of the trial or to consider all the available options simultaneously [71]. A good example of the former statement is the cost-effectiveness analysis alongside the Scandinavian Simvastatin Survival Study [69], where the authors used a previous model of Coronary Heart Disease to project beyond the five-year horizon of the study. An example of the latter statement is the supplement that strategies of annual analysis of fecal occult blood testing (FOBT) with five-year sigmoidoscopy, recommended by the American Cancer Society [60], could bring to an analysis of a clinical trial of annual FOBT versus no testing, such a s the Minnesota Colon Cancer Control Study [73]

In-silico models of complex diseases are exploited to predict the evolution (i.e. the appearance of events or the persistence in a state devoid of severe complications) of an individual (or a population), providing a probability distribution for the individual (or the population) to develop a certain complication.

The aims of complication models do not limit to predict time courses. It is of interest also evaluating possible variations of the quality of life during the lifetime that is predicted for a patient and the costs that the treatments that are administered to him require, since multiple treatments are often possible for the same disease [90]. The choice of the best strategy involves the evaluation of both clinical outcomes and costs of the different

available strategies. The formal process for comparing available alternative strategies is called *decision analysis*.
The final aim of an *in-silico* model is to support decision analysis, helping clinicians in taking the best choice among the available ones [90].
The steps of constructing, evaluating and interpreting the model are done iteratively to give a progressive refinement.

A requirement for diabetes simulation models has been identified in the medical and healthcare policy community, and, as a result, a number of models, mainly based on the Markov Models, have been developed and reported in the literature [79], [63], [74]. In the next section, a brief introduction on Markov Models and a rapid overview on the available models of diabetes complications will be presented.

3.2 *Markov Models*

As concerns mathematical aspects, the complication models of major importance are all composition of Markov models, each of them representing a complication.

A Markov Model (MM), also called state transition model, is used to represent *recursive* events [85]. Discrete MMs enumerate a finite set of mutually exclusive possible states such that, in any given interval of time (called a cycle or *stage*), an individual member of the Markov cohort is in only one of the states.

A Markov model is a stochastic model that assumes the Markov property, which is the following memoryless property: the state of the system at time instant *t* depends only on the state of the system at time instant *t-1*; in other words, it does not depend on previous time instants.

A set of *initial probabilities* is used to specify the distribution of the cohort (group of individuals that is homogenous for a set of demographic and clinical aspects) among the possible states at the beginning of the process. A matrix of *transition probabilities* is used to specify the transitions among states.

In the original graphical representation of a MM (Figure 3.2), sometimes called “bubble diagram”, each state is represented using a circle while arrows represent transitions from a state to another one.

A transition arrow pointing back to the state from which it originates indicates that it is possible for a cohort member to remain in the same state for more than one stage. The numbers along the arrows indicate the transition probabilities. The probabilities of the transition arrows emanating from any state must sum to 1.

Each complication is represented by a Markov model similar to that in Figure 3.2:

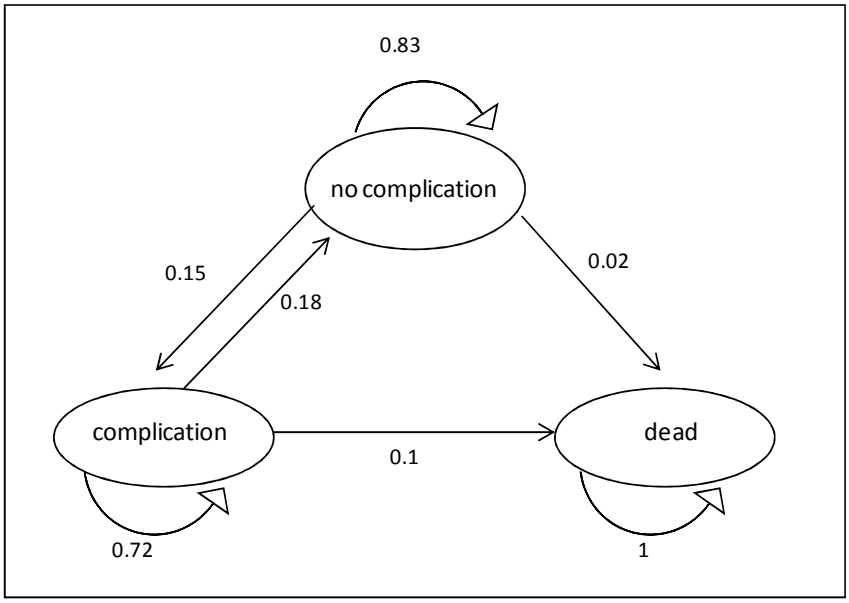


Figure 3.2: *3-state Markov model for a generic complication. Circles represent possible states, i.e. clinical conditions that can characterize a person. Arrows indicate possible transitions.*

The model of Figure 3.2 is characterized by 3 states: “No complication” representing diabetic people without any severe complication; “Complication” representing the people that reach the considered endpoint; “Dead” represents death caused by the complication.

Therefore, Markov models allow representing the evolution over time of a diabetic population that is often simulated with time step of 1 year. The Markov models allow simulating over time the evolution of a cohort of patients in its mean behaviour and performing individual-level simulations. The first one is generally called expected-value simulation and is based on a deterministic approach, i.e. if the probability of transition

from state “No Complication” to state “Complication” is p=0.03, then the 3% of patients in the former will transfer to the latter at next 1-year step of the simulation. On the opposite, the second is a stochastic simulation where the single *in silico* individual transfers only if the number provided by a random number generator is smaller than transition probability p. That is why this kind of simulation is also named Monte-Carlo.

Obviously, also a high number of single *in silico* individuals can be considered and the results of these multiple simulations can be averaged. Theoretically, higher is the number of individual-level simulations, more similar are their average results to the expected-value predictions.

On the other hand, a high number of individual-level simulations allows quantifying the variability in model outcomes resulting very helpful to establish the reliability of average predictions and of the expected-value simulation.

For instance, the evolution over time of an *in silico* population obtained with a toy-model is reported in Figure 3.3:

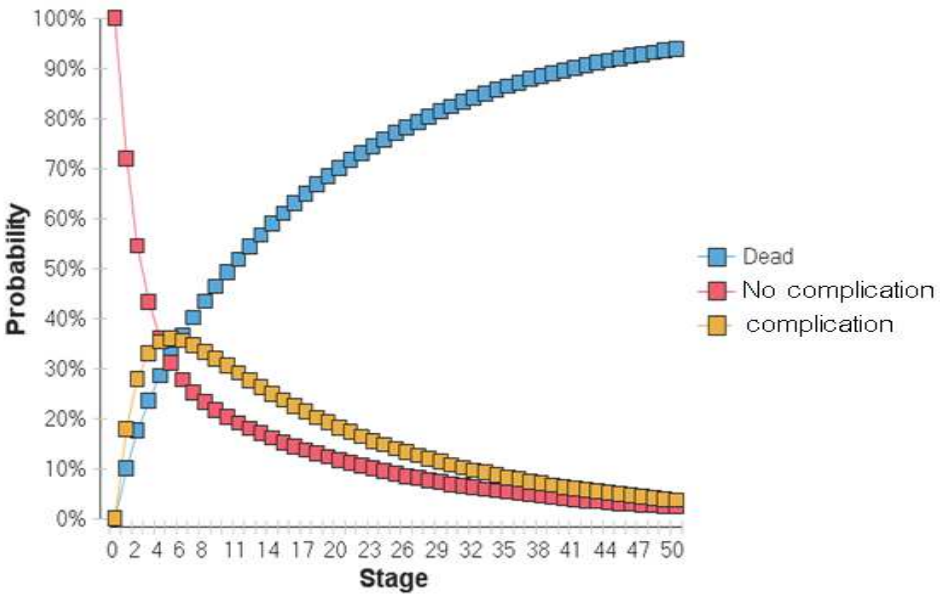


Figure 3.3: *Evolution over time of a diabetic population obtained with a toy, 3-state Markov model.*

Automatically, a survival curve can be obtained from these simulations:

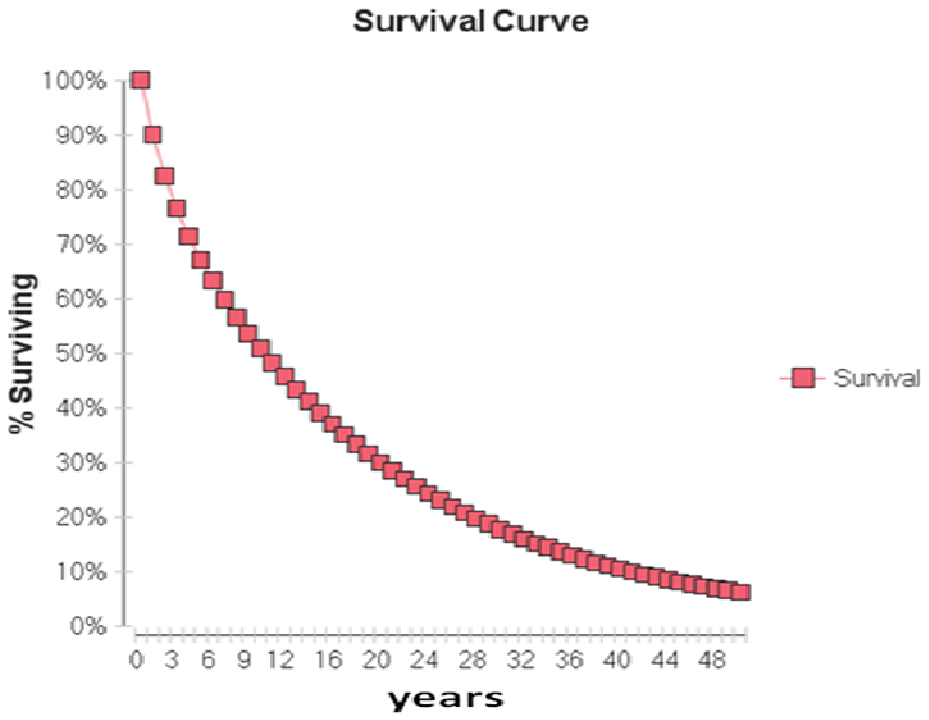


Figure 3.4: *Survival curve obtained from the simulations of Figure 3.3.*

Transition probabilities are estimated from the data collected during clinical trials. Wrong values of these probabilities mean wrong predictions; this is why the availability of high quantity of data, derived from homogenous cohorts, is critical.

Simulations do not limit to the predictions of events but extend to the time course of risk factors. When data are available, risk factors worsening is based on them, otherwise a gradual worsening is usually implemented. This latter can be slowed by treatments, whose administration can be implemented in the software. Since treatments are characterized by a rate of failure, also this occurrence is sometimes inserted in the model together with the possibility to administer multiple and subsequent, distinct treatments.

The most interesting diabetes complication models that were developed thanks to data collected by clinical trials are the Palmer model [79], the Eastman model [63], and the EAGLE model [74], which will be described in the following.

3.2.1 The Palmer model

One of the most known complication models is the Palmer model (also called CORE model), which is based on multiple interconnected sub-models: one for each of the considered complications [79], [80], [81]. All the sub-models are Markov models characterized by 2, 3 or more states depending on the specific modeled complication. The dependences among events are generally expressed by changes in the worsening rate of risk factors or in the values of transition probabilities.

The considered complications or endpoints are: myocardial infarction, angina, heart failure, stroke, peripheral vascular disease, neuropathy, foot ulcer, macular edema, cataract, limb amputation, blindness, microvascular complications (retinopathy, vitreous haemorrhage, nephropathy and renal failure), hypoglycemia, ketoacidosis, and lactic acidosis, to which add "surrogate endpoints", i.e. stages of disease worsening that allowed monitoring precisely the evolution over time of neuropathy, orthostatic hypotension, impotence.

Palmer model is also one of the most appreciated models since it is based on original data derived from the most recent databases. However, model predictions do not base only on these clinical database but on treatment and economics databases. The former stores data on treatment pathways, treatment effects and on the change in each physiological parameter in the simulation, as a consequence of treatment or patient management. The latter is used to perform economic analysis, i.e. to evaluate the cost of patient treatment (with and without complication) over the considered time period. In addition, the economics database allows evaluating the quality-adjusted life years, which briefly consist in life expectancy corrected for the quality of life.

Palmer model allows simulating both the time evolution of a cohort of patients in its mean behavior and the individual-level simulation.

As regards transition probabilities, they are derived from event rates registered during the clinical trials. However, probabilities can depend also on some risk factors [79].

3.2.2 The Eastman model

As Palmer model, also the Eastman model is based on Markov type models and exploits a Monte-Carlo approach to simulate possible complication events in single *in silico* individuals [62], [63]. Again, as Palmer model, multiple sub-models are present, each one

for a single complication. Considered endpoints are: retinopathy, nephropathy, neuropathy and cardiovascular disease.

A high number of the implementation principles characterizing this model were taken up by Palmer and the EAGLE modelers and this highlights the importance of the Eastman model in the history of complication model development. In particular, a methodology to estimate incidence rates, to be used as transition probabilities of the Markov models, from the cumulative incidence observed during the survey, was detailed. The presented reasoning that is based on the fit of an exponential model to collected data laid the fundamentals for the use of more complicate models exploited by the following modelers, like the Cox proportional model [86].

3.2.3 The EAGLE model

The most recent model among the three reported here is the EAGLE model. It implements an object-oriented probabilistic Monte Carlo simulation, which is based on a Markov process with yearly intervals. Transition probabilities are dependent on the status of the simulated patient, with related calculations defined internally.

Twenty outcomes (e.g., hypoglycemia, retinopathy, macular edema, end stage renal disease, neuropathy, diabetic foot syndrome, MI, and stroke) are projected based on data from epidemiological and clinical trials.

The EAGLE model is capable of simulating the progression of type 1 and type 2 diabetes and this is the major difference with respect to previous models. In fact, the reader can find in this description the basic principles already enunciated for the older Eastman model. On the other hand, the model author’s made clear that the EAGLE was not developed on original data, but on a subset derived from previous people's publications. This is the main drawback of the model.

A systematic comparison of Palmer, EAGLE and other models is detailed in [78].

3.3 *Objective of the work*

As resulting from the overview presented in the previous section, models able to integrate accumulating –omics knowledge (metabolomics, proteomics, genomics) into a clinical macro-level for multifactorial diseases are still missing and, so far, the most interesting

complication models, developed from data collected by clinical trials, are based on Markov Models and use only phenotypic variables as features to describe the cohort of interest.

The aim of the present work is to model the progression of diabetes vascular complications relying on the powerful framework of *Dynamic Bayesian Networks* (DBNs). DBNs provide a more flexible structure with respect to Markov Models, and allows to easily enlarge the model with additional information. This is why such a model will represents a solid base for future developments, such as the enrichment with genotypic information, as it will be discussed at the end of the chapter.

3.4 *Methods*

3.4.1 Bayesian Networks

Bayesian networks are now being used in a variety of applications. The interest in general instruments able to compute posterior probability distributions has been quite high in the Bioengineering and Biomedical Informatics community. As a matter of fact, DBNs allow dealing with a variety of crucial problems in biomedicine, ranging from classification to prediction, and from simulation to parameter estimation. Recently, Bayesian network approaches were successfully employed in the context of genome biology and in biomedical research [64]. One of the most common application is diagnosis problems, as in case of medical diagnosis. An example is PATHFINDER [67], a program to diagnose diseases of the lymph node by means of Bayesian network approach.

A Bayesian Network (*BN*) is a probabilistic graphical model that represents conditional dependence over a set of random variables in a compact and human-readable form. Probabilistic graphical model possess two important characteristics: i) they clearly express the conditional independence between the variables, thus allowing an intuitive but sound way to describe the assumptions underlying the modeling process; ii) they associate to the graph a probabilistic model that can be used for performing inference, and, thus, estimation, simulation and prediction [68].

A Bayesian Network is completely determined by a Directed Acyclic Graph (*DAG*), known also as the *network structure*, and by a set of conditional probability distributions: each node of the network corresponds to a random variable and each edge corresponds to a probabilistic dependence between the two nodes (the terms *node* and *random variable* are used as synonyms).

More precisely, a Bayesian Network represents a joint probability distribution between its nodes for which the *Markov condition* holds: any node in a Bayesian Network is conditionally independent of its non-descendants, given its parents.

The Markov condition implies that the joint probability distribution of the nodes can be decomposed as

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P\left(X_i \middle| Pa_{X_i}\right). \qquad (3.1)$$

Where Pa_X denotes the set of parents of X: whenever we have an edge $X_i \rightarrow X_j$, we say that X_j is a *child* of X_i and that X_i is a *parent* of X_j.

The decomposition of equation (3.1) is called *chain rule for Bayesian Networks* and allows a more compact representation of the full joint probability distribution, requiring fewer parameters to be completely specified: the probability distribution of each node can in fact be expressed simply as a function of the states of its parent nodes.

Figure 3.5 reports an example of discrete Bayesian Network with 4 nodes, modeling the hypothetical probabilistic relations between the variables *HighFatDiet* (*HFD*), *GlucoseTolerance* (*GT*), *Obesity* (*OB*) and *RiskOfCardiovascularDisease* (*CVD*). As it is clear from the figure, the probability distribution of each node is expressed as a function of all possible combination of values of its parents, in the form of a c*onditional probability table* (*CPT*). A natural application of this network is to use it as instrument to compute posterior probability distributions, i.e. the posterior probability of any of the problem variables given knowledge about any of the other variables of the problem. This theme is called *inference*. For example, such a network can be used to answer queries like: “What is the probability of being obese, if on a high fat diet?”, “What is the probability of having impaired glucose tolerance, if at risk of cardiovascular disease?”, “If obese and on a high fat diet, what is the probability of being at risk of cardiovascular

disease?". Moreover, the Markov condition can be used to infer conditional independence relations from the network. For example, we can infer from the network structure that, once the values of *GlucoseTolerance* and *Obesity* are known, *RiskOfCardiovascularDisease* becomes independent of *HighFatDiet*.

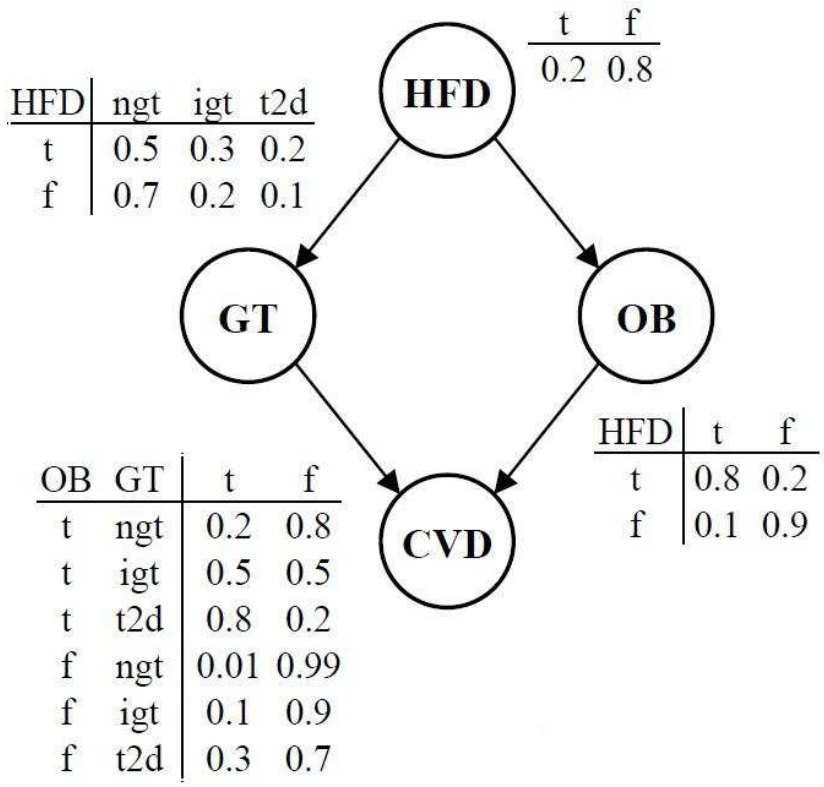


Figure 3.5: *Example of a simple Bayesian Network with 4 discrete variables, representing the interactions between High Fat Diet (HFD), Glucose Tolerance (GT), Obesity (OB) and Cardio Vascular Disease (CVD).*

The example in Figure 3.5 is a *BN* with discrete variables, i.e. variables with a finite number of possible values. Conditional distributions in discrete variable *BNs* can be conveniently represented with probability tables and are able to model dependencies between variables without making any assumption on the underlying relationship (e.g. linearity). Many real-world variables are of a continuous nature (e.g. blood glucose concentration or gene expression levels). In these cases, a possible solution is to discretize these variables and resort to discrete *BNs*. In some cases, though, discretization would lead to a major loss of information, unless a high number of discrete states is employed,

which would significantly increase model complexity. The other solution is to employ continuous-variable *BNs*. The general *BN* framework so far presented holds for both discrete and continuous variables, as long as the conditional distribution $P(X|Pa_X)$ assigned to each node represents for each possible value pa_X of Pa_X a distribution over *X*. When all variables in the network are continuous, the most commonly employed distribution is a linear Gaussian distribution model. Given the continuous variable *Y* with continuous parents $X_1, \ldots, X_k$, the probability density of Y as a function of its parents is:

$$P(Y|x_1, \ldots, x_k) = N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k; \sigma^2) \qquad (3.2)$$

This simple model can be extended to cases in which the mean of *Y* depends on its parents in a nonlinear way, or in which the variance also depends on the parent values. Hybrid models are also possible, which incorporate both discrete and continuous variables.

Both the graphical structure of a *BN* and the parameters of the conditional probability distributions can be learned from the available data. However, learning these networks is often non-trivial due to the high number of variables to be taken into account in the model, with respect to the instances of the dataset.

Structure learning of a *BN* is NP-complete problem in the general case: given a dataset $\mathcal{D}$, containing multiple samples of a set of random variables, the objective is to find the best, or the most probable, *BN* structure in the exponential space of all possible structures. Several scoring functions have been proposed to assess the quality of a *BN* structure: some of the most notable are the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Bayesian Dirichlet equivalent (BDe) [67]. Regarding the learning approaches, from the vast literature, three main approaches can be identified to *BN* structure learning: greedy search, complete search and search based on independence tests. Briefly, *greedy search* attempts to construct a *BN* structure starting with a network without any edge and iteratively adding the "best" set of parents to each node, according to a local score; *complete search*, conversely, explores the entire space of possible networks and is guaranteed to return an optimal network, but the huge memory and time requirements limit its application to small sized networks; approaches based on *independence tests* start with a complete network and aim at forbidding as many edges as possible, by assessing conditional independence between variables with statistical tests.

In the present work, an approach based on independence tests evaluated by a Bayesian Dirichlet equivalent with uniform priors (BDeu), has been adopted, as described in section 3.4.5.3.

3.4.2 Dynamic Bayesian Networks

Being interested in modeling the history of diabetic patients, the dynamics of the disease will be explicitly modeled by relying on Dynamic Bayesian Networks (DBNs).

DBNs are an extension of BNs that represent the temporal evolution of variables over time. Nodes in the directed acyclic graph associated with a DBN continue to represent random variables, while edges represent temporal dependencies. The key assumption is that the probability distributions describing the temporal dependencies are time invariant so that the overall temporal evolution of the analyzed process can be entirely reconstructed by knowing the temporal dependencies represented in the DBN graph [83].

Figure 3.6 shows an example of DBN describing the evolution of the expression values of three genes *G1*, *G2*, and *G3*. The graph shows that the expression value of each gene at time (t + 1) is assumed to depend on the gene's expression at time t as well as on the expression of one or two of the other genes. Furthermore, the example shows that the temporal dimension of DBNs allows encoding feedback regulation such as the one occurring between *G1* and *G2*, which is not possible in static BNs because of the required acyclicity of the graph. The example in the Figure is a DBN of order 1, as all temporal dependencies occur between consecutive time points; yet DBNs are not restrict to dependencies of order 1 but can represent also higher order dependencies.

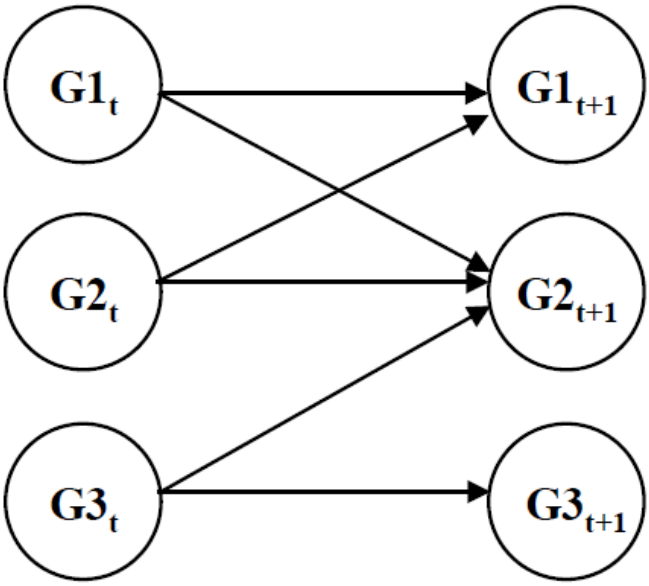


Figure 3.6: *Example of a Dynamic Bayesian Network for three genes G1, G2 and G3.*

DBNs are advantageous with respect to MMs since each variable is represented by one node, whereas MMs require as many nodes as the number of combinations of variable values [75], [92]. Thus, extending the DBN model with the addition of new variables just requires adding as many nodes.

3.4.3 Model general structure

The general scheme of a DBN for modeling the dynamics of a complex disease such as diabetes is represented in Figure 3.7.

In the scheme of Figure 3.7, the input variables are called *covariates* (referred to as *U*), while the output variables are called *outcomes* (referred to as *Y*). In particular, the time depending covariates are called *dynamic* (referred to as *Udyn*), while the remaining covariates are called *static* (referred to as *Ustat*), since they are not time dependent. To be precise, static variables are either constant or time varying, but their variation across time is completely predictable (e.g. age, which deterministically increases of 1 year every time step). All the outcomes are time-dependent.

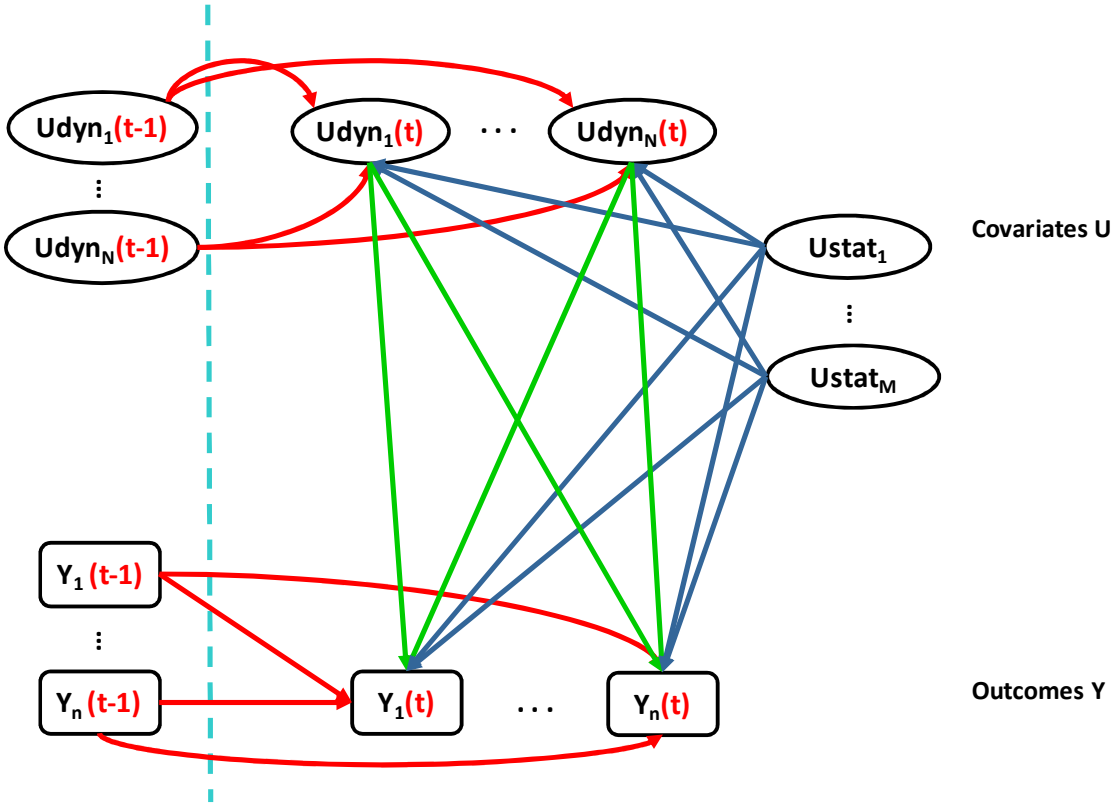


Figure 3.7: *General scheme of a Dynamic Bayesian Network for the dynamics of a multifactorial disease*

The dependences between variables can be summarized as follows:

- Each dynamic covariate at time *t* can depend on each other dynamic covariate at time *t-1* and on each static covariate;
- Each outcome at time *t* can depend on each other outcome a time *t-1*, on each dynamic covariate at time *t-1* and on each static covariate.

This network represents the *a priori* structure of the model.

The specific structure of the model, i.e. the set of variables used as nodes of the network and edges representing their conditional probabilities, depends on the information contained in the available data, and will be presented in section 3.5.

3.4.4 Data

3.4.4.1 Datasets

Databases collecting data over more than ten years allow estimating the event rates that are basic for the development of complication models. Therefore datasets represent the fundamental starting point for any predictive model. Three of them, i.e. the Framingham

Heart Study database, the Diabetes Control and Complications Trial (DCCT) database, and the UK Prospective Diabetes Study (UKPDS) resulted particularly important for the field [61], [89] and have been used to implement predictive models described in the previous section.

The Framingham Heart Study is a cornerstone of epidemiological studies and, after more than 50 years from its beginning, it remains the most famous and influential investigation in cardiovascular disease epidemiology. Nowadays, it is considered the epitome of a successful epidemiological research, productive of insights and applications and the prototype of the cohort study [77].

The DCCT was a multicenter, randomized clinical trial designed to compare intensive with conventional diabetes therapy with regard to their effects on the development and progression of the early vascular and neurologic complications of insulin-dependent diabetes mellitus [61].

Similarly, the UKPDS was designed to establish whether, in patients with type 2 diabetes, intensive blood-glucose control reduced the risk of macrovascular or microvascular complications, and whether any particular therapy was advantageous [89].

The three databases differed for many aspects, among which the pathologies of interests, the surveyed patients, and the duration of the study. However, all of them allowed establishing the importance of some clinical factors for the development of micro/macrovascular complications in the long period, discarding others. In particular, the Framingham Study was the first one to suggest a relationship between diabetes and cardiovascular diseases, on the basis of statistical evaluations, laying the foundations for the subsequent two trials, more focused on diabetes.

Most of all, they share the approach that is based on a survey of the population of interest along time, periodical measurements of factors of clinical interest (systolic blood pressure, plasma insulin...), which are usually called “risk factors”, and on the effort of relating these latter with the observed incidence of micro/macrovascular events. That is why all of them were followed by predictive models of complications, which were developed on collected data.

A forth database that is often used for the implementation of complication models is the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR), which can provide additional information for the specific complication retinopathy [70]. Both Type I and Type II diabetes were considered in the survey allowing accomplishing two parallel studies.

With the objective to model the progression of diabetes complications modeling the combined effect of phenotype and treatment (and the future prospect of use genotypic information too), data needed to learn the model have to satisfy some precise requirements:

- the number of patients composing the cohort has to assure a robust learning;
- the dataset has to contain information about the main diabetic complications and the correlated events, as well as information on phenotypic variables and the main treatments;
- data have to be collected through a longitudinal study over a period of medium-long duration (e.g., ten years);
- the dataset has to contain also genetic information, in particular SNP data.

Among the available dataset previously described, only the DCCT satisfies all these requirements.

The next section provides a brief description of the DCCT dataset, mainly focused on the relevant characteristics for building the prediction model.

3.4.4.2 DCCT/EDIC description

The Diabetes Control and Complications Trial (DCCT, 1982-93) and the Epidemiology of Diabetes Interventions and Complications (EDIC, 1994-2006) follow-up study have been ongoing for more than twenty years [61]. The clinical trial and subsequent follow-up have provided the scientific community with invaluable information regarding the effect of glycemia and glycemic control on long-term diabetes complications.

The DCCT studied a cohort of 1,441 subjects between 13 and 39 years old which had suffered type 1 diabetes mellitus (T1DM) for 1-15 years at the time of recruitment. All participants were relatively healthy except for diabetes and were free of severe diabetes-related complications. The Primary Prevention cohort consisted of 726 subjects with

T1DM for 1-5 years and no diabetes related complications (no microaneurysms on fundus photography and urine albumin excretion <40 mg/day). The Secondary Intervention Cohort consisted of 715 subjects with T1DM for 1-15 years and mild to moderate non-proliferative retinopathy and a urinary albumin excretion rate <200 mg/day. Subjects were randomized to conventional (CT) or intensive diabetes therapy (IT) (Figure 3.8). The intent of IT was to achieve blood glucose levels of 70-120 mg/dL in the morning and before meals, <180 mg/dL after meals, and an HbA1c in the non-diabetic range (<6.05%). Although it was not feasible to achieve these glycemic targets consistently in the majority of the subjects assigned to the IT group (fewer than 5% maintained an average HbA1c <6.05%), there was a substantial difference in glycemic control between the IT and the CT groups. The CT group maintained an average HbA1c of about 9.0% (similar to their baseline value) throughout the 3-9 (mean 6.5) years of follow-up. Those in the IT group lowered their HbA1c to about 7.0% and maintained this for the duration of the study (Figure 3.9).

Following the end of the DCCT in 1993, and a transitional period during which the conventional treatment group was taught intensive therapy and the clinical care of all of the subjects was transferred to their own health care providers, an observational study of the DCCT cohort, entitled Epidemiology of Diabetes Interventions and Complications, was launched. The goal of the EDIC follow-up was to examine the longer term effects of the original DCCT interventions, especially concerning complications, such as cardiovascular disease and more advanced stages of retinal and renal disease, that require a longer period of time to develop. During the transition from the DCCT clinical trial to the EDIC observational study, the average difference in glycemic control, measured by HbA1c, that had been approximately 2% during the DCCT (7.2% in the intensive treatment group compared with 9.1% in the conventional treatment group) narrowed (7.9% vs. 8.1% in IT and CT groups, respectively). The difference in mean HbA1c between the two original treatment groups has become statistically indistinguishable during the most recent six years of EDIC follow-up. (Figure 3.9) Phase 1 of the EDIC follow-up study spanned twelve years. The total mean follow-up of the original cohort was approximately 16 (range 13-20) years. Retention of the DCCT cohort remained outstanding: 96% of the surviving DCCT cohort joined EDIC in 1994 and 94% of the

original cohort (n= 1357 of 1441) remained active throughout the first phase of EDIC (Figure 3.9)

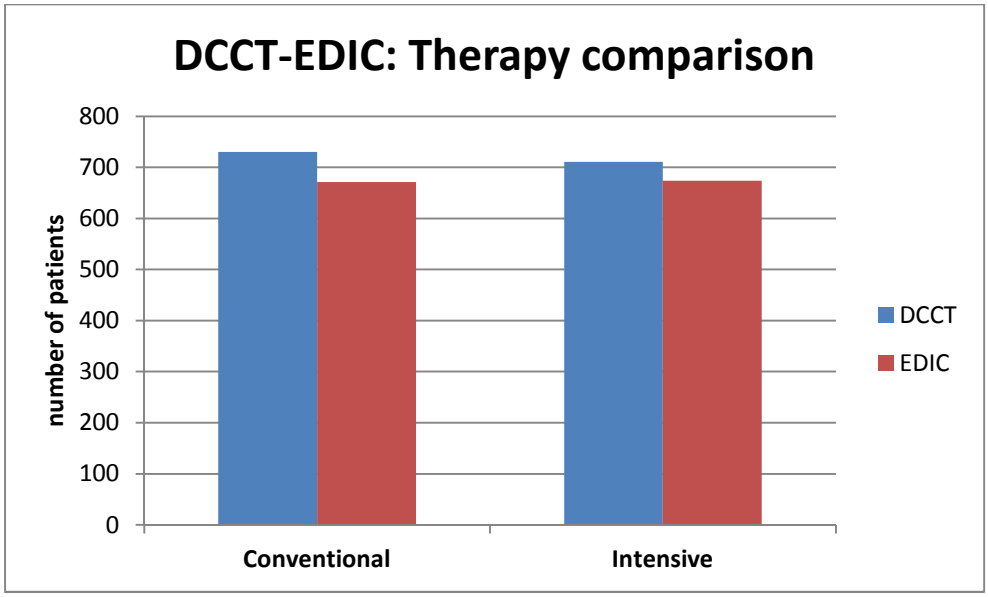


Figure 3.8: *number of patients for therapy during DCCT and EDIC studies. the notations "Conventional" and "Intensive" referred to EDIC, have to be meant as "EDIC patients who were treated with Conventional therapy during DCCT" and "EDIC patients who were treated with Intensive therapy during DCCT".*

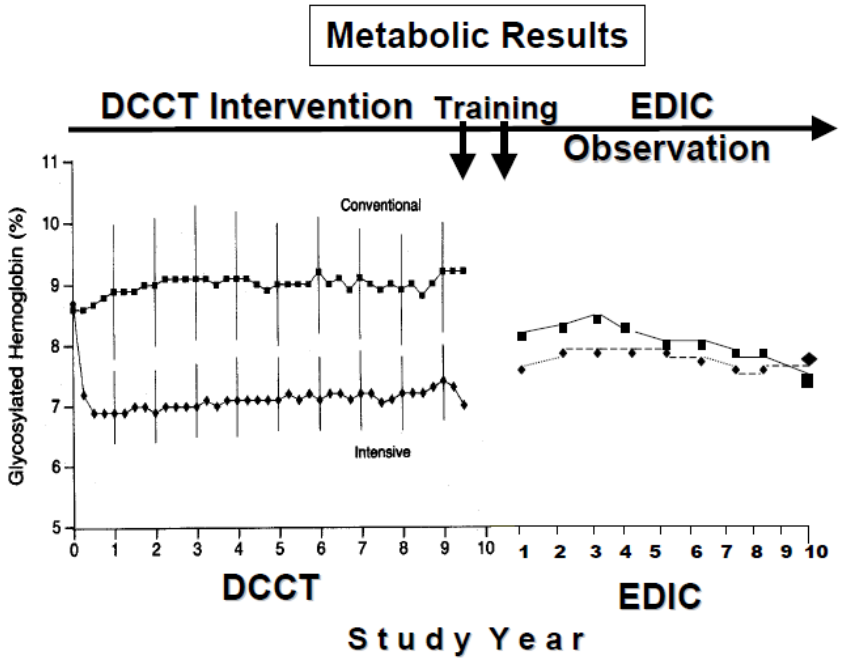


Figure 3.9: *Glycemic Levels during DCCT/EDIC as measured by glycosylated hemoglobin (HbA1c). Medians with 25th to 75th percentiles shown.*

3.4.4.3 Data analysis and preprocessing

Data from both DCCT and EDIC studies were used, thus having the possibility to train the model on a longer time period. Since the 1441 patients entered the DCCT in different years (thus having different follow-up time periods), for each patient the number of DCCT-years was computed by comparing the individual date of DCCT enrolling and DCCT close-out. For each patient, individual EDIC data were then appended at the end of the last DCCT year, as shown in Figure 3.10.

The computed mean follow-up period was 15.3 years.

Considering the set of the measures of all the variables for a single patient on a single year as an instance of the dataset, that instances for which all the dynamic covariates were missing have been discarded, in order to reduce the missingness of the dataset and thus avoidind the need for massive imputation. Thus, the number of available (or valid) years for each patient was computed. The mean value for the number of available years was 15.

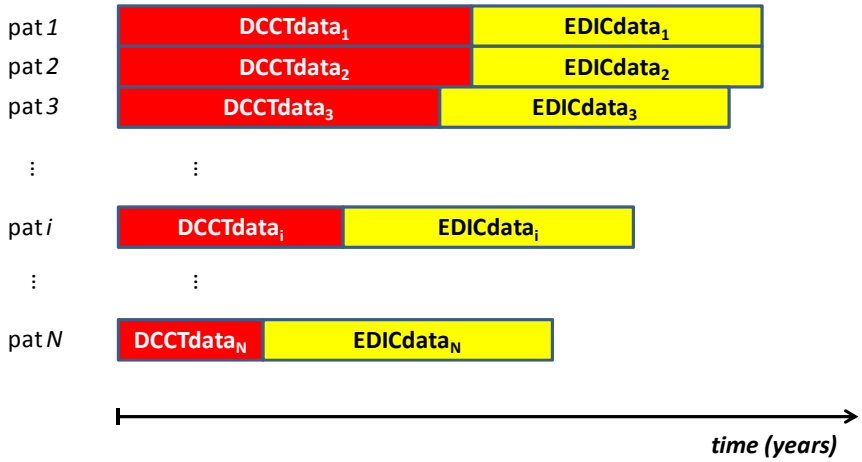


Figure 3.10: *For each patient, EDIC data were appended at the end of the DCCT data. The resulting mean follow-up period was 15.3 years.*

Relaying on previous literature works (see section 3.2) and on data availability in the DCCT/EDCI dataset, the following variables were used as covariates for the DBN model:

Static Covariates

This group includes both actually static variables, such as patient gender, and variables that vary in time but in a completely predictable way (e.g. age) or dependent by external decisions (e.g. treatment).

1. *SEX*:

   Patient's gender, assuming 2 possible discrete values: *Male* and *Female*

2. *AGE*:

   Patient's age in years

3. *DURATION*:

   Number of years since diabetes was first diagnosed

4. *TREATMENT*:

   Annual treatment received by the patient. This variable represents the type of treatment the patient received during the year ending with the current visit, and it can assume 3 possible discrete values: *Conventional treatment*, *Intensive self-treatment* (i.e. intensive treatment managed by the patient himself) or *Intensive forced treatment* (i.e. intensive treatment managed by clinicians). Treatment received during the DCCT period belongs to the *Intensive forced treatment* class, while treatment during the EDIC period belongs to the *Intensive self-treatment* class. This variable is considered an "intervention variable", i.e. an independent variable that is known to influence Glycosylated Hemoglobin (HbA_{1C}) value and thus is forced to contain this link in the DBN.

5. *THERAPY*:

   Number of years of diabetes not treated with intensive therapy, either forced or self-treatment. This variable initially represented the number of years of intensive therapy. Then, it was converted into a more informative variable, accounting for the total number of years of diabetes not treated with intensive therapy. The variable was computed combining information from the original variable *THERPAY* and the variable *DURATION*

6. *SMOKE*:

   Patient's smoking status, assuming 2 possible discrete values: *Never smoked* or *Ever Smoked*. The variable was initially a dynamic covariate, assuming 3 possible values according to the DCCT/EDIC codification: *Smoker* (if the patient was actually smoking at the visit time), *Non Smoker* (if the patient never smoked or quit smoking more than 3 months before the visit time), *Former Smoker* (if the patient had smoked but quit less than 3 months before the visit time). Since there's little difference

between the states *Smoker* and *Former Smoker* from a clinical point of view, this latter status was incorporated by the former. Given the high missingness throughout the study, the variable was then converted into a static covariate, computed as the mode of the available values for each patient, thus giving 2 possible states: *Never Smoked* or *Ever Smoked*

7. *EXERCISE*:

   Patient's physical activity level, assuming 3 possible values: *Sedentary*, *Moderate* and *Vigorous*. This variable was initially a dynamic covariate, assuming 4 possible values according to the DCCT codification and 3 possible values according to the EDIC codification: *Sedentary* (less than 5 hours of moderate activity per week), *Moderate* (more than 5 hours of moderate activity per week), *Vigorous* (more than 8 hour of hard activity per week) and *Strenuous* (more than 5 hours of very hard activity per week) in the DCCT dataset, and *Sedentary* (occasional physical activity), *Moderate* (considerable, but not constant, physical activity) and *Strenuous* (constant physical activity) in the EDIC dataset. In order to have uniform information, the third status of DCCT codification (*Vigorous*) was incorporated with the last one, thus giving the single state *Strenuous*, as in the EDIC dataset. A correspondence between homonymous states in the DCCT and EDIC codifications was then assumed. Given the high missingness throughout the study, the variable was then converted into a static covariate, computed as the mode of the available values for each patient.

Dynamic Covariates

8. *WHR*:

   Waist circumference to hip circumference ratio; information on WHR in the DCCT was available only at the screen visit, while annual measurements were available in the EDIC. This variable was thus imputed, for each patient, by performing a linear interpolation of the available values, thus obtaining the WHR time-course within the study.

9. *HBA1C*:

   Glycosylated Hemoglobin (HbA_{1C}) value, expressed as percentage of the total hemoglobin. HbA_{1C} is strictly connected to diabetes mellitus, since it is a form of hemoglobin that is correlated to the average plasma glucose concentration over

prolonged periods of time, thus serving as a marker for average glycemic values. In diabetes mellitus, high levels for glycosylated hemoglobin indicate a poor control of blood glucose levels, and have been associated with cardiovascular disease, nephropathy, and retinopathy [72];

10. *SBP*:

    Systolic Blood Pressure, expressed in millimeters of mercury (mmHg);

11. *TRIG*:

    Triglycerides value, expressed in mg/dl.

12. *LDL:*

    Low-Density Lipoproteins value, expressed in mg/dl

13. HDL:

    High-Density Lipoproteins value, expressed in mg/dl.

    Since measures of TRIG, LDL and HDL were available every 2 years in EDIC, we decided to impute isolated missing values (i.e. missing values placed between two valid measures at the previous and the following year) with a linear interpolation of the 2 adjacent measures, as reported in the example of Figure 3.11;

14. *BMI*:

    Body Mass Index, given by mass/height^2 and thus expressed in Kg/m^2.

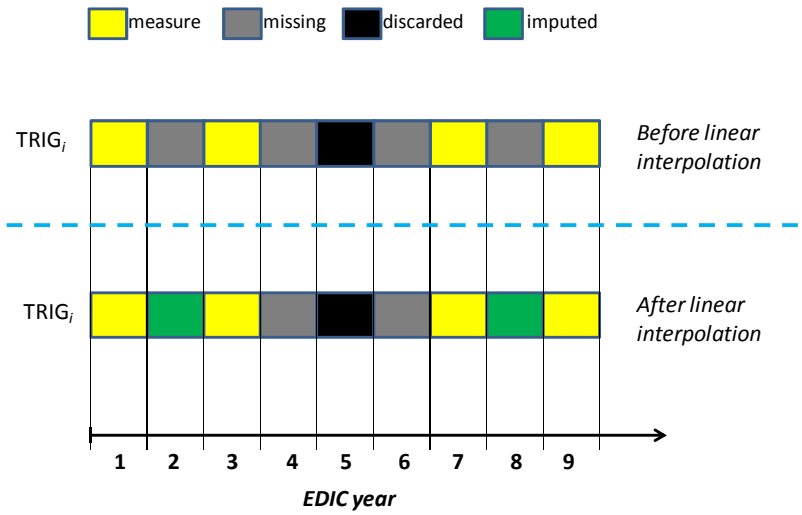


Figure 3.11: *Example of imputation for the covariate TRIG for the i-th patient. Missing values (in gray) placed between two valid measures (in yellow) are replaced by a linear interpolation (in green). Discarded measures (in black) were not used for imputation.*

Table 3.1 reports the final list of the variables used as covariates for the model.

Variable name	Variable Description	Covariate Type	Variable Nature	Unit of measure / Possible States
SEX	Patient's gender	Static	Discrete	M/F
AGE	Patient's age	Static	Continuous	years
DURATION	Patient's duration of diabetes	Static	Continuous	years
TREATMENT	Annual treatment	Static	Discrete	CONVENTIONAL/ INTENSIVE-SELF/ INTENSIVE-FORCED
THERAPY	Total number of years of diabetes not treated with intensive therapy	Static	Continuous	years
SMOKE	smoking status	Static	Discrete	NEVER/EVER
EXERCISE	physical activity status	Static	Discrete	SEDENTARY/MODERATE/STRENUOUS
WHR	Waist-Hip Ratio	Dynamic	Continuous	Unit-less
HBA1C	Glycosylated Hemoglobin value	Dynamic	Continuous	%
SBP	Systolic blood Pressure value	Dynamic	Continuous	mm Hg
LDL	Low-Density Lipoproteins value	Dynamic	Continuous	mg/dl
HDL	High-Density Lipoproteins value	Dynamic	Continuous	mg/dl
TRIG	Triglycerides value	Dynamic	Continuous	mg/dl
BMI	Body-Mass Index value	Dynamic	Continuous	Kg/m^2

Table 3.1: *Variables used for the DBN model of diabetes complications.*

Since the DBN wants to model the transition from a year to the following, a variable is needed to have a valid measure both at time *t* and at time *t+1*. Thus, a variable was considered to have a non-missing value only if it had a valid measure both at year *t* and at year *t+1*. Missingness for each covariate *c*, for each couple of consecutive years *yy*, was computed as the rate of the number of patients with a missing measure for covariate *c* to the total number of available patients for the couple of consecutive years *yy*. The resulting percentages of missingness for the model's covariates are reported in Table 3.2, together with the number of valid patient for each couple of consecutive years. Covariates with no missing values are not reported in the Table.

% missing	WHR	HBA1C	SBP	LDL	HDL	TRIG	BMI	# valid patients
year 1-2	3.96	0.28	0.07	0.14	0.00	0.07	0.07	1439
year 2-3	3.77	0.42	0.14	0.07	0.07	0.07	0.00	1433
year 3-4	3.51	0.77	0.28	0.00	0.00	0.00	0.07	1426
year 4-5	3.18	0.64	2.05	0.35	0.35	0.35	0.07	1417
year 5-6	2.93	1.29	5.00	1.50	1.50	1.50	0.71	1400
year 6-7	2.49	1.54	6.44	1.61	1.24	1.24	1.54	1366
year 7-8	2.32	0.97	5.62	3.00	2.55	2.55	3.00	1335
year 8-9	1.99	1.15	3.51	2.29	1.76	1.76	3.28	1309
year 9-10	1.00	1.61	2.07	5.74	4.98	4.98	3.14	1306
year 10-11	0.78	1.41	1.41	3.53	2.59	2.59	4.00	1274
year 11-12	0.56	1.11	1.43	4.05	3.18	3.18	4.53	1259
year 12-13	0.00	0.95	1.75	4.44	3.33	3.33	4.68	1260
year 13-14	0.00	0.82	1.80	16.99	15.60	15.60	5.07	1224
year 14-15	0.00	0.99	2.17	13.69	12.22	12.22	5.02	1015
year 15-16	0.00	0.58	0.86	30.22	29.06	29.06	3.02	695
year 16-17	0.00	1.77	0.51	15.40	13.89	13.89	2.27	396
year 17-18	0.00	1.61	1.20	6.02	4.02	4.02	3.21	249
year 18-19	0.00	3.36	1.68	60.92	59.66	59.66	6.72	238

Table 3.2: *Missingness throughout the DCCT-EDIC dataset.*

Analyzing Table 3.2, we decided to use the first 15 years of the data, for which the missingness level is always lower than 20%.

Outcomes

As reported in section 1.1, the main diabetic vascular complications are cardiovascular disease, nephropathy and retinopathy. Since there was no uniformity of information between retinopathy status in DCCT and retinopathy status in EDIC, only cardiovascular disease and nephropathy were considered as outcomes for the model. As illustrated in section 3.2, each complication can be modeled by a state transition model, allowing representing the evolution over time of the patients.

1. *CARDIOVASCULAR DISEASE (CVD):*

   According to the DCCT design and protocols, the following cardiovascular episodes were recorded during the study: Myocardial Infarction, Angina Pectoris, Heart Failure, Stroke (or Cerebro-Vascular Accident) and Coronary Artery Disease. Only 64 CVD episodes occurred during the entire DCCT/EDIC study, involving 42 patients. Given this small number, the CVD status of a patient was modeled as a discrete outcome with 2 possible values, as reported in Table 3.3. The possible states transitions are reported in the scheme of Figure 3.12: once a patient suffers a CVD episode, he is

considered a patient at CVD risk, thus having no possibility to come back to the control condition.

CVD STATUS	Description
1: Control	no CVD episodes
2: CVD	any CVD episode

Table 3.3: *Possible values for the CVD status.*

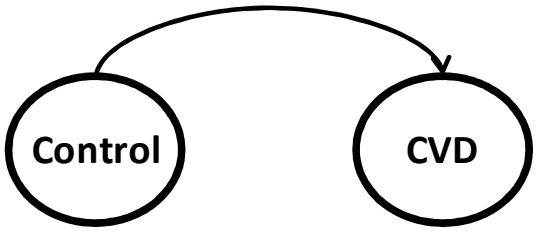


Figure 3.12: *Possible states transitions for CVD.*

2. NEPHROPATHY:

According to the DCCT design and protocols, the nephropathy status was computed combining the values of 2 clinical variables, respectively Albumin Excretion Rate (AER) expressed in mg/day, and Creatinine Clearance (CR-CL) expressed in ml/min/1.73m^2, and coded in 6 levels of increasing severity, as shown in Table 3.4.

DCCT nephropathy severity level	AER (mg/day)	CR-CL (ml/min/1.73m^2)
1	< 40	
2	[40, 100)	
3	[100,200)	
4	[200, 300)	
5	≥ 300	≥ 70
6	≥ 300	< 70

Table 3.4: *Nephropathy severity levels according DCCT criteria.*

Following the guidelines for the outcomes codification defined within the SUMMIT project by SAIL (Sample AILability system), the nephropathy status of a patient was modeled as a discrete outcome with 4 possible values, combining information on the patient's Albumniuria status and End-Stage Renal Disease (ESRD) status, as reported in Table 3.5. Both the individual Albuminuria status and ESRD status are coded according to the variable codification defined by SAIL and reported in Table 3.6. The

Albuminuria status is computed on the basis of the Albumin Excretion Rate value (in mg/day). The ESRD status is computed on the basis of estimated Glomerular Filtration Rate (eGFR) value (expressed in mg/min) or on the basis of episodes of renal insufficiency (dialysis or kidney transplantation).

The possible states transitions are reported in the scheme of Figure 3.13: any transition is allowed, except for any backward step from the last status, which represents a clinical condition in which kidney is no longer able to perform its function.

NEPHROPATY STATUS	Description
1: Control	NormoAlbuminuria and no ESRD episodes
2: microAlbuminuria	microAlbuminuria and no ESRD episodes
3: macroAlbuminuria	macroAlbuminuria and no ESRD episodes
4: ESRD	any ESRD episode

Table 3.5: *Possible values for the Nephropathy status.*

Albuminuria status	AER (mg/day)
Control	< 30
MicroAlbuminuria	$[30, 300)$
MacroAlbuminuria	≥ 300
ESRD status	**eGFR (mg/min)**
Control	> 15
ESRD	≤ 15, or episode of renal insufficiency (kidney transplant or dialysis)

Table 3.6: *Albuminuria and ESRD status according to the SAIL definitions.*

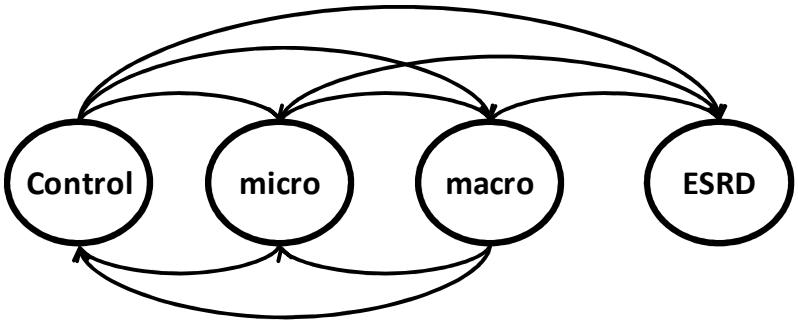


Figure 3.13: *Possible states transitions for Nephropathy.*

Table 3.7 lists the number of patients in each status of Diabetic Nephropathy for each year of the study, according to codification reported in Table 3.5.

Nephropathy numbers in the DCCT/EDIC dataset				
Year	# Controls	# Micro	# Macro	# ESRD
1	1284	157	0	0
2	1292	140	6	0
3	1253	173	7	0
4	1244	165	16	0
5	1240	157	21	1
6	1216	157	25	1
7	1171	166	32	0
8	1133	171	35	1
9	1122	166	33	3
10	1088	194	35	4
11	1072	171	39	6
12	1055	176	51	5
13	996	179	58	8
14	946	174	62	12
15	631	117	48	13
16	464	87	32	12
17	222	58	15	8
18	173	35	18	7
19	100	19	12	4

Table 3.7: *Number of patients for each state of Diabetic Nephropathy throughout the DCCT/EDIC study.*

3.4.5 Learning

3.4.5.1 Split TRAIN/TEST

In order to train and further test the model, the entire dataset was partitioned into 2 subsets of subjects to be used, respectively, as training set, on which learn the model, and as test set, on which test the model. The split train/test was performed stratifying patients by the following 3 variables: *age*, *sex* and *treatment*. The TRAIN:TEST proportion is 9:1 (1298 subjects in the training set and 143 in the test set).

Missingness is uniformly distributed between the training and the test set, exhibiting similar percentages to the ones computed for the entire dataset (see Table 3.8 and Table 3.9 compared to Table 3.2).

The proportion 9:1 was respected also in the number of patients which suffered CVDs episodes (37 in the training set and 5 in the test set) and renal insufficiency episodes (21 in the training set and 3 in the test set).

% missing TRAIN	WHR	HBA1C	SBP	LDL	HDL	TRIG	BMI	# valid patients
year 1-2	4.01	0.23	0.08	0.15	0.00	0.08	0.08	1296
year 2-3	3.80	0.31	0.08	0.08	0.08	0.08	0.00	1290
year 3-4	3.50	0.70	0.23	0.00	0.00	0.00	0.08	1284
year 4-5	3.14	0.63	2.20	0.31	0.31	0.31	0.08	1275
year 5-6	2.86	1.19	5.32	1.59	1.59	1.59	0.79	1259
year 6-7	2.44	1.39	6.76	1.63	1.22	1.22	1.63	1227
year 7-8	2.25	0.92	5.76	3.09	2.59	2.59	3.17	1198
year 8-9	1.95	1.19	3.48	2.21	1.61	1.61	3.48	1177
year 9-10	1.02	1.70	1.96	5.79	4.94	4.94	3.15	1175
year 10-11	0.87	1.48	1.40	3.67	2.71	2.71	4.28	1145
year 11-12	0.62	1.24	1.50	3.98	3.10	3.10	4.78	1130
year 12-13	0.00	1.06	1.95	4.69	3.54	3.54	4.77	1131
year 13-14	0.00	0.91	2.01	17.50	16.13	16.13	5.29	1097
year 14-15	0.00	1.10	2.20	13.86	12.32	12.32	5.39	909
year 15-16	0.00	0.49	0.81	30.26	29.13	29.13	3.24	618
year 16-17	0.00	1.70	0.57	15.01	13.60	13.60	2.55	353
year 17-18	0.00	1.79	1.34	6.25	4.02	4.02	2.68	224
year 18-19	0.00	3.72	1.86	62.33	60.93	60.93	6.51	215

Table 3.8: *Missingness throughout the training set.*

% missing TEST	WHR	HBA1C	SBP	LDL	HDL	TRIG	BMI	# valid patients
year 1-2	3.50	0.70	0.00	0.00	0.00	0.00	0.00	143
year 2-3	3.50	1.40	0.70	0.00	0.00	0.00	0.00	143
year 3-4	3.52	1.41	0.70	0.00	0.00	0.00	0.00	142
year 4-5	3.52	0.70	0.70	0.70	0.70	0.70	0.00	142
year 5-6	3.55	2.13	2.13	0.71	0.71	0.71	0.00	141
year 6-7	2.88	2.88	3.60	1.44	1.44	1.44	0.72	139
year 7-8	2.92	1.46	4.38	2.19	2.19	2.19	1.46	137
year 8-9	2.27	0.76	3.79	3.03	3.03	3.03	1.52	132
year 9-10	0.76	0.76	3.05	5.34	5.34	5.34	3.05	131
year 10-11	0.00	0.78	1.55	2.33	1.55	1.55	1.55	129
year 11-12	0.00	0.00	0.78	4.65	3.88	3.88	2.33	129
year 12-13	0.00	0.00	0.00	2.33	1.55	1.55	3.88	129
year 13-14	0.00	0.00	0.00	12.60	11.02	11.02	3.15	127
year 14-15	0.00	0.00	1.89	12.26	11.32	11.32	1.89	106
year 15-16	0.00	1.30	1.30	29.87	28.57	28.57	1.30	77
year 16-17	0.00	2.33	0.00	18.60	16.28	16.28	0.00	43
year 17-18	0.00	0.00	0.00	4.00	4.00	4.00	8.00	25
year 18-19	0.00	0.00	0.00	47.83	47.83	47.83	8.70	23

Table 3.9: *Missingness throughout the test set.*

3.4.5.2 Discretization of continuous covariates

Dealing with both discrete and continuous variables, an hybrid DBN could appear the most appropriate choice. However, since modeling an hybrid DBN requires specific assumptions on the distribution of continuous variables, a discrete DBN was adopted. Thus, each continuous variable was discretized using specific cut-offs and the whole model was fully specified by a set of Conditional Probability Tables (CPTs). Variables *WHR*, *SBP*, *LDL*, *HDL*, *TRIG* and *BMI* were discretized according to literature cut-offs reported in Table 3.10.

Variable	Cut-offs	Number of cut-offs	Number of states	Reference
WHR	0.9 (men) and 0.85 (women)	1	2	[76]
SBP	120 mmHg and 140 mmHg	2	3	[59]
LDL	100 mg/dl	1	2	[84]
HDL	40 mg/dl (men) and 50 mg/dl (women)	1	2	[84]
TRIG	150 mg/dl	1	2	[84]
BMI	20 Kg/m^2 and 25 Kg/m^2	2	3	[91]

Table 3.10: Literature **c***ut-offs used for continuous variables.*

Since there are no literature guidelines concerning cut-offs for the age, the duration of the disease and the number of years of diabetes not treated with intensive therapy, a search strategy to identify the optimal cut-off values while learning the DBN structure was defined, and it will be described in the next section. The same cut-off search strategy was applied to HbA_{1c}, being it the direct intervention target of the DCCT and EDIC studies.

3.4.5.3 Structure and cut-offs learning

The DBN here implemented aimed to merge the data-driven information with literature knowledge. Therefore, the DBN structure was learned directly from data, but incorporating some constraints derived from the literature both in the network structure (i.e. allowing only certain edges to be learned, as detailed in the following) and in discretization cut-offs, as explained in the previous section (see Table 3.10).

Nodes of the DBN can be classified into four classes, each of them with specific edge constraints:

- Static Nodes:

  Each static covariate (see section 3.4.4.3) is represented by a static node (St) in the network (except for the covariate *TREATMENT*, which will discussed later). Thus, the static nodes are: *SEX*, *AGE*, *DURATION*, *THERAPY*, *SMOKE* and *EXERCISE*. These nodes cannot be influenced by other nodes, i.e. they can be parent but not child nodes. Edges from static nodes can be directed to dynamic nodes at time *t*, $Dyn_v(t)$, or outcome nodes at time *t*, $Out_v(t)$.

- Dynamic Nodes:

  Each dynamic covariate (see section 3.4.4.3) results in 2 dynamic nodes in the network, representing the value at time *t*, $Dyn_v(t)$, and *t-1*, $Dyn_v(t-1)$, respectively, where *t* is a positive integer representing the year ($2 < t < 15$). Thus, the dynamic nodes are: *HBA1C(t)*, *HBA1C(t-1)*, *SBP(t)*, *SBP(t-1)*, *LDL(t)*, *LDL(t-1)*, *HDL(t)*, *HDL(t-1)*, *TRIG(t)*, *TRIG(t-1)*, *WHR(t)*, *WHR(t-1)*, *BMI(t)* and *BMI(t-1)*. Each $Dyn_v(t-1)$ node cannot be a child node. Each $Dyn_v(t)$ node is forced to be a child of its correspondent $Dyn_v(t-1)$ node and is a candidate child of every other $Dyn_i(t-1)$ node and St node.

- Outcome Nodes:

  As for dynamic covariates, each outcome (see section 3.4.4.3) results in 2 outcome nodes in the networks, representing the value at time *t*, $Out_v(t)$, and *t-1*, $Out_v(t-1)$, respectively. Thus, the outcome nodes are: *CVD(t)*, *CVD(t-1)*, *NEPHRO(t)* and *NEPHRO(t-1)*. Each $Out_v(t-1)$ node cannot have parents. Each $Out_v(t)$ node is forced to be a child of its correspondent $Out_i(t-1)$ node, and can be child of every other $Dyn_i(t-1)$ node, St and $Out_i(t-1)$ node.

- Intervention Nodes:

  The covariate *TREATMENT* (see section 3.4.4.3) is forced to result in 2 static intervention nodes, Int(t) and Int(t-1), representing the state of the treatment at time *t* and *t-1* respectively, since, from a clinical point of view, it is relevant not only information on the current treatment but also on the recent change in treatment. Thus, the intervention nodes are: *TREATMENT(t)* and *TREATMENT(t-1)*, for a total of 26 nodes in the network. Since the covariate *TREATMENT* represents the intervention variable of the DCCT and EDIC studies, where the intent of the intensive treatment was to achieve HbA1c level in the non-diabetic range (<6.05%) [61], the nodes *TREATMENT(t)* and *TREATMENT(t-1)* can affect only the *HBA1C(t)* node.

Table 3.11 summarizes all node types and the related candidate parent/child node types according to the edge constraints.

Type	Forced parent edge	Forced child edge	Possible parent edge	Possible child edge
St	None	None	None	$Dyn_v(t)$, $Out_v(t)$
$Dyn_v(t-1)$	None	$Dyn_v(t)$	None	$Dyn_v(t)$
$Dyn_v(t)$	$Dyn_v(t-1)$	None	$Dyn_v(t)$, St, Int(t)*, Int(t-1)*	$Out_v(t)$
$Out_v(t-1)$	None	$Out_v(t)$	None	$Out_v(t-1)$
$Out_v(t)$	$Out_v(t-1)$	None	$Out_v(t-1)$, $Dyn_v(t)$, St	None
Int(t)	None	None	None	HBA1C(t)
Int(t-1)	None	None	None	HBA1C(t)

Table 3.11: *Nodes type and edges constraints. * Int(t) and Int(t-1) nodes are possible parents only for the HBA1C(t) node.*

The DBN structure was learned from data by searching the space of all possible network structures with a Tabu Search algorithm [66], identical to the *Hill Climbing* step of the *Max-Min Hill Climbing* (MMHC) algorithm for Bayesian Networks structure learning described by Tsamardinos et al. in [88]. Basically, the search aims to find both edges and discretization thresholds in order to maximize the DBN's prediction ability.

The search shapes edges connecting $Dyn_v(t-1)$, $Out_v(t-1)$, St, Int(t) and Int(t-1) nodes to $Dyn_v(t)$ and $Out_v(t)$ ones. Considering the constraints listed above, each $Dyn_v(t)$ and $Out_v(t)$ node has one forced parent node, and a number of candidate ones. For each node, the goal is to find the best parent combination, determined by the likelihood-equivalence Bayesian Dirichlet score with uniform priors (BDeu), with Equivalent Sample Size $\alpha = 5$ [88], [67]. The goal is reached by searching the combination of parent edges that maximizes BDeu for each candidate node on the training data. The forced parent is always included when computing the score for a combination of candidate parents. Each node's parents can be searched independently.

The implemented search is stepwise. At each step the algorithm determines if an edge is to be added or removed from the optimal parent combination obtained at the previous step. Since an edge can be either present or absent, each combination of candidate n parents can be represented by a binary vector p with a size n (the forced parent is not considered in p). The initialization step assumes no candidate parent selected, i.e. the BDeu score for a node is computed considering only the edge of its forced parent. Then the search algorithm proceeds by evaluating n possible steps, each one determined by switching a single binary value of p. For example, the search for a node with $n = 3$ begins setting p equal to [0, 0, 0]. The very first considered steps are [1, 0, 0], [0, 1, 0] and [0, 0, 1]. The step associated to the highest BDeu score is then selected, and the search continues.

Once a step is selected, its vector p is compared to the elements of a Tabu list with maximum size t. If p is already present in the Tabu list, its score is set as *-Inf*. If not, p is pushed into the Tabu list. Once the list is full, new vectors push out the previously inserted ones, following a first-in-first-out approach. If more than s steps are completed without a BDeu score improvement, the search stops. Values for t and s were 100 and 15 respectively, according to the literature gold standards [66].

Note that in the original MMHC algorithm the edges, at each step, could undergo three possible perturbations, namely addition, removal or reversal. In the present case, however, edge direction is fixed, so there is no need to test for edge reversal. Furthermore, the network is acyclic by construction. This yield two major consequences: on the one hand, there is no need to check for acyclicity after every step of the Tabu Search algorithm; on the other hand, the optimal parent set for each node can be identified independently of the other nodes, thus greatly reducing computational complexity.

As reported at the end of the previous section, for some nodes discretization thresholds have been taken from the literature (Table 3.10), while for variables *AGE*, *DURATION*, *THERAPY* and *HBA1C* the thresholds have been inferred by the data. For these latter variables, each one was assumed to have three possible states (low = 1, medium = 2 and high = 3), and the thresholds could be computes as 4 possible combinations: (a) the 33^{th} and 66^{th} percentile; (b) the 25^{th} and 50^{th} percentile; (c) the 25^{th} and 75^{th} percentile; or (d) the 50^{th} and 75^{th} percentile. Thresholds combinations have been explored during the BN structure learning described above: in particular, the learning of the edges was performed for each possible threshold combination on the variables *AGE*, *DURATION*, *THERAPY* and *HBA1C*, for a total of $4^4 = 256$ combination. For each combination, the whole BN score was computed as the sum of each individual $Dyn_v(y)$ node and $Out_v(t)$ node BDeu score.

Table 3.12 reports the learned thresholds for *AGE*, *DURATION*, *THERAPY* and *HBA1C*, thus completing information of Table 3.10.

Variable	Cut-offs	Number of cut-offs	Number of states
AGE	28 years and 40 years	2	3
DURATION	100 months and 157 months	2	3
THERAPY	4.92 years and 10.17 years	2	3
HBA1C	7.1 % and 9.1 %	2	3

Table 3.12: Learned **c***ut-offs for continuous variables*

3.4.5.4 Parameters learning

Once the structured is fixed, the phase of learning the conditional distributions implied by the network consists in estimating, for each variable *X*, a set of parameters $\boldsymbol{\theta}_{X|Pa_X}$ describing the dependency of *X* over its parents Pa_X.

In the case of a discrete network, the parameters to be estimated are all the entries of the CPT of each variable, *i.e.* the values $\theta_{x|pa_x} = P(X = x|Pa_X = pa_X)$ for each of the possible values of *X* and of its parents Pa_X.

To this aim, the Bayesian *maximum a posteriori* (MAP) estimates was exploited. The Bayesian MAP estimates consists on a *maximum likelihood* estimates (based on calculating the relative frequencies of the different events in the data), augmenting this observed data with prior distributions over the values of these parameters.

The maximum likelihood estimate of $\theta_{x|pa_x}$ is given by:

$$\hat{\theta}_{x|pa_x} = \frac{N(X = x \wedge Pa_X = pa_X)}{N(Pa_X = pa_X)}, \qquad (3.3)$$

where $N(c)$ counts the number of observations in the dataset satisfying condition *c*.

One of the risks of maximum likelihood is that it can sometimes return estimates equal to zero, in case no example satisfying the condition at the numerator is observed in the dataset. To avoid this situation, it is often preferred to smooth the estimate with a coefficient α known as *Equivalent Sample Size* (ESS). The smoothed estimate is given by:

$$\hat{\theta}_{x|pa_x} = \frac{N(X = x \wedge Pa_X = pa_X) + \alpha}{N(Pa_X = pa_X) + |Val(Pa_X)| \cdot \alpha}, \qquad (3.4)$$

where $|Val(Pa_X)|$ is the number of distinct values Pa_X can take.

This expression corresponds to a MAP estimate of $\theta_{x|pa_x}$, assuming a Dirichlet prior distribution with equal-valued hyper-parameters α. An intuitive interpretation of α is the number of imaginary samples, for each combination of values of *X* and Pa_X, assumed to have been observed before estimating $\theta_{x|pa_x}$ from the data, as already explained in section 2.5.1. In this context, α was set equal to 5, as already mentioned in the previous section.

3.4.6 Prediction

Once the model was fully specified through the learning phase, it was applied on the cohort of subjects of the test set to predict the evolution of the patients' state.

Given the initial discretized values for both covariates and outcomes, the CPTs allow predicting the values of all the variables for the following years and thus the probability of developing a certain complication. For example, in Figure 3.14 the value of discretized HbA1c level at time point *t* depends only on the values of the same variable at time *t-1*, thus the correspondent CPT is represented by a 3x3 matrix, each column representing the conditional probability of HbA1c at time *t* given the value of its parent (HbA1c at time *t-1*). Assuming for the *i-th* patient that HbA1c has low level at time point *t-1*, the level at time *t* will be predicted by a roulette wheel selection method where the chance for every possible value is given by the conditional probability in the correspondent cells. In the case shown in Figure 3.14, given the trained CPT and a low level of HbA1c at time *t-1* for the *i-th* patient, the probabilities for low, medium and high level at time *t* are 5%, 74% and 21%, respectively, and the simulation predicts a medium level for HbA1c at time *t*.

By applying iteratively this procedure to all the covariates and outcomes, the evolution of each patient belonging to the cohort of interest can be predicted year by year. This approach permits to predict the progression of a complication also over long period of time.

The described approach was applied to the cohort of patients of the test set, starting from the initial values and using a prediction horizon of 15 years. For each patient, 100 simulations were performed in order to obtain a probability distribution for each variable and for each year. The basic idea of this stochastic simulation approach is to run a simulation process that, starting from the observations and following a topological order, samples a new value of each unobserved variable given the values of all the other variables sampled so far. In this way a chain of values is generated. Such chain is known to converge to the posterior distribution of the variables given the observations [65].

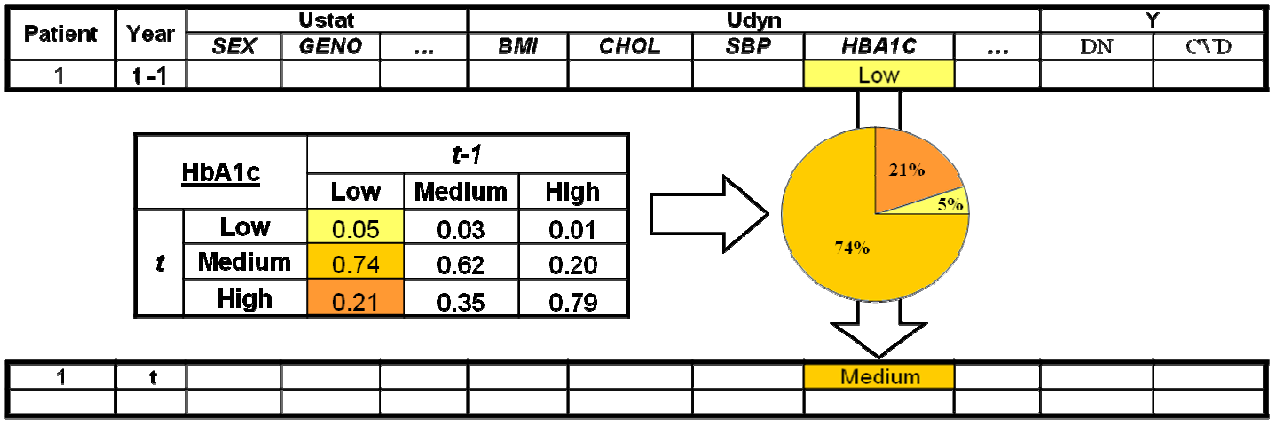


Figure 3.14: *Example of a single variable prediction by exploiting the correspondent CPT.*

3.5 *Results*

The final network resulting from the learning step is represented in Figure 3.15.

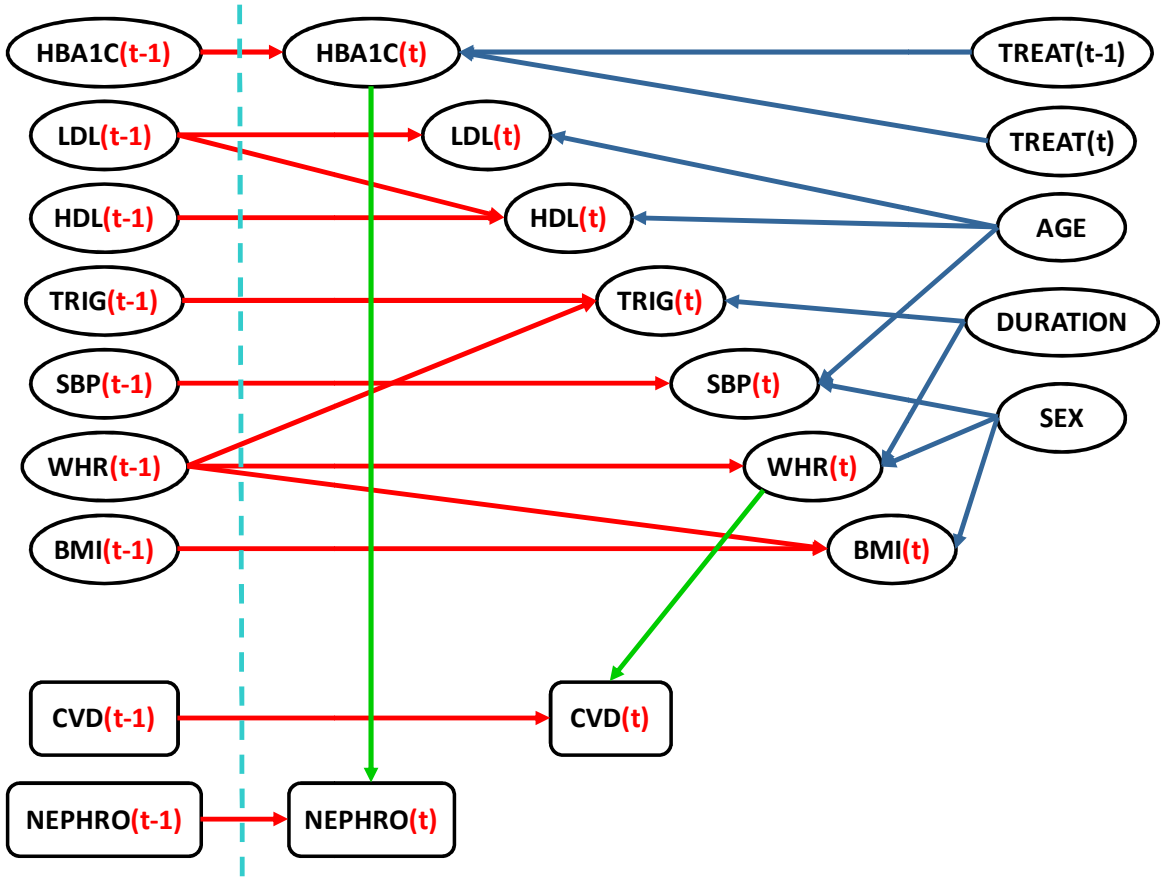


Figure 3.15: *Final DBN structure*

Analyzing Figure 3.15, 2 well-distinct blocks can be observed in the network: a sub-network for Nephropathy (Figure 3.16), in which a clear short-term effect is played by the variable *TREATMENT* on the nephropathy state through the intermediate effect on the variable *HBA1C*, and a sub-network for the Cardiovascular Disease (Figure 3.17), in which a clear long-term effect is played by the variable *DURATION* on the CVD state through an intermediate effect on the variable *WHR*. It is interesting to note how the lipid variables (i.e. *TRIG*, *LDL* and *HDL*) as well as the anthropometric variables (*WHR* and *BMI*) belong to the same sub-network of CVDs, thus showing a certain consistency with clinical knowledge [76]. The variables *THERAPY* (i.e. the number of years of diabetes not treated with intensive therapy), *SMOKE* and *EXERCISE* were left out from the network, since their effect was likely overcome by the stronger influence of other variables.

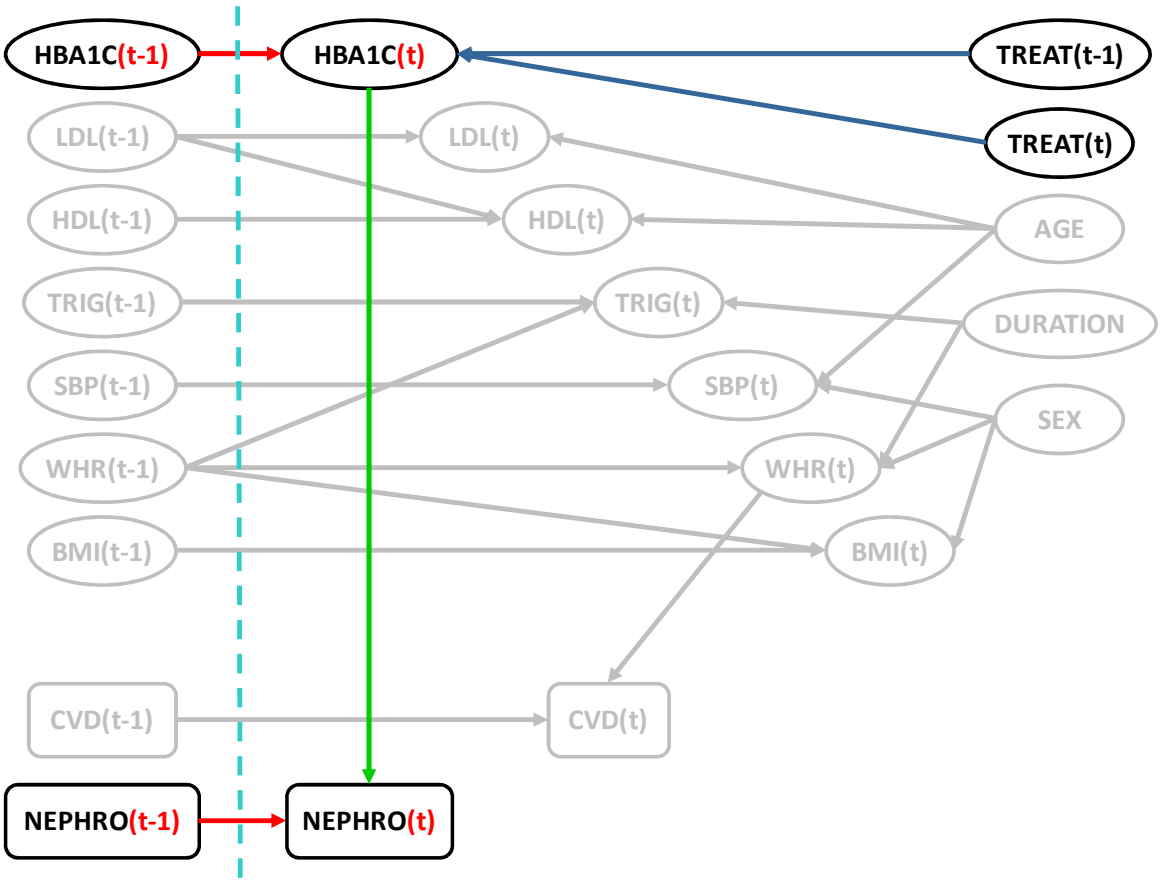


Figure 3.16: *Nephropathy sub-network.*

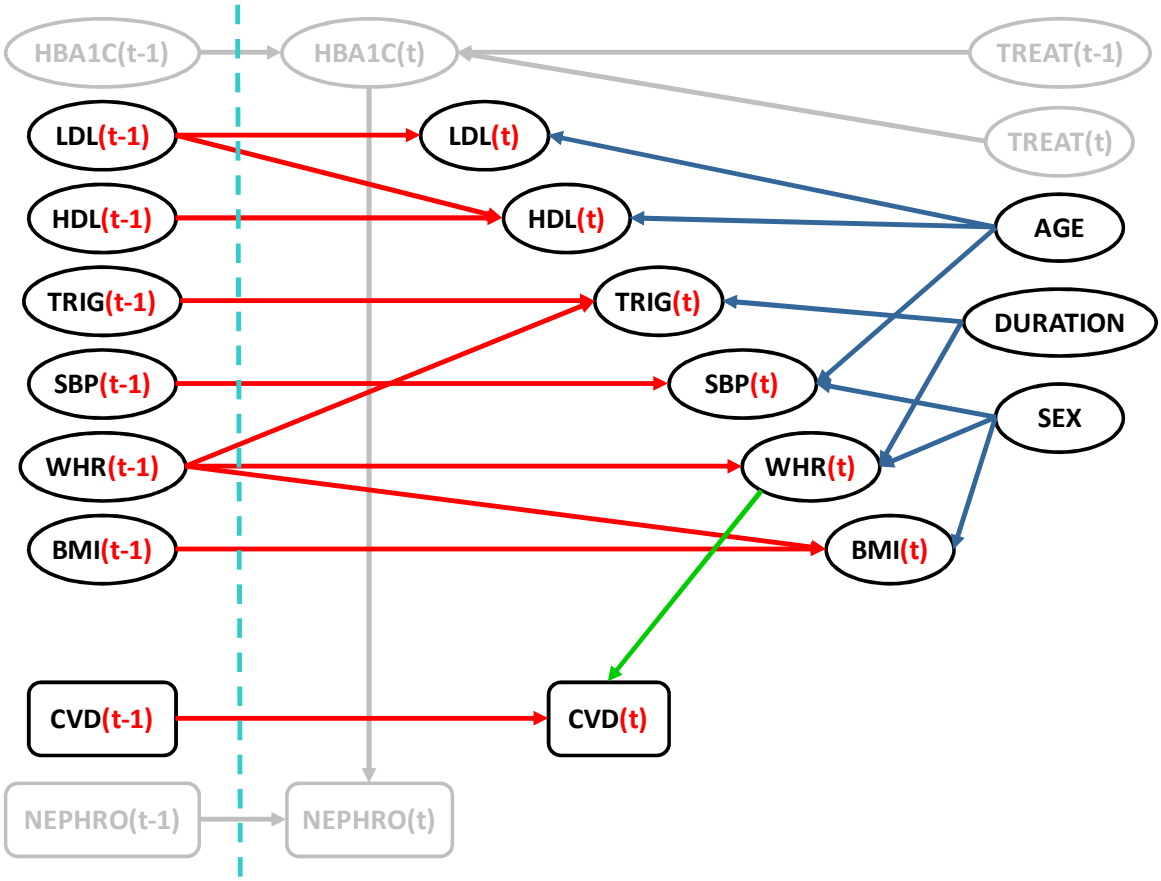


Figure 3.17: *CardioVascular Disease sub-network.*

The validation step consisted in comparing results of simulations run on the initial population of the test set to real data. In particular, the comparison was performed at a population level: for each dynamic variable, the annual population distributions computed on simulated data were compared to the annual population distributions computed on real data. The annual distribution was computed considering all the 100 simulated values for each patient. Figure 3.18 to Figure 3.26 show real distributions (top panel) and percentage difference with simulated distributions (bottom panel) for all the dynamic variables in order to quantify prediction accuracy. Considering that only the first 15 years of the data were used to train the model, the population predictions for the first 15 years fit very well real data, exhibiting percentage differences not greater than 10% for all the dynamic variables. In particular for the outcomes, the differences are lower than 5% for all the 15 years. These results are similar to the ones obtained by [74], where the authors defined a valid model as one in which the mean simulate event rates correspond to the mean published event rates within a range of ±10%.

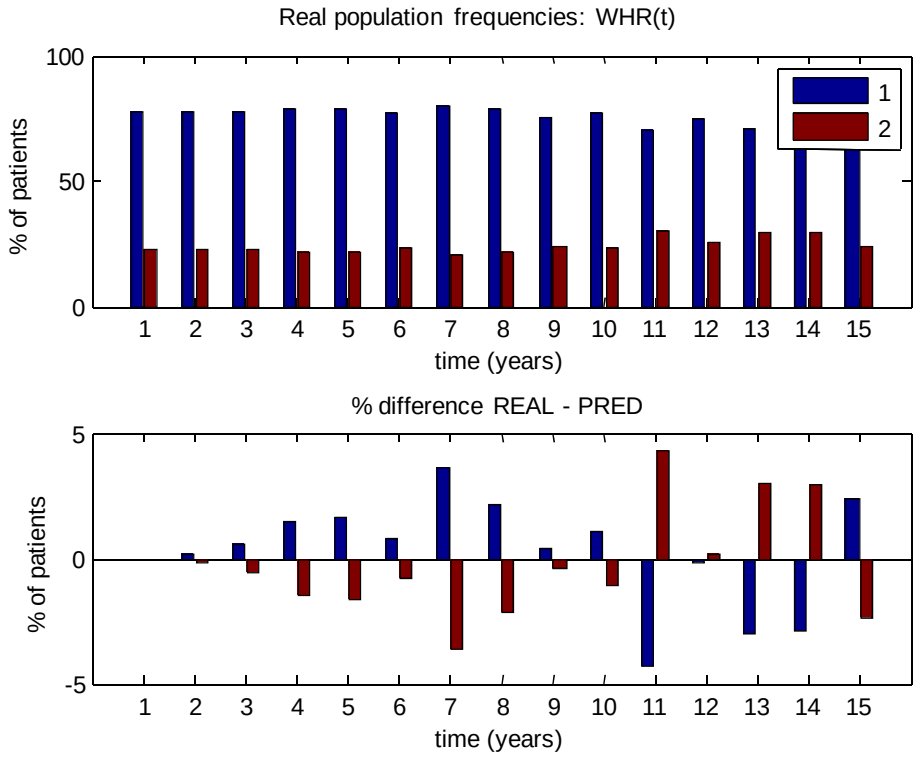


Figure 3.18: *Real population distribution of WHR and percentage difference with predicted one for each year.*

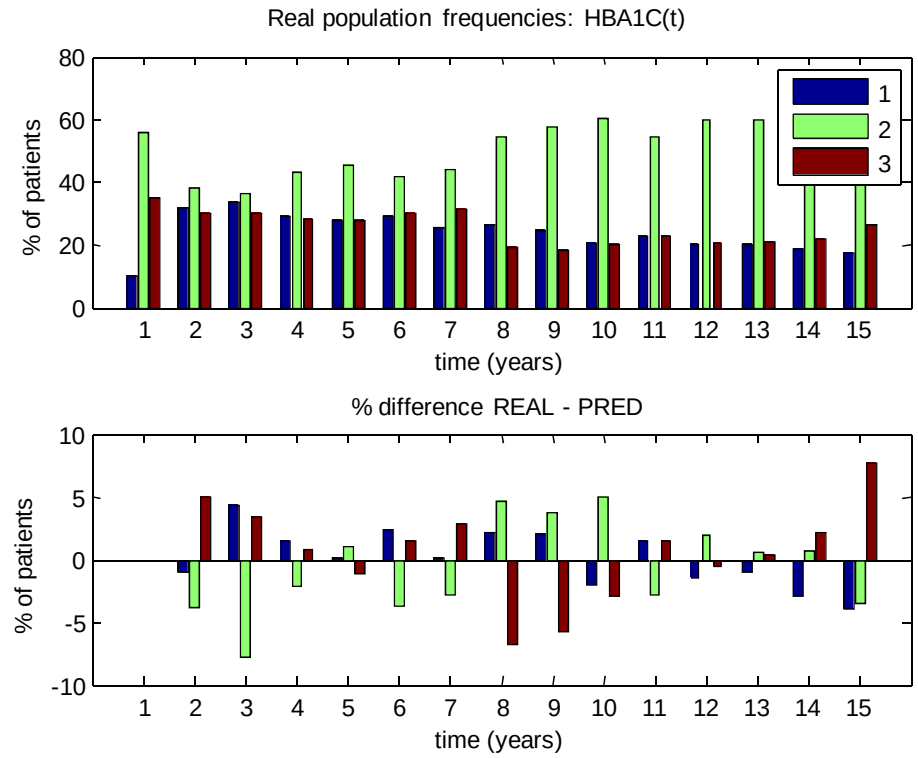


Figure 3.19: *Real population distribution of HBA1C and percentage difference with predicted one for each year.*

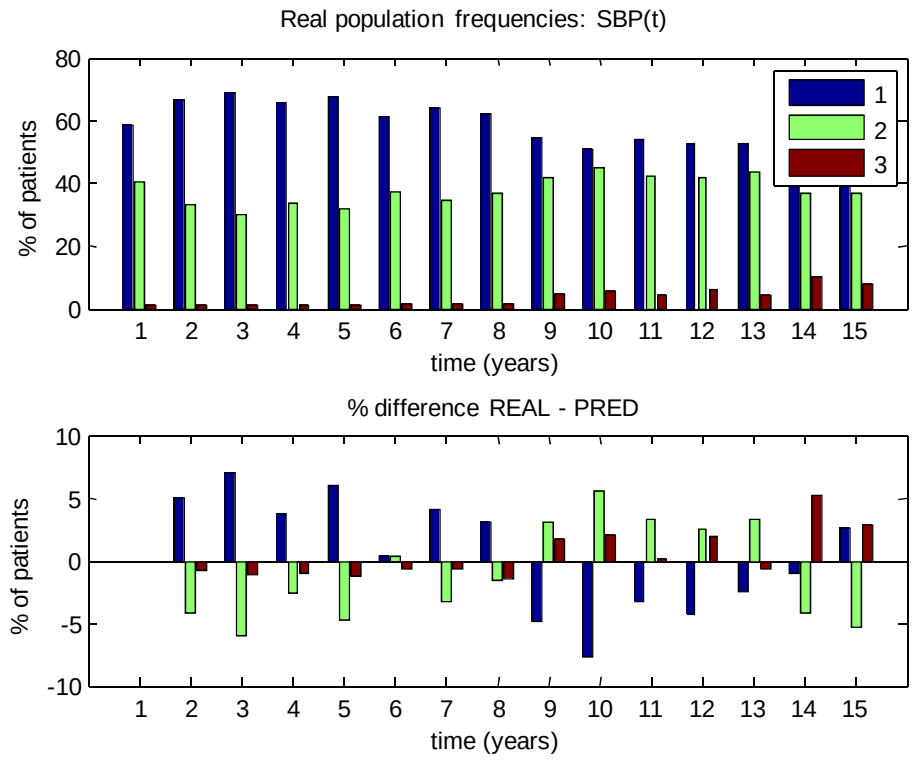


Figure 3.20: *Real population distribution of SBP and percentage difference with predicted one for each year.*

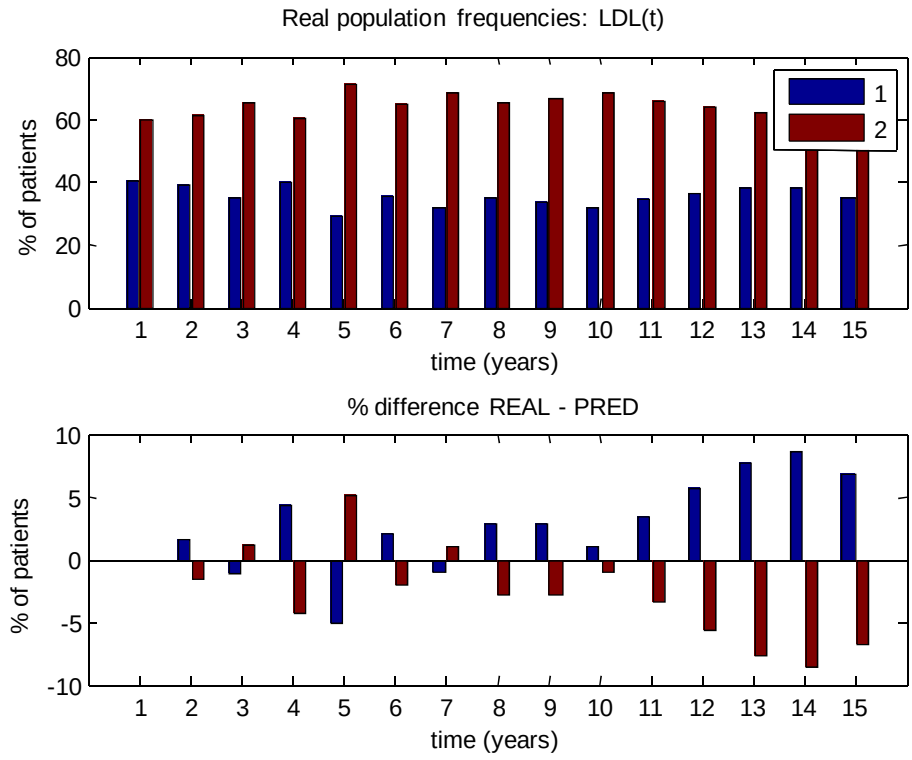


Figure 3.21: *Real population distribution of LDL and percentage difference with predicted one for each year.*

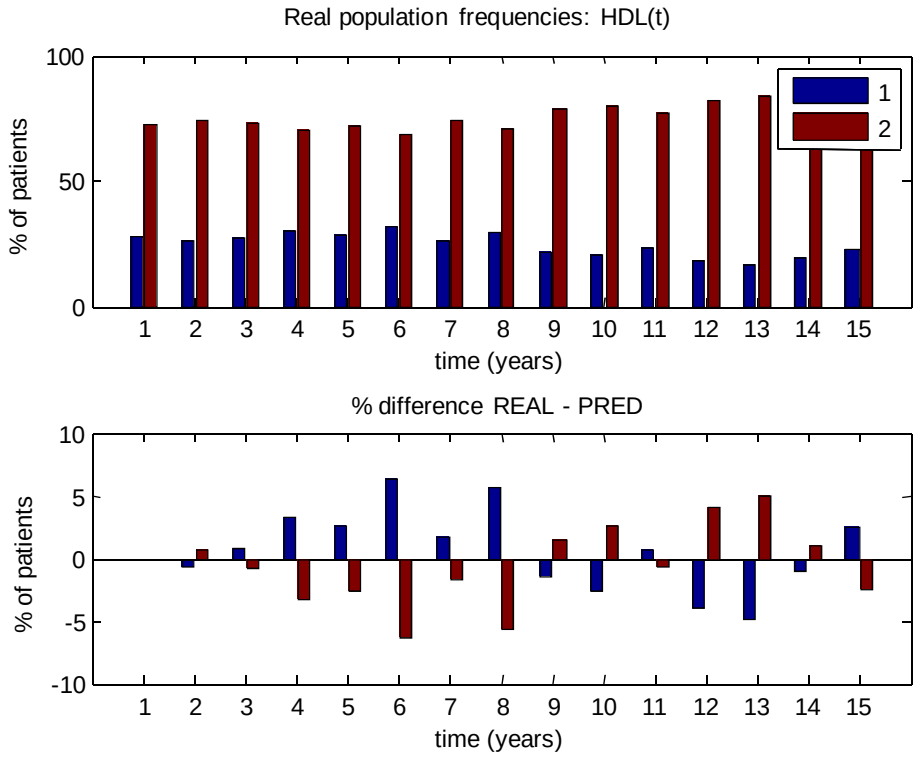


Figure 3.22: *Real population distribution of HDL and percentage difference with predicted one for each year.*

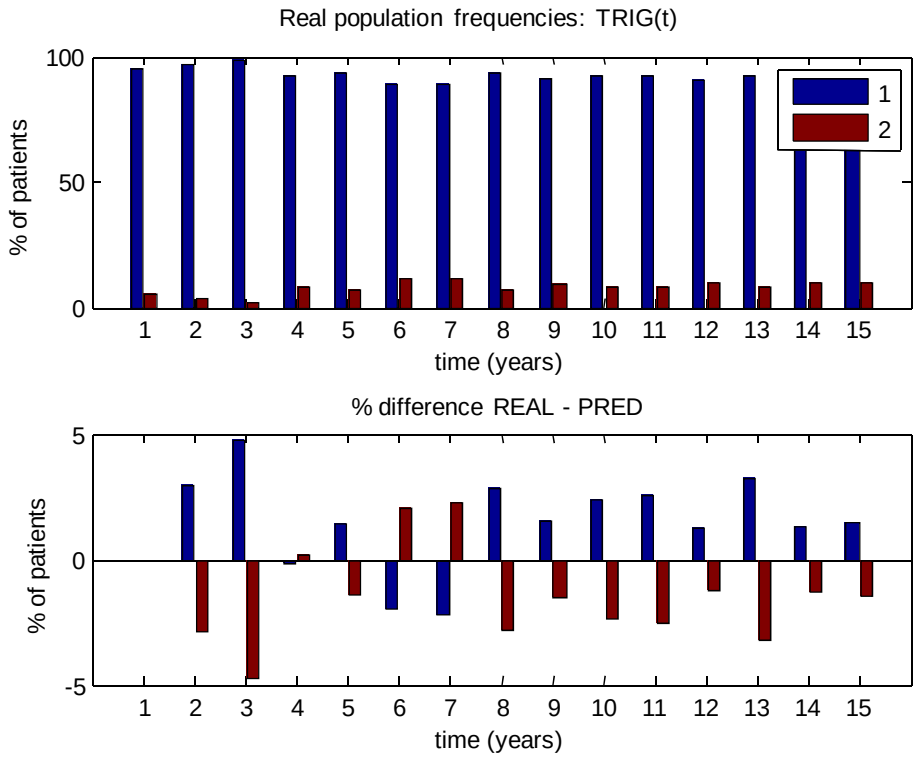


Figure 3.23: *Real population distribution of TRIG and percentage difference with predicted one for each year.*

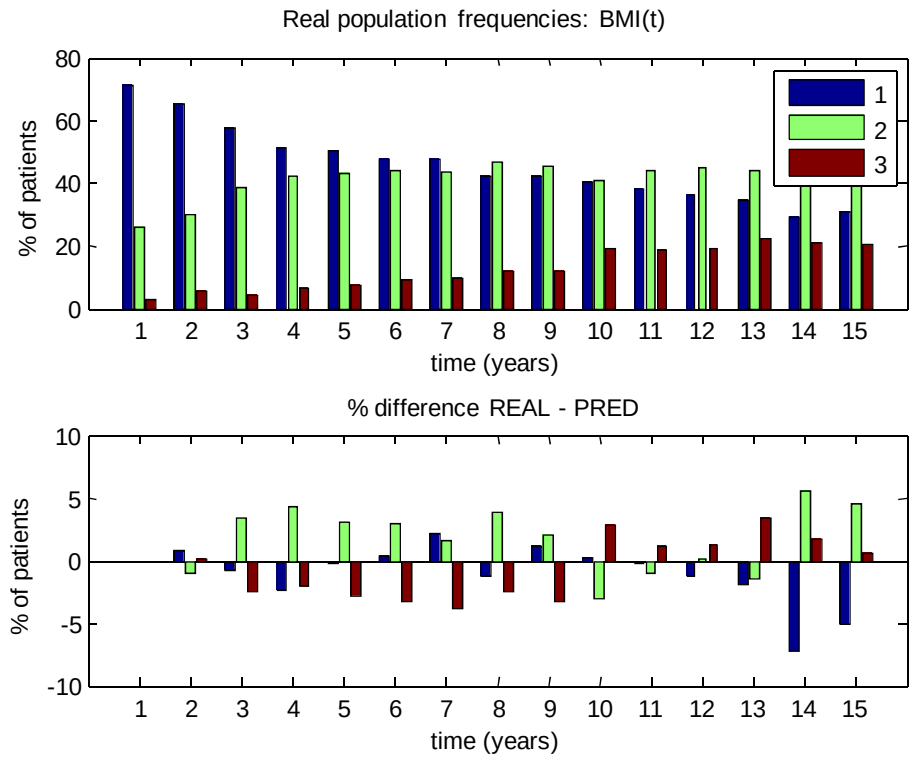


Figure 3.24: *Real population distribution of BMI and percentage difference with predicted one for each year.*

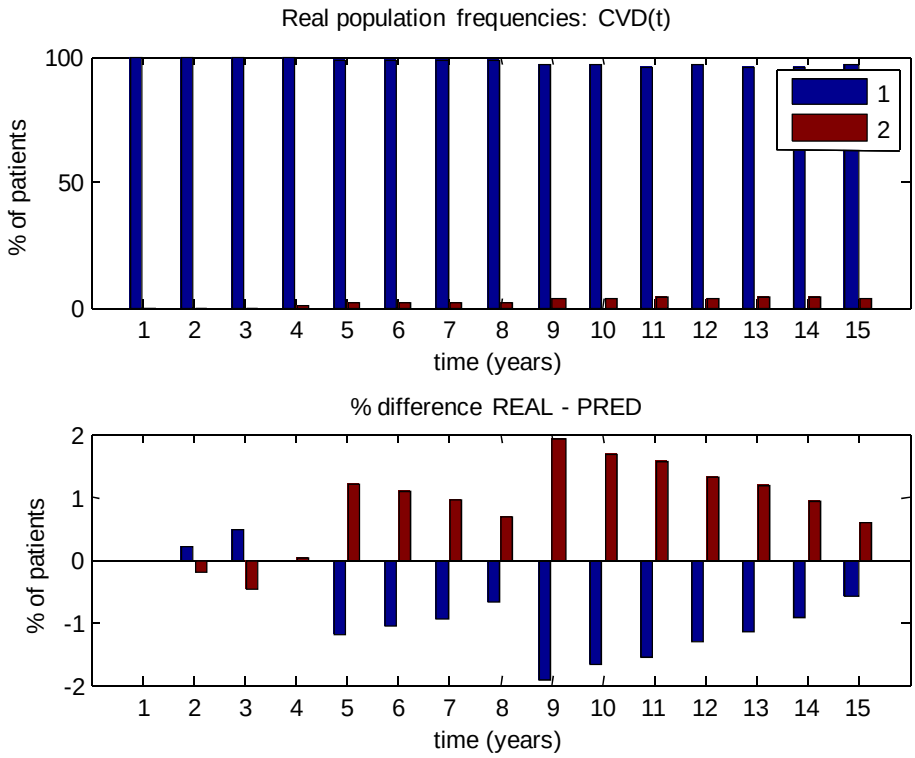


Figure 3.25: *Real population distribution of CVD and percentage difference with predicted one for each year.*

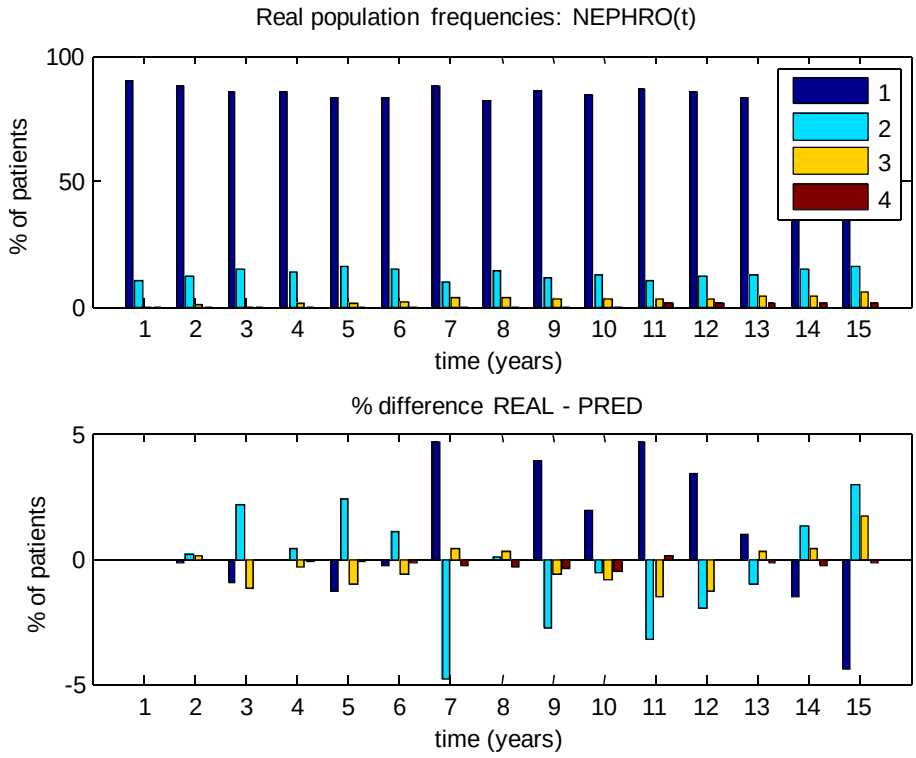


Figure 3.26: *Real population distribution of NEPHRO and percentage difference with predicted one for each year.*

3.6 *Software tool for diabetes care professionals*

As already pointed out in the Introduction of the current Chapter, the final aim of an *in-silico* model is to provide clinicians with a tool for supporting decision analysis, in order to predict the risk for long-term complications, thus optimizing clinical trials and avoiding invasive and expensive tests. In this context, the final step of the present work is representing by the development of a web application to simulate the progression of diabetes long-term complications. In particular, here we developed a standalone Java application that implements stochastic simulation based on Bayesian network structure and parameters learned from the DCCT and EDIC datasets. The main goal of the application is to provide a tool to simulate single patient or population evolution dynamics, toward nephropathy and cardiovascular disease. The tool will allow professionals involved in diabetes care to analyze and predict the onset of pathologies such as coronary heart disease, stroke and nephropathy with a certain belief, based on patients or population anamnestic evidence.

In order to reach the highest compatibility with existing operative systems and an easy to install and run deployment strategy, we employed Java Standard Edition (Java SE) technology to develop this application. Indeed, Java SE Platform lets you develop and deploy Java applications on desktops and servers, offering rich user interface, performances and portability that today application require. Having as objective the distribution of the software over the internet, Java Web Start technology has been selected. This technology is being developed as Java Network Launching Protocol & API (JNLP), which provides a browser-independent architecture for deploying Java technology-based applications to the client desktop.

3.6.1 Methods

3.6.1.1 Packages

Classes and methods are grouped in four Java packages. The package *dbn* is the main package. It provides initialization of the main classes and the application layer control. The package *libDBN* contains all the classes and methods developed to simulate single patient or population evolution dynamics and all storage data structures. The package *dbnGUI* has been developed to provide a graphical user interface (GUI) employing Java AWT and SWING libraries. Implementing the interfaces ActionListener and ItemListener provided by these libraries, it makes possible the interaction with user and events. The package *visualDBN* provides the facilities for visualization of network structure and outputs analysis graphs.

3.6.1.2 Data structures

The structure and parameters of the Bayesian network learned are stored in an object, instance of the class NetStructure. It stores the conditional probability tables (CPTs) values and adjacency matrix in matrices and variable names, nodes arity (i.e. number of possible states) and discretization levels for each variable in arrays. The object can be serialized, indeed the class implements the interface Serializable, and saved to disk. This allows loading the entire network structure in a single step, making faster the application start up process. In case of network modification due to subsequent learning processes it is possible to re-load the structure and all parameters into the *NetStructure* object and save it again to disk. Access to CPTs values is allowed by means of a function that maps a combination of indexes given the topological order of a variable to the linear index of the CPTs matrix.

The data employed for and obtained from the simulation process are stored in a data structure implemented with the class *evolutionMatrix*. Each year of simulation is stored in a generic list container, an *ArrayList<T>* container of the Java Generics library. Several auxiliary and temporary data structures are employed to perform internal computation.

3.6.2 Simulations

Inference in Bayesian network can be accomplished in several ways, such as exact inference by enumeration or by variable elimination, message passing algorithms and stochastic simulation. Exact inferences need to rewrite a query P(X|e) in terms of CPT entries product. Given any subset of *Xi* setting them to certain values due to evidence, we can calculate the probability distribution of some other subset of *Xi* by marginalizing over the joint. This is costly due to calculating an exponential number of joint probability combinations. In this application, we implemented a Markov Chain Monte Carlo stochastic simulation, as described in section 3.4.6. Briefly, for each variable the distribution of possible values is obtained from the CPT tables. Inserting the covariates initial values for the single patient or for the population, distribution probabilities may be generated according with the observed data. The new value of the considered variable is obtained from a random sampling among possible values according to its conditional probability distribution. A stochastically selected value is assigned to current variable. This procedure is repeated for each variable to generate a complete dataset for the selected number of patients. By this way, users can insert initial values and simulate the entire dynamic evolution process of the cohort of patients.

3.6.3 Visualization

Simulation results may be visualized in the main windows of the application where a table reports the distribution per year of each variable. Incidence of nephropathy and cardiovascular disease over the year may be visualized as a graph. We are currently finalizing the implementation of the DAG visualization and the exporting functions for graph and tables. A preliminary mock-up of the interface is shown in Figure 3.27.

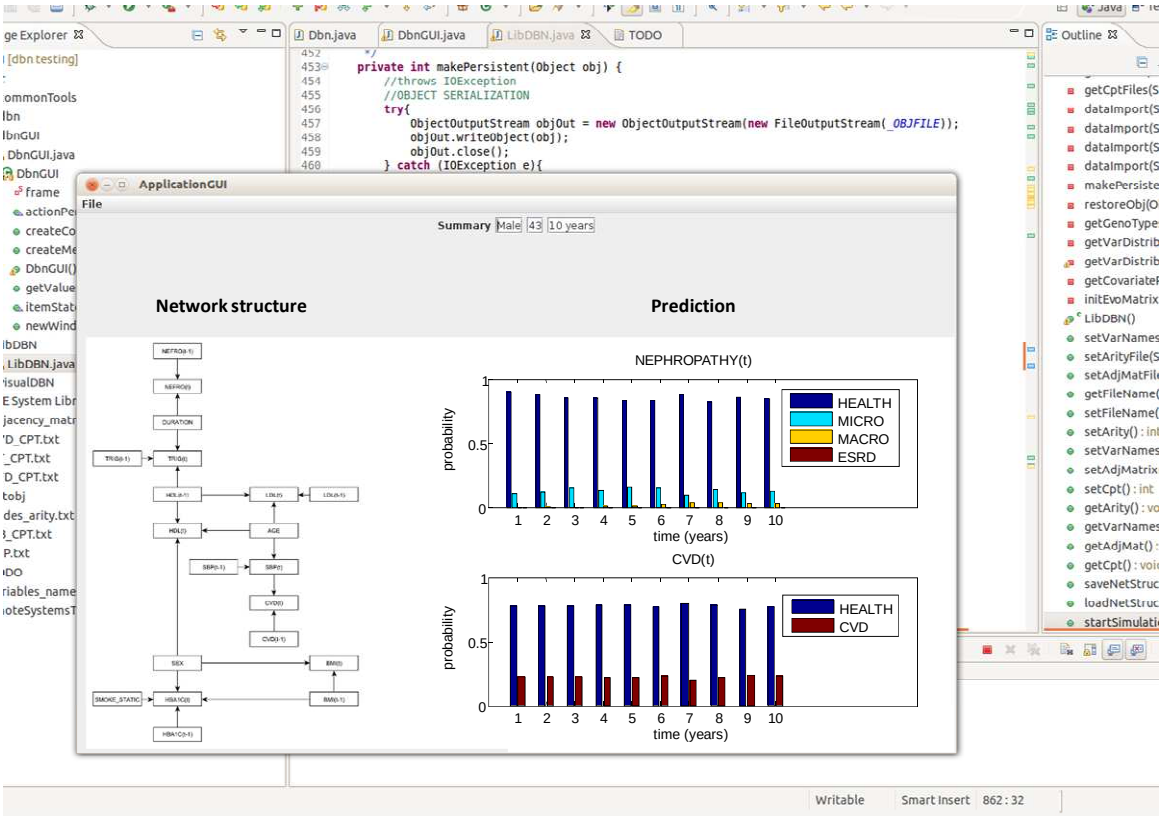


Figure 3.27: *Preliminary mock-up for the user interface of the software tool.*

3.7 *Discussion*

In this chapter the problem of modeling the combined effect of phenotype and treatment on the outcome was discussed.

In-silico models of complex diseases are essential to predict the risk for long-term complications, thus optimizing clinical trials and avoiding invasive and expensive tests.
A requirement for diabetes simulation models has been identified in the medical and healthcare policy community, and, as a result, a number of models, have been developed and reported in the literature. However, these models are mainly based on Markov Models, thus requiring as many nodes as the number of combinations of variable values.

In the present work, the progression of two vascular diabetes complications, Cardiovascular disease and Nephropathy, was modeled using Dynamic Bayesian Networks and integrating in the model phenotypic information as well as information on treatment. A Bayesian Network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph. *DBNs*

are a special class of *BNs* that model the stochastic evolution of a group of random variables over time. *DBNs* are advantageous with respect to Markov Models since each variable is represented by one node, thus, extending a *DBN* model with the addition of new variables just requires adding as many nodes.

The *DBN* model was developed on the data collected from the DCCT (Diabetes Control and Complication Trial), a randomized clinical trial which involved 1441 type 1 diabetic volunteers between 1982 and 1993, with the aim of comparing the effects of standard control of blood glucose versus intensive control on the complications of diabetes.

In particular, relying on an *a-priori* information on the network general structure, the model was learn directly from a subset of real data, and validated on the subset left out during the learning phase.

Results regarding the network structure show a good consistency with clinical knowledge, exhibiting 2 well-distinct blocks of effects: a first block with a stronger short-term effect for Nephropathy, regulated by the indirect effect of treatment on HbA1c, and a second block with a stronger long-term effect for Cardiovascular Disease, regulated by the indirect effect of the duration of diabetes on Waist-Hip Ratio, and involving also all the lipid variables.
Results regarding the simulated progression of complications show very good performances, exhibiting a prediction accuracy greater than 90 % for all the dynamic covariates and greater than 95 % for the outcomes, Nephropathy and Cardiovascular Disease, thus proving the effectiveness of the model.

The good prediction performances of the model make it rather suitable to be use as a tool for support clinical decision analysis. To such an aim, a web Java application that implements stochastic simulation based on the structure and parameters learned from the DCCT and EDIC datasets is currently under development. The web application development is still ongoing, but the current version represents a good starting point for future extensions and improvements.
Future developments, in particular, can regard the extension of the *DBN* model and the refinement of the web application based on it.
The flexible structure of the *DBN* will in fact allow the easy introduction of other variables: the most interesting variables to be exploited are diabetic Retinopathy, as an

additional outcome of the model, and the genotypic information as a potential mean to improve predictions.

A cost-effectiveness analysis to evaluate costs and consequences of possible treatments, as well as a cost-utility analysis to quantify eventual improvements in the patients' quality of life, will be implemented, in order to better address the supporting function of the web application in the decision analysis process.

Chapter 4

Modeling the effect of treatment on diabetes phenotype: a compartmental model of aspirin action

Referring to the multi-level scheme presented in Figure 1.1, this chapter will focus on the effect of treatment on phenotype, as shown in Figure 4.1.

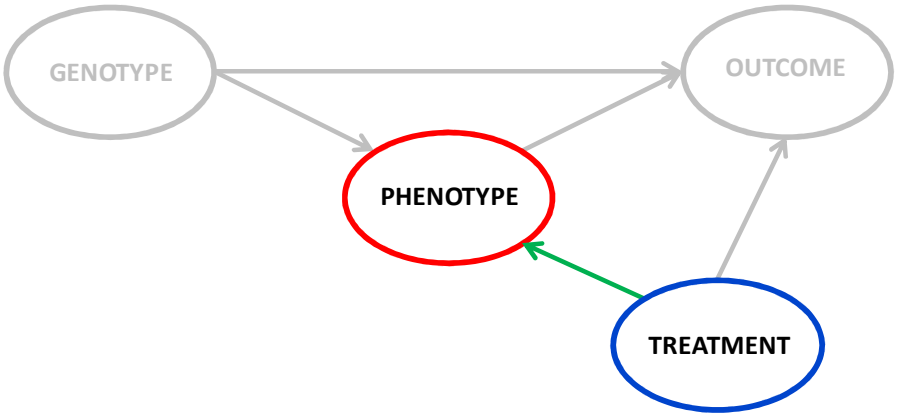


Figure 4.1: *Modeling the effect of treatment on phenotype.*

Aspirin represents an important component of cardiovascular prevention in diabetic patients. The biological background regarding the physiological mechanisms of action of aspirin as antiplatelet agent will be firstly introduced, then the most relevant results from clinical trials and epidemiological studies of aspirin as a therapy for patients at high cardiovascular risk will be shown. A compartmental model of aspirin action developed to qualitatively explain experimental evidence will be finally presented and its performance evaluated by a sensitivity analysis approach.

4.1 *Introduction*

4.1.1 Atherothrombosis

Atherosclerosis is a chronic inflammatory disorder in which immune mechanisms interact with metabolic risk factors to initiate, propagate, and activate vascular lesions, and represents the major cause of ischemic coronary artery disease and cerebrovascular disease [104].

Arterial thrombosis, an acute complication that develops on the surface of a ruptured atheromatous plaque or as a consequence of endothelial erosion, may cause myocardial infarction or ischemic stroke. Platelets are key cellular components of arterial occlusive thrombi and may participate in the development and progression of atheromatous plaques [131].

Platelets originate from megakaryocytes in bone marrow and are vital components of hemostasis, the physiologic process that arrests hemorrhage after tissue trauma and vascular injury. Although the adhesion and activation of platelets can be viewed as a repair-oriented response to sudden fissuring or rupture of an atheromatous plaque, uncontrolled progression of such a process through a series of self-sustaining amplification loops may lead to the intraluminal formation of thrombus, vascular occlusion, and transient ischemia or infarction. The ability of platelets to participate in both normal hemostasis and atherothrombosis depends on their adhesive properties and their capacity to become activated very quickly in response to various stimuli [131].

Currently available antiplatelet drugs, such as aspirin, interfere with certain steps in the activation process by selectively blocking key platelet enzymes or receptors, reducing the risk of arterial thrombosis through mechanisms that cannot be dissociated from an increased risk of bleeding complications [121].

In particular, randomized trials indicate that low-dose aspirin can prevent arterial thrombosis under various circumstances, including first vascular events among low-risk, healthy subjects and recurrent vascular events among patients with known acute or chronic occlusive vascular disease [121]. However, a diminished responsiveness has been reported in patient with type 2 diabetes [127], [100], [101], with the suggestion that this

might explain the apparent failure of the drug to reduce the risk of atherothrombotic events in individual trials and meta-analysis of aspirin in diabetes [94], [115], [98], [125]. A more comprehensive picture of the main results from clinical trials will be presented in section 4.2.

4.1.2 Aspirin as antiplatelet agent

Aspirin, also known as *acetylsalicylic acid* ($C_9H_8O_4$), is a salicylate drug belonging to a group of medications called nonsteroidal anti-inflammatory drugs (NSAIDs). It was first synthesized by Felix Hoffman, a chemist with the German company Bayer, in 1897, even if the active metabolite of aspirin, *salicylic acid*, was first extracted from the bark of the willow in 1763 by Edward Stone of Wadham College, Oxford University. Today, aspirin is one of the most widely used medications in the world as an analgesic to relieve minor aches and pains, as an antipyretic to reduce fever, and as an anti-inflammatory medication, with an estimated 40,000 tones of it being consumed each year [139].

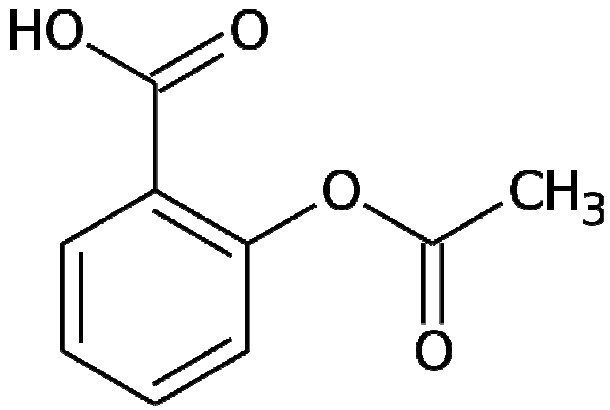


Figure 4.2: *Structural formula of aspirin.*

The best-characterized mechanism of action of aspirin is the inhibition of thromboxane-dependent platelet function, through permanent inactivation of the cyclooxygenase (COX) activity of prostaglandin H synthase 1 (also referred to as COX-1) [122]. As shown in Figure 4.3, prostaglandin H synthases, which have both cyclooxygenase and hydroperoxidase (HOX) activity, converts arachidonic acid (AA), a precursor primarily involved in cellular signaling and inflammatory process, to a complex set of derivatives which are collectively known as the ‘arachidonic acid cascade’. One of the final product of the cascade is thromboxane, an enzyme which stimulates platelets to produce the

coagulation factors as well as increasing platelet aggregation (the enzyme is in fact named for its role in clot formation, i.e. thrombosis).

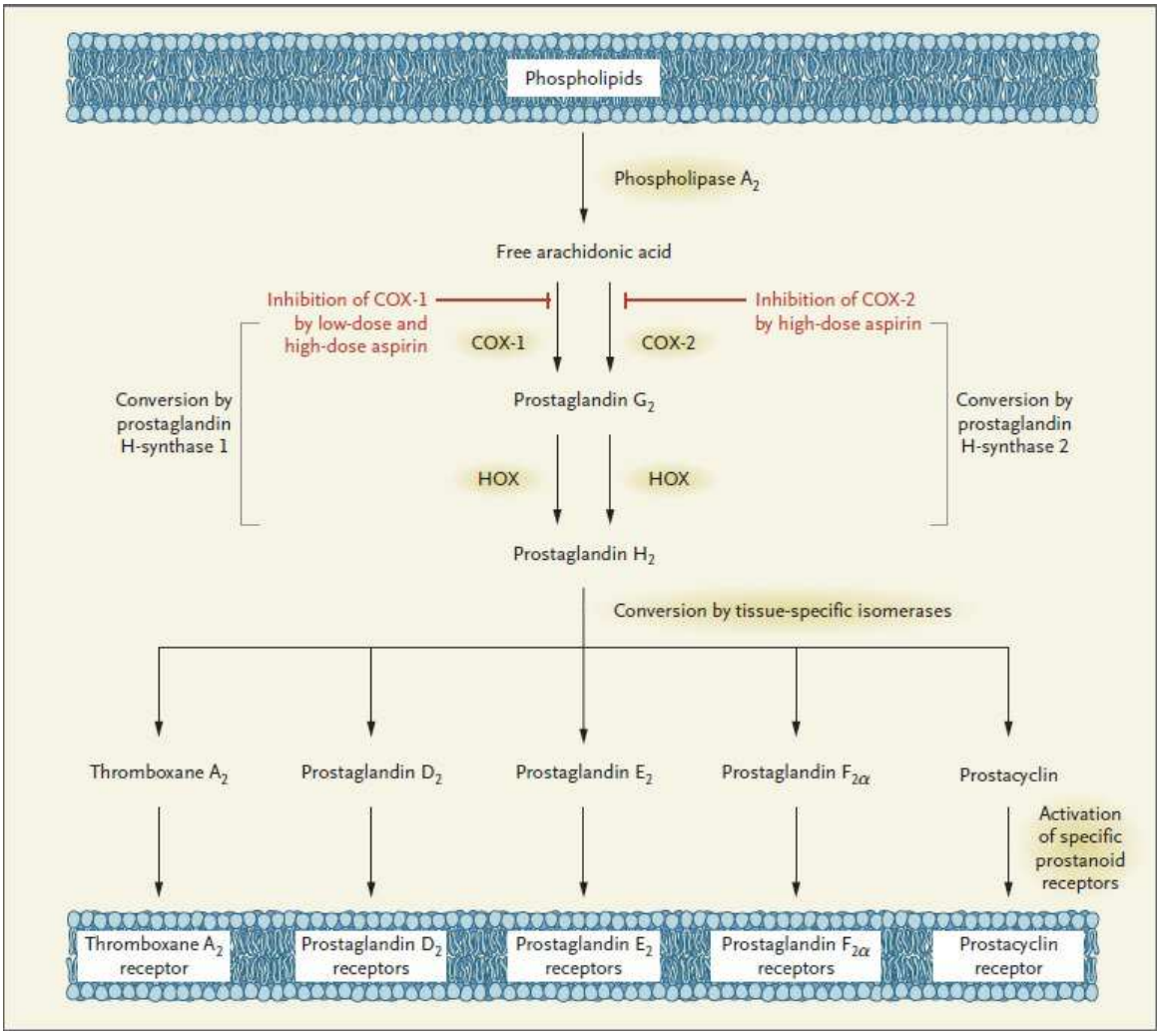


Figure 4.3: *Mechanism of action of aspirin on the arachidonic acid cascade. Figure from [122].*

The synthases are colloquially termed cyclooxygenases and exist in two forms, cyclooxygenase-1 (COX-1), which is the constitutive form of the enzyme, and cyclooxygenase-2 (COX-2), which is an inducible form. Low-dose aspirin mostly inhibits COX-1, whereas high-dose aspirin inhibits both COX-1 and COX-2 [122]. In particular, by diffusing through cell membranes, aspirin enters the COX channel, a narrow hydrophobic channel connecting the cell membrane to the catalytic pocket of the enzyme. Aspirin acts on COX-1 permanently inactivating it, trough an irreversible acetylation process: a single molecule of aspirin reacts with a single molecule of COX-1 producing a

single molecule of salicylic acid and a single molecule of acetylated COX-1 (Figure 4.4), thus preventing AA to bind the catalytic site of the enzyme to start the AA cascade [140].

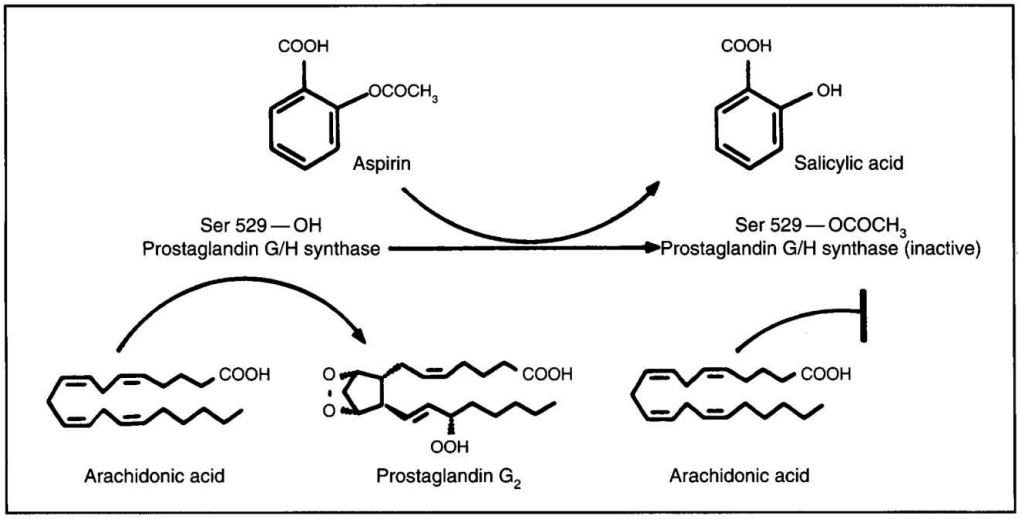


Figure 4.4: *Reaction between aspirin and COX-1: aspirin acetylates the hydroxyl group of a serine residue at position 529 (Ser529) in the polypeptide chain of platelet prostaglandin G/H synthase, thus inactivating the cyclooxygenase catalytic activity of the enzyme which leads to formation of prostaglandin G_2 from arachidonic acyd. Figure from [140].*

This process is irreversible and its effect is long-lasting for the entire single platelet lifespan, since platelets are not able to synthesize *de novo* COX-1 and, thus, only new platelet generation from megakaryocytes in bone marrow can recover pre-aspirin COX-1 levels [140].

In the next section, exemplifying results from clinical trials of aspirin in cardiovascular prevention will be briefly presented.

4.2 *Results from clinical trials*

In the context of the multi-level analysis adopted in this thesis, the main clinical trials of aspirin can be separated in 2 classes, on the basis of the final end-points considered:

1. *Trials focusing on the outcome* (Figure 4.5.A), in which the goal of the study is to analyze the effect of aspirin on cardiovascular events;

2. *Trials focusing on the phenotype* (Figure 4.5.B), in which the goal of the study is to characterize the effect of aspirin on COX-dependent platelet activity.

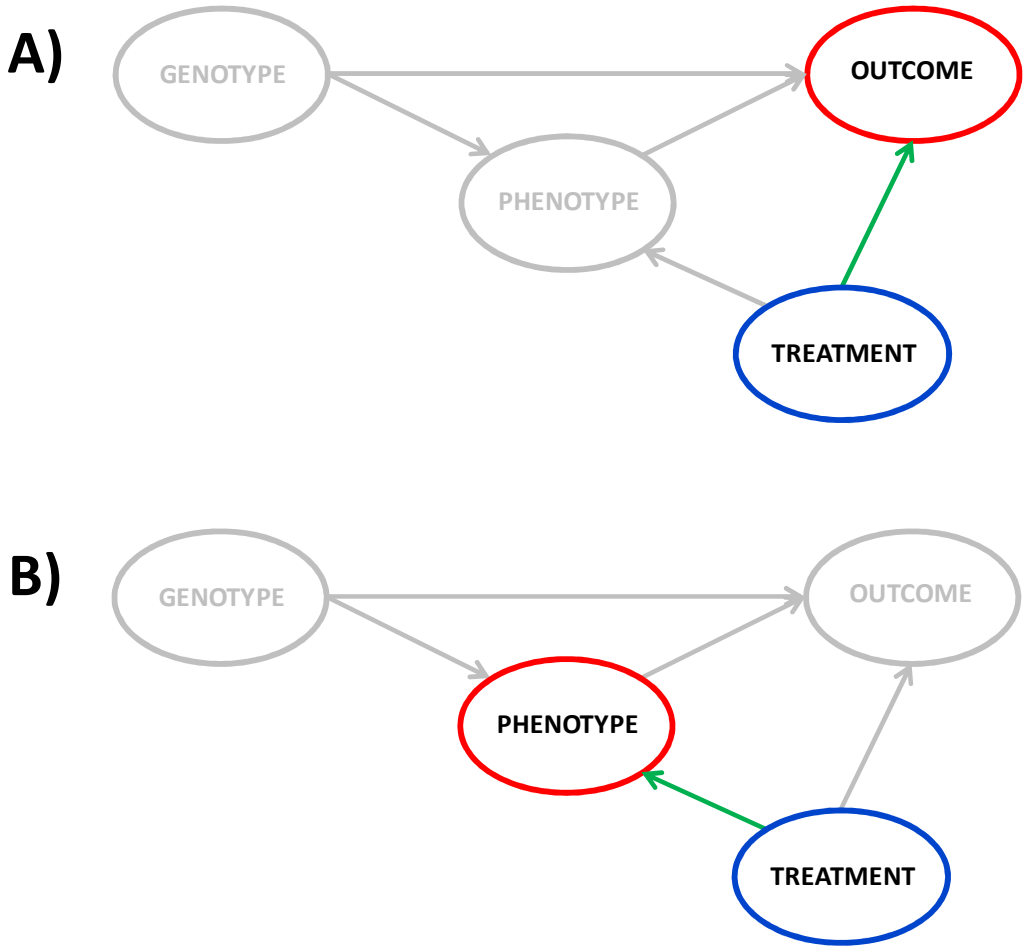


Figure 4.5: *Scheme for trials focusing on the outcome (A) and for trials focusing on the phenotype (B).*

4.2.1 Trials focusing on the outcome

The efficacy and safety of aspirin on non diabetic patients are document from analysis of many randomized clinical trials that included patients at variable risk of thrombotic complications of atherosclerosis [120]. Aspirin has been tested in patient demonstrating the whole spectrum of atherosclerosis, from apparently healthy low-risk individuals to patients presenting acute vascular events. Among patients with occlusive vascular disease, both individual studies and meta-analysis of trials of antiplatelet therapy indicate that aspirin significantly reduces the risk of a serious vascular event (nonfatal myocardial, infarction, nonfatal stroke, or death from vascular causes) [122]. For example, in [137] a meta-analysis of 287 studies involving 135 000 patients in comparisons of aspirin therapy

versus control, showed that among a wide range of patients with vascular disease, for which the annual risk of a serious vascular event ranges from 4 to 8 percent, aspirin significantly prevented at least 10 to 20 fatal and nonfatal vascular events for every 1000 patients treated for one year (Figure 4.6).

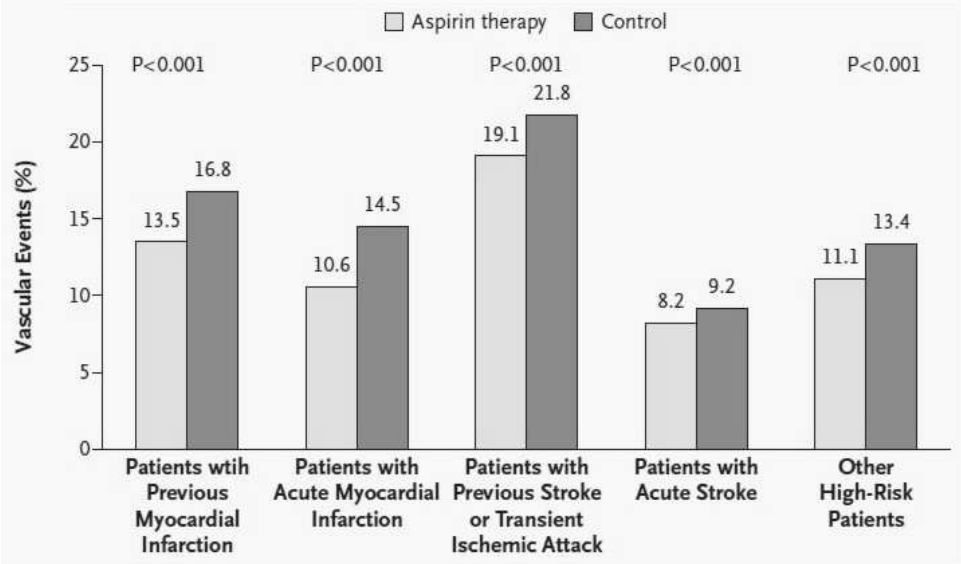


Figure 4.6: *Absolute effects of antiplatelet therapy with aspirin on the risk of vascularеEvents (nonfatal myocardial infarction, nonfatal stroke, or death from vascular causes) in five groups of high-risk non diabetic patients. Figure from [122].*

The inhibition of thromboxane-dependent platelet function by aspirin may lead to the prevention of thrombosis as well as to excess bleeding, thus assessing the net effect requires an estimation of the absolute thrombotic risk versus the hemorrhagic risk of the individual patient. In [121], aspirin has been evaluated in six primary prevention trials of aspirin versus placebo (the Primary Prevention Project trial on high-risk men and women [99], the Hypertension Optimal Treatment trial on hypertensive patients [105], the Thrombosis Prevention Trial on high-risk men [136], the Swedish Angina Pectoris Aspirin Trial on stable angina patients [112], the Physicians' Health Study trial on healthy men [110] and the United Kingdom Doctors trial on healthy men [124]) for a total of approximately 58000 patients who were at variable cardiovascular risk. Results show that as the risk of experiencing a major vascular event increases, so does the absolute benefit of antiplatelet prophylaxis with aspirin for a number of clinical conditions, including

stable and unstable angina pectoris and patients who suffered a myocardial infarction (Figure 4.7).

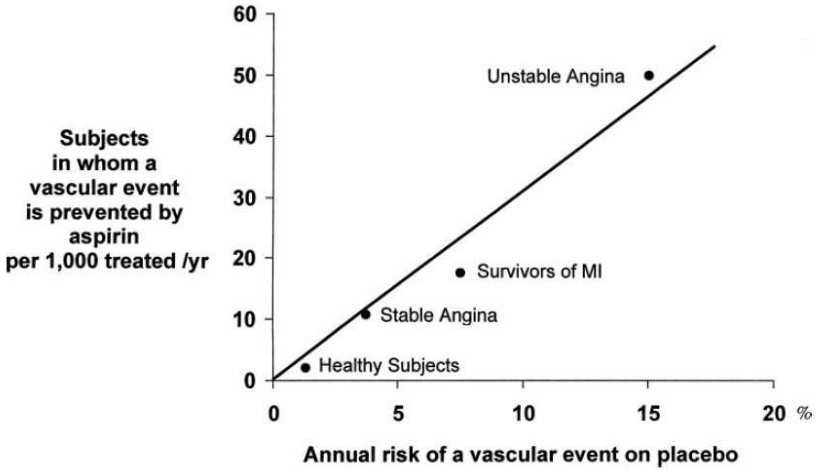


Figure 4.7: *For each category of patients, the abscissa denotes the absolute risk of experiencing a major vascular event. The absolute benefit of antiplatelet treatment is reported on the ordinate as the number of subjects in whom an important vascular event is actually prevented by treating 1,000 subjects with aspirin for 1 year. Figure from [121].*

In contrast to non-diabetic subjects, for which aspirin has been proofed to have a significant effect, a clear benefit of aspirin in the prevention of major cardiovascular events in people with diabetes remains unproved [98].

For example, in order to examine the efficacy of aspirin for the primary prevention of atherosclerotic events in patients with type 2 diabetes, Ogawa and Nakayama studied results from the Japanese Primary Prevention of Atherosclerosis With Aspirin for Diabetes (JPAD) trial [115], a randomized controlled trial in which patients were randomly assigned to assume low-dose aspirin (81 or 100 mg per day) or not. End-points were atherosclerotic events, including fatal or nonfatal ischemic heart disease, fatal or nonfatal stroke, and peripheral arterial disease as well as death from any cause. The incidence of the primary end point of any atherosclerotic event was not significantly different in the aspirin group than in the non-aspirin group (log-rank test, *p-value* = 0.16), as shown in Figure 4.8, thus the authors concluded that aspirin as primary prevention did not reduce the risk of cardiovascular events [115].

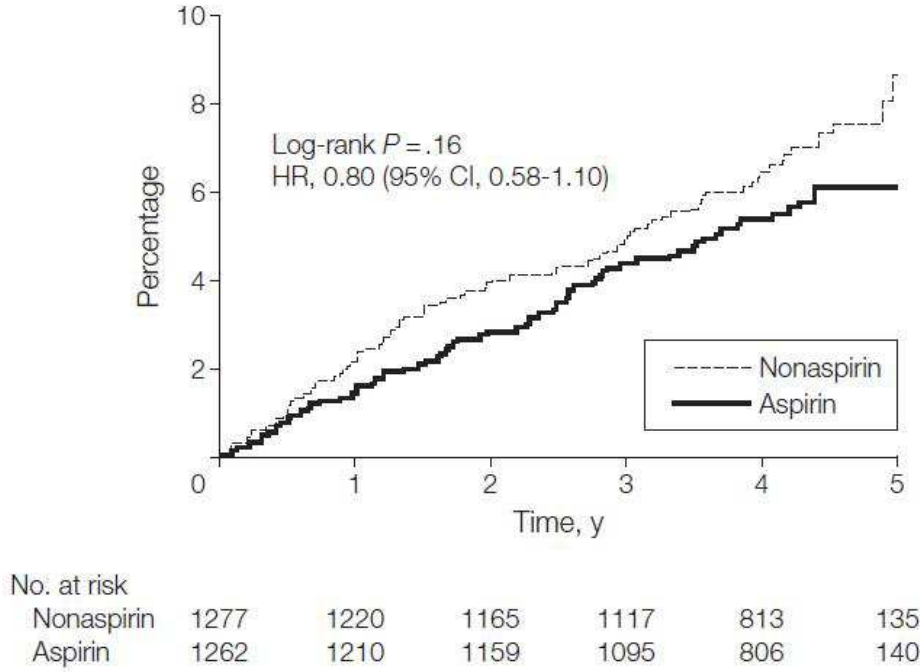


Figure 4.8: *Total percentage of atherosclerotic events according to treatment group in the JPAD trial. CI indicates confidence interval; HR, hazard ratio. Figure from [115].*

Similar results were obtained by Belch from the analysis of results from the Prevention of Progression of Arterial Disease and Diabetes (POPADAD) trial [94], a multicentre, randomized, placebo controlled trial conducted to determine whether aspirin was more effective than placebo in reducing the development of cardiovascular events in patients with diabetes mellitus and asymptomatic peripheral arterial disease. Two hierarchical composite primary end points of death from coronary heart disease or stroke, non-fatal myocardial infarction or stroke, or amputation above the ankle for critical limb ischemia, and death from coronary heart disease or stroke were the main outcomes measured. Overall, the authors concluded that specific adverse events were not significantly different between the aspirin and no-aspirin groups [94].

Pignone and Alberts performed a meta-analysis that added data from three trials performed specifically in patients with diabetes (the already mentioned JPAD [115] and POPADAD [94], and the Early Treatment of Diabetic Retinopathy Study [113]) to the data from subgroups of patients with diabetes from the six large trials of aspirin for primary prevention in the general population investigated also in [121], as already described. Using a random-effect model, the authors found that aspirin was associated

with a 9% decrease in risk of coronary heart disease events (non fatal and fatal myocardial infarction) and with a 15% decrease in the risk of stroke, both decreases not being statistically significant (Figure 4.9). The authors concluded that aspirin likely produces a modest reduction in CVD risk in patients with diabetes, but not statistically significant compared to diabetic patients not treated with aspirin [125].

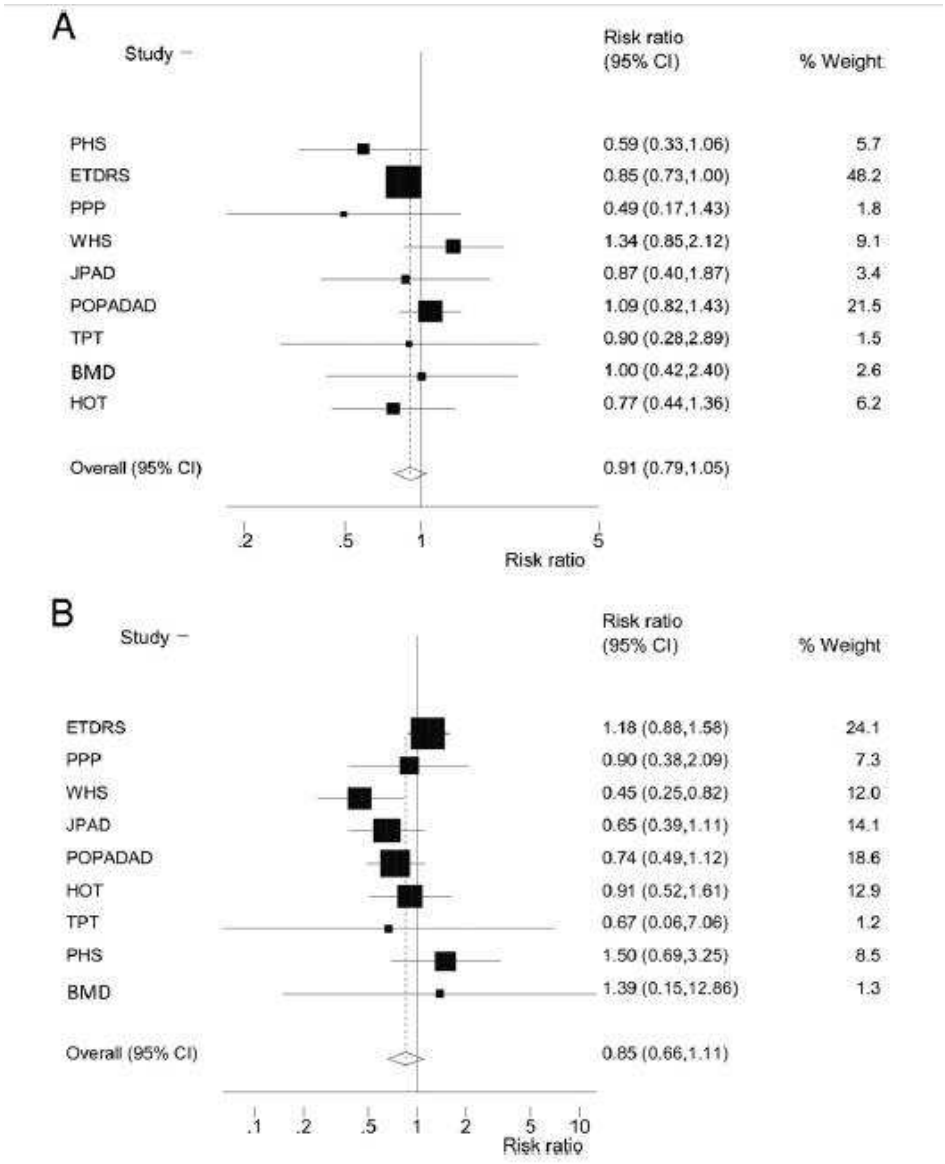


Figure 4.9: *Effect of aspirin on coronary heart disease events, tests for heterogeneity: P =0.367 (A). Effect of aspirin on risk of stroke in patients with diabetes, tests for heterogeneity: P =0.131 (B). CI stands for confidence interval. BMD indicates British Medical Doctors; ETDRS, Early Treatment of Diabetic Retinopathy Study; HOT, Hypertension Optimal Treatment; JPAD, Japanese Primary Prevention of Atherosclerosis with aspirin for Diabetes; PHS, Physicians' Health Study; POPADAD, Prevention of Progression of Arterial Disease and Diabetes; PPP, Primary Prevention Project; TPT, Thrombosis Prevention Trial; and WHS, Women's Health Study. Figure from [125].*

In the medical literature, the interindividual variability in response to aspirin, which brings to treatment failure, has been indicated with the term ‘aspirin resistance’ [118]. To be precise, today scientists argue about the term to be used, since the term ‘resistance’ implies that something can be measured that has a direct bearing on clinical efficacy of aspirin and that may lead to a change in the therapy. Since, so far, a such a relationship has not been discovered yet, the term ‘treatment failure’ may be more correct [102]. In this thesis, for simplicity, the term ‘aspirin resistance’ will be used henceforth.

4.2.2 Trials focusing on the phenotype

Moving from black box approach adopted by clinical trials focusing on the outcomes, in this sub-section the main results from clinical trials focusing on the phenotype will be presented. In particular, 2 trials conducted by Rocca and Santilli will be described; the former conducted on healthy subjects, and the latter on diabetic ones.

In both trials, the effect of aspirin on the activity of platelet COX-1 (referred to as simply COX henceforth) have been characterized through measurements of serum thromboxane B_2 (TxB_2), which is an indirect measure of the COX activity in serum [133]. More precisely, particular attention was paid to the recovery of serum TxB_2 during and after aspirin therapy.

4.2.2.1 Healthy subjects

In the first trial, the authors randomized 48 healthy Caucasian subjects to 1 to 8 groups, according to treatment duration, ranging from 1 to 8 weeks [133]. Each patient received enteric-coated aspirin 100 mg once a day and was instructed to take tablets at the same time of the day. Serum TxB_2 (in ng/ml) was measured (together with other blood and urine samples) at the end of each week of aspirin, and at days 1, 2, 3 and 7 after withdrawal. The authors found that:

- serum TxB_2 was steadily suppressed over 8 weeks, the average percent inhibition being constantly above 99% of the baseline, without significant intergroup differences: 1-week treatment caused 99.3% ±0.7% inhibition, and 8-weeks treatment produced 99.6 ± 0.3% inhibition (Figure 4.10).
- initial recovery of serumTxB_2 levels seem to differ among groups: at days 1 and 2 following aspirin withdrawal, TxB_2 values were similar in the subjects treated for 1 and 2 weeks and significantly higher than the corresponding values of longer

treatment groups (2-factor repeated measurements analysis of variance with the post hoc Holm-Sidak test for pairwise comparison, *p-value* < 0.05). Exposure to aspirin for at least 3 weeks showed a 2-day delay before detectable recovery (Figure 4.11);

- the overall kinetics of TxB_2 recovery showed a complex sigmoidal pattern, not appropriately described by a simple first-order kinetics (Figure 4.11).

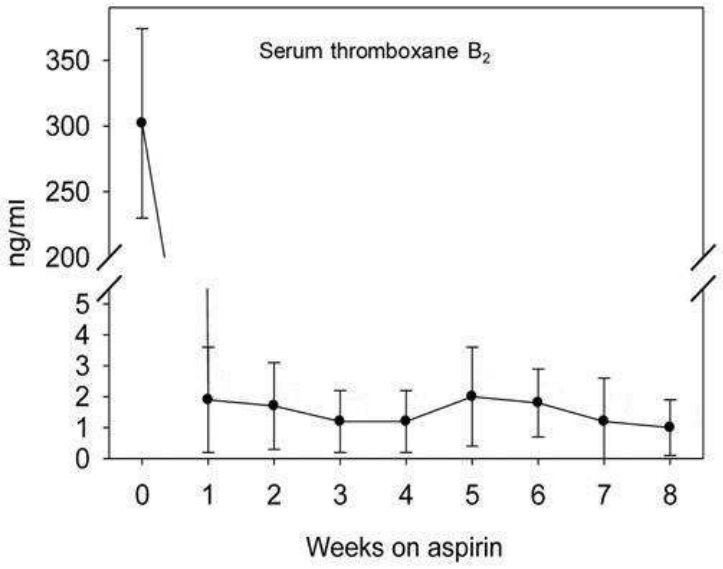


Figure 4.10: *Absolute values of TxB_2 (mean ± sd) of baseline (week 0) and for each week of treatment. Figure from [133].*

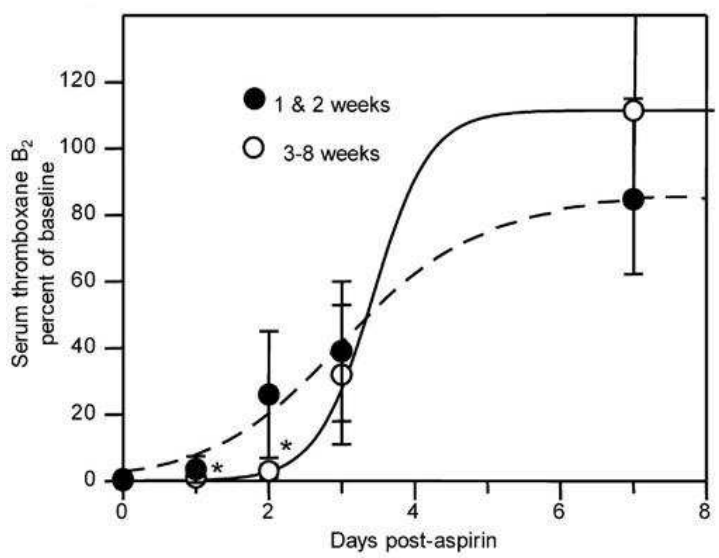


Figure 4.11: *TxB_2 data pooled from 1 to 2 weeks versus 3 to 8 weeks of treatment for the whole post-treatment period.* indicates significant difference. Figure from [133].*

These findings made the authors conclude that:

- to suppress COX recovery for 2 days after aspirin withdrawal, at least 3 weeks of treatment 100 mg once daily are needed;
- the 2-days delay, exhibited by COX recovery after more than 2 weeks of treatment 100 mg daily, is interpreted as the evidence that aspirin acetylates COX in megakaryocytes, thus leading to generation of inactivated platelets within the first 2 days after aspirin withdrawal;

4.2.2.2 Diabetic subjects

In the second trial, the authors characterized the kinetics of platelet COX recovery in aspirin-treated diabetic (and non diabetic) patients [128]. In the first phase of the trial, one hundred type 2 diabetic patients on chronic aspirin 100 mg daily were studied and serum TxB_2 measured every 3 hours, between 12 and 24 hours after a witnessed aspirin intake. The linear slope of serum TxB_2 recovery between 12 and 24 hours was computed for each patient (Figure 4.13). Patients with the fastest TxB_2 recovery (i.e. the ones in the upper tertile of the slope distribution) underwent phase 2: they were randomized to aspirin 100 mg once a day, 200 mg once a day or 100 mg twice a day, for 28 days and TxB_2 was reassessed. The protocol scheme is represented in Figure 4.12. Results from the first versus second phase of the study are presented in Figure 4.14. The authors found that:

- the median serum TxB_2 concentration measured at 12 hours after aspirin dosing in the 100 diabetic patients was comparable to the median value reported in the first clinical trial on healthy subjects, treated with the same dose and formulation of aspirin;
- about one third of the 100 diabetic patients showed a COX recovery significantly higher than healthy subjects in the 12-24 interval after aspirin intake;
- a twice-daily regimen with 100 mg aspirin is significantly more effectiveness with respect to a once-daily regimen and a 200 mg once-daily regimen (Figure 4.14).

The authors, thus, concluded that:

- aspirin maximal effectiveness in the suppression of COX-dependent platelet function is not different between healthy and diabetic patients;
- the main difference between healthy subjects and a fraction of diabetic patients is represented by a faster COX recovery during the 12-24 hours dosing interval;

- inadequate thromboxane inhibition by low-dose aspirin can be corrected by a twice-daily regimen.

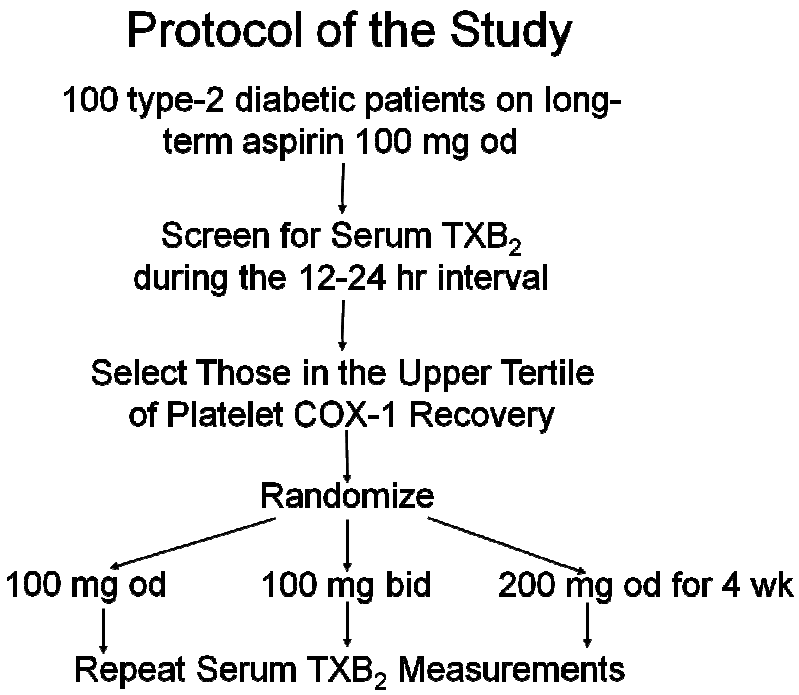


Figure 4.12: *Protocol of the study. Figure from [128].*

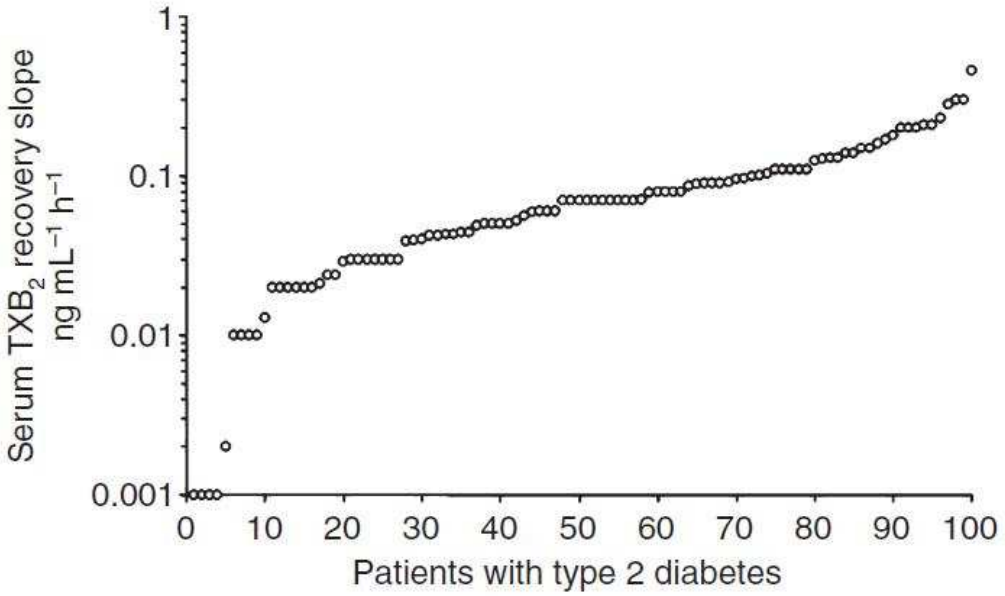


Figure 4.13: *Individual recovery slope of serum TxB_2 over the 12-24 hours interval of aspirin 100 mg once daily administration in patients with type 2 diabetes. Figure from [128].*

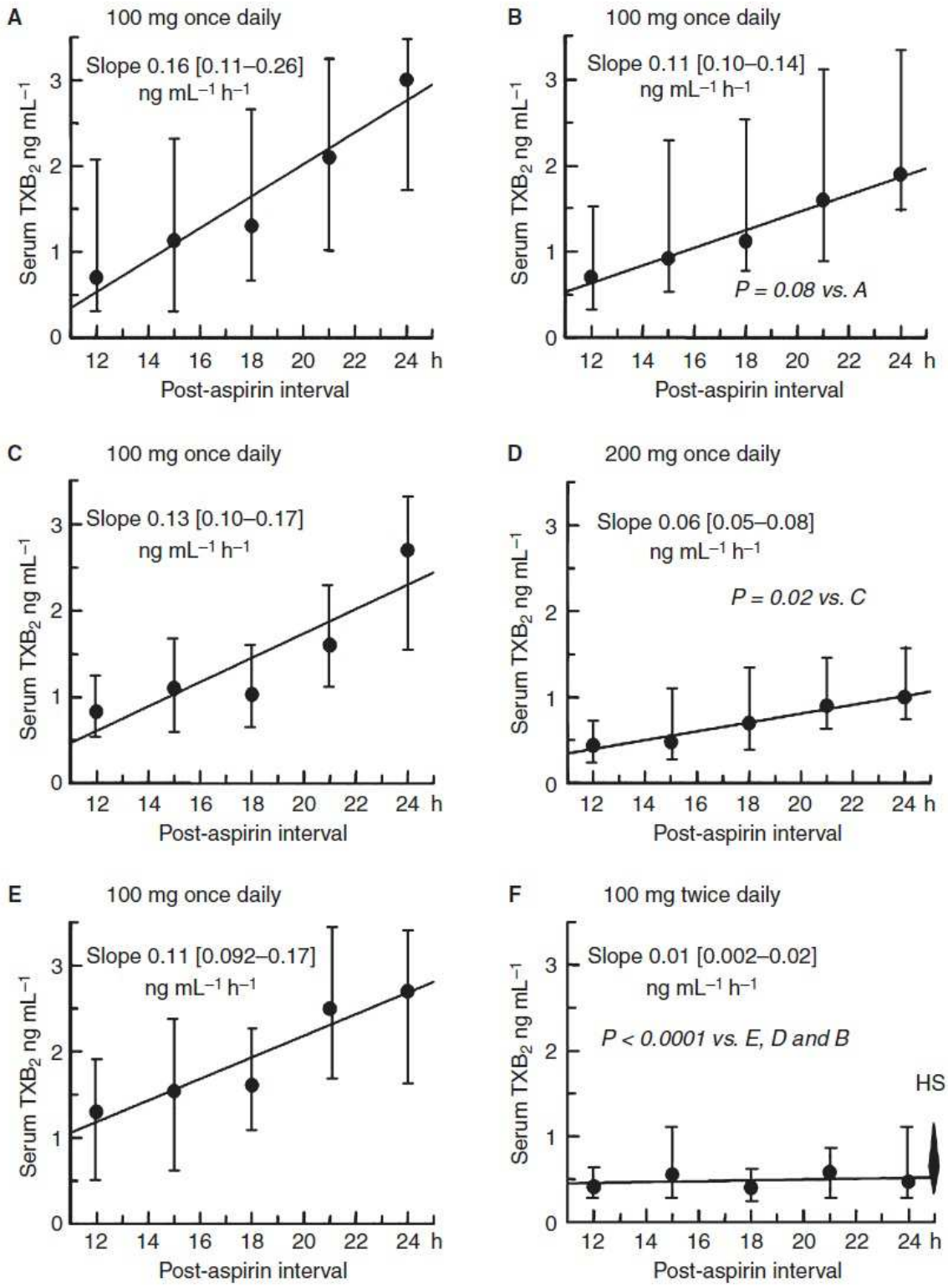


Figure 4.14: S*erum TxB_2 recovery slope between 12 and 24 hours after aspirin dosing in diabetic patients in the upper tertile, before (left panels) and after (right panels) the randomized phase of the study. (A-D) patients randomized to 100 mg once a day; (B-E) patients randomized to 200 mg once a day; (C-F) patients randomized to 100 mg twice a day. Figure from [128].*

4.2.2.3 Conclusions: potential mechanisms

Combining results from the 2 clinical trials described above, the authors concluded that the main difference between healthy and diabetic patients is represented by a faster COX recovery between 12 and 24 hours after aspirin intake and that this is likely the cause of the so-called 'aspirin resistance' leading to treatment failure in most diabetic patients treated with aspirin [128]. The authors hypothesize that this mechanism may be caused by an enhanced thromboxane biosynthesis in type 2 diabetics with macrovascular disease (as indicated by other literature works [97]), most likely reflecting variable platelet turnover. Moreover, they also hypothesize that a reduced systemic bioavailability of enteric-coated aspirin can limit the extent of COX acetylation in megakaryocytes [128].

4.3 *Objective of the work*

If, on one hand, a possible approach to the analysis of the 'aspirin resistence' problem is the same dynamic statistical analysis used for the *in silico* model of diabetes complications adopted in Chapter 3, on the other hand it is interesting to study the problem from a physiological point of view as well, searching for the biological mechanisms responsible for the different responses to drug, observed in experimental data.

Since, due to limited access to bone marrow megakaryocytes, it is difficult to clinically investigate both the causes of experimental evidence as well as the adequacy of different aspirin regimens, an *in silico* model of aspirin responsiveness can be useful to simulate interaction between aspirin and COX, and might help designing personalized antiplatelet regimens in T2DM.

Though some works have tried to explain this process from a mathematical point of view [119], [111], a detailed and complete characterization is still missing.

The object of the work presented in this chapter is, thus, to develop an *in silico* model of aspirin action, able to:

- explain data in healthy subjects;
- test hypothesis for faster recovery in diabetic patients;
- predict correct response to different aspirin regiments.

4.4 *Methods*

A compartmental model to describe and simulate the processes of COX inhibition and reappearance in serum in response to an aspirin therapy has been developed.
The model consists of four key-elements:

1. Thrombopoiesis mechanism
2. COX kinetics
3. Aspirin pharmacodynamics (PD)
4. Aspirin pharmacokinetics

The following section will describe, for each key-element, the hypotheses and assumption used to build the model.

4.4.1 Thrombopoiesis mechanism

The term thrombopoiesis refers to the process of thrombocyte generation, i.e. generation of platelets from megakaryocytes in bone marrow. The developed model of thrombopoiesis is based on available literature, in particular on the most recent work by Patel [116], [117] and Thon [135], [134].
According to most recent findings, platelets (PLTs) are generated from megakaryocytes (MKs) by fragmentation: each MK is generated in bone marrow by a precursor cell, and, after a megakaryocyte-maturation-period (*MK_matur*) during which each MK increases its dimension and becomes proliferative, i.e. able to generate PLTs. Each MK generates a certain number of ProPLTs (*N_ProPLTs_per_MK*), an intermediate form of platelet, over a subsequent time interval called megakaryocyte-proliferation-period (*MK_prolif*). The generation of ProPLTs takes place during the entire *MK_prolif*, until the complete fragmentation of the MK. Each ProPLT, although physically connected to the MK, is functionally disconnected, and, after a short period during which it stretches and elongates its structure, it detaches from the MK. The detached ProPlLT is a barbell cell, which, after a ProPlatelet-life-period (*PrePLT_life*), finally generates 2 Platelets (PLTs).
While megakaryocytes and proplatelets are in bone marrow, platelets are released in systemic circulation.
A MK is supposed to reach its mature state when it starts generating ProPLTs, and to die when the last ProPLT is generated. So, the Megakaryocyte life is given by the sum of the *MK_matur* and the *MK_prolif*.

For simplicity, ProPLTs are supposed to form consecutively over the *MK_prolif*, i.e., the number of ProPLTs generated by the MK in the time unit is a constant.

A ProPLT is supposed to form when it starts protruding from the MK and to die when it is divided into 2 Platelets.

A Platelet simply dies after the Platelet-life-period (*PLT_life*).

Figure 4.15 represents the chronologic order of the events for a single MK.

The population of MKs at the generic time *t* is supposed to be, with respect to the stage of maturation, without any privileged stage, as shown in Figure 4.16.

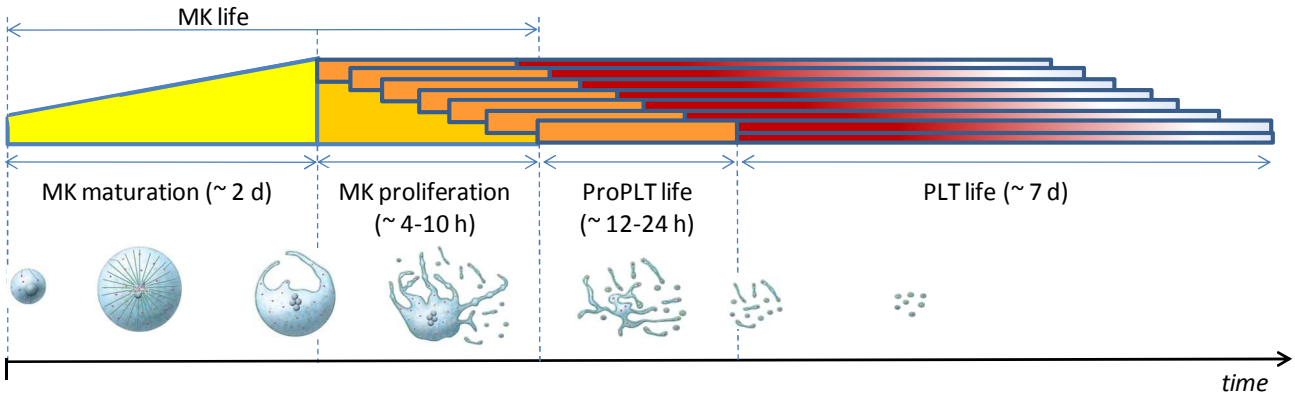


Figure 4.15: *timeline of the process of platelets generation from a single megakaryocyte. The different stages of the process, with representative literature values, are represented with different colors: megakaryocyte-maturation-period (yellow), megakaryocyte-proliferation-period (light orange), proplatelets-life-period (dark orange), and platelet-life-period (magenta).*

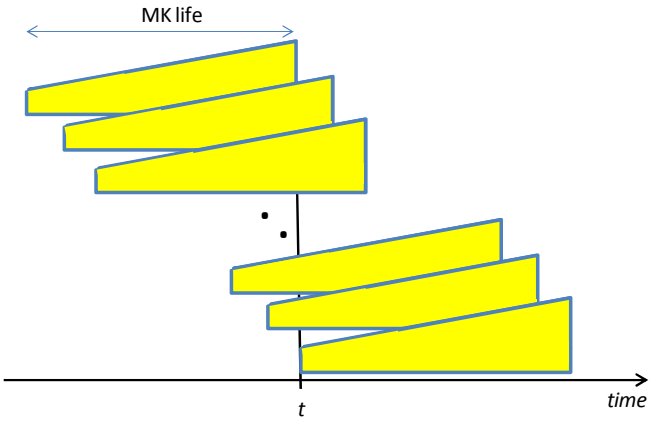


Figure 4.16: *Whole population of Megakaryocytes uniformly distributed over the interval 0 ÷ MK_life*

Table 4.1 lists the physiological parameters of the thrombopoiesis mechanism with literature ranges.

Parameter	Description	Literature value
N_MKs_per_Kg	Number of Megakaryocytes per kilogram of subject weight	~ $6.1e^6$ [106]
MK_matur	Maturation period of Megakaryocytes	~ 2-4 days [117]
MK_prolif	Proliferation period of Megakaryocytes	~ 4-10 hours [106]-[116]
ProPLT_life	Lifespan of ProPlatelets	~ 12-24 hours [117]
PLT_life	Lifespan of Platelets	~ 7-10 days [135]
N_ProPLTs_per_MK	Number of ProPlatelets generated from each Megakaryocyte	~ 1000 ÷ 1500 [116]
N_PLTs_per_ProPLT	Number of Platelets generated from each ProPlatelet	2 [116]

Table 4.1: *Literature values for the parameters of thrombopoiesis.*

4.4.2 COX kinetics

Since PLTs are not able to synthesize *de novo* COX [140], the enzyme is supposed to be constantly produced only within MKs during their maturation period. When the MK terminates its maturation and enters the proliferation period, COX synthesis is supposed to stop. Each ProPLT is supposed to inherit a certain amount of COX from its MK father. The total amount of COX within the MK is supposed to be equally distributed to all the ProPLTs generated from the MK, so that COX amount received by each ProPLT is the same. COX inherited from the MK remains inside the ProPLT throughout its life period, during which the ProPLT stretches and detaches from the MK, without any *de-novo* synthesis occurring. Then, when the ProPLT divides into 2 PLTs, COX is simply supposed to be equally divided between the 2 new-forming PLTs. COX degradation is supposed to be negligible in MKs and ProPLTs, since COX is a housekeeping enzyme (i.e. an enzyme present in all the cells to perform essential metabolic functions), while a nonzero degradation is supposed to take place in PLTs, representing enzyme utilization and elimination thought platelet death.

To model the processes of synthesis and transfer of COX from megakaryocytes in bone marrow to platelets in blood, accounting also for the temporal dimension of the processes involved in thrombopoiesis (see previous section), the *compartmental distributed model* of Figure 4.17 has been developed.

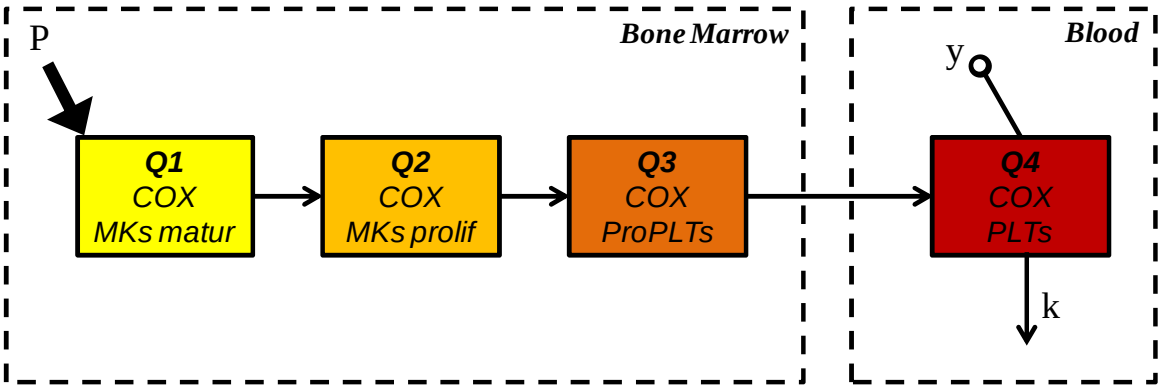


Figure 4.17: *The compartmental model for COX kinetics. Each compartment represent COX total amount in a different cell population: maturing megakaryocytes (yellow), proliferating megakaryocytes (light orange), proplatelets (dark orange) and Platelets (magenta). Solid arrows denote fluxes of COX: P represent COX synthesis, while k denote its degradation rate.*

The model is described by the following partial differential equations:

$$
\begin{cases}
\dfrac{\partial Q_1(s,t)}{\partial s} = P(s), & Q_1(0,t) = 0 \quad (4.1)\\[2ex]
\dfrac{\partial Q_1(s,t)}{\partial t} = 0, & Q_1(s,0) = Q_{10} \quad (4.2)\\[2ex]
COX_1(t) = \displaystyle\int_0^{MK_{matur}} Q_1(s,t)ds, & COX_1(0) = COX_{10} \quad (4.3)\\[2ex]
\dfrac{\partial Q_2(s,t)}{\partial s} = f[Q_2(s,t), s], & Q_2(0,t) = f_2[Q_1(s,t)] \quad (4.4)\\[2ex]
\dfrac{\partial Q_2(s,t)}{\partial t} = 0, & Q_2(s,0) = Q_{20} \quad (4.5)\\[2ex]
COX_2(t) = \displaystyle\int_0^{MK_{prolif}} Q_2(s,t)ds, & COX_2(0) = COX_{20} \quad (4.6)\\[2ex]
\dfrac{\partial Q_3(s,t)}{\partial s} = 0, & Q_3(0,t) = f_3[Q_2(s,t)] \quad (4.7)\\[2ex]
\dfrac{\partial Q_3(s,t)}{\partial t} = 0, & Q_3(s,0) = Q_{30} \quad (4.8)\\[2ex]
COX_3(t) = \displaystyle\int_0^{ProPLT_{life}} Q_3(s,t)ds, & COX_3(0) = COX_{30} \quad (4.9)\\[2ex]
\dfrac{\partial Q_4(s,t)}{\partial s} = -k \cdot Q_4(s,t), & Q_4(0,t) = f_4[Q_3(s,t)] \quad (4.10)\\[2ex]
\dfrac{\partial Q_4(s,t)}{\partial t} = 0, & Q_4(s,0) = Q_{40} \quad (4.11)\\[2ex]
COX_4(t) = \displaystyle\int_0^{PLT_{life}} Q_4(s,t)ds, & COX_4(0) = COX_{40} \quad (4.12)\\[2ex]
y = COX_4(t) & (4.13)
\end{cases}
$$

Where:

- *t* and *s* are the 2 independent variables, representing 'time' and 'cell maturation state' respectively. The unit of measure is [*time*] for both the variables, since 'cell maturation state' represents the age of the cell.
- Q_1, Q_2, Q_3, and Q_4 represent active-COX (i.e. non acetylated COX) distribution over *s* in, all the maturing megakaryocytes, all the proliferating megakaryocytes, all the proplatelets and all the platelets, respectively. The unit of measure for COX is [*mass*]. The Q_1, Q_2, Q_3, and Q_4 are time-dependent distributions, i.e. function of the 2 independent variables, and this is why the model is not lumped (i.e. dependent on one variable only) but distributed.
- Q_{10}, Q_{20}, Q_{30}, and Q_{40} represent the initial distribution over *s* of, all the maturing megakaryocytes, all the proliferating megakaryocytes, all the proplatelets and all the platelets, respectively.
- COX_1, COX_2, COX_3, and COX_4 represent active-COX total amount in, all the maturing megakaryocytes, all the proliferating megakaryocytes, all the proplatelets and all the platelets, respectively, and are function of time only, simply given by the integral over *s* of their respective distributions.
- COX_{10}, COX_{20}, COX_{30}, and COX_{40} represent the initial total amount of active-COX in, all the maturing megakaryocytes, all the proliferating megakaryocytes, all the proplatelets and all the platelets, respectively.
- *P(s)* represents the overall synthesis of new COX (which is supposed to take place within maturing megakaryocytes only) as a function of the maturation state *s*. *P* is considered to be constant over time. The unit of measure for *P* is [*mass/time*].
- *f* is a function expressing the partial derivative of the active-COX distribution in all the proliferating megakaryocytes (Q_2) with respect to the maturation state of the proliferating megakaryocytes *s*.
- the f_i (i = 1,2,3) are functions expressing the dependency of the initial value of the distribution over *s* (where 'initial' stands for 'in the initial maturation state', i.e. for *s* = 0) of compartment *i* on the previous variable state $Q_{i-1.}$
- *k* represents the degradation rate coefficient of COX. COX degradation is supposed to be a linear process taking place only in the platelets compartment, i.e. the degradation flux is proportional to COX amount in the compartment via a rate coefficient *k*, which

is supposed to be a constant. The unit of measure for k is [$time^{-1}$]. Information from [103] support an indicative value for COX half-life $t½$ of about 0.8÷1 day, i.e. a value for k equal to $ln(2) / t½ \sim 4.8e^{-4} \div 6.0e^{-4}$ min^{-1}.

- y represents the output of the model, i.e. the measure, which is the time-course of active-COX total amount in blood (i.e. in all the circulating platelets). The unit of measure for y is [*mass*].

Analytical expression for the $P(s)$, Q_{io}, COX_{i0}, f and f_i are derived from the physiological parameters of thrombopoiesis mechanism, shown in Table 4.1, and from the COX degradation rate coefficient k and the rate of new COX synthesis in the single megakaryocyte p_{MK}, following the rationale detailed in the following section

4.4.2.1 Mathematical formulation

To derive the mathematical formulation of the model, the steady-states of COX distributions separately for each cell type (i.e. maturing megakaryocytes, proliferating megakaryocytes, proplatelets and platelets) need to be considered and described. Analytical expression for the $P(s)$, Q_{io}, COX_{i0}, f and f_i will be highlighted in bold.

- Maturing megakaryocytes

COX kinetics within the single maturing MK is simply given by the constant production of COX taking place in the single MK (p_{MK}). Thus, the differential equation describing the COX time-course of the single maturing MK ($q_{MKm}(t)$) is given by:

$$\frac{dq_{MKm}}{dt} = p_{MK}, \qquad q_{MKm}(0) = 0 \qquad (4.14)$$

Since in the initial maturation state (s = 0) COX = 0, the analytical solution for $q_{MKm}(t)$ is algebraically described by a linear equation (see Figure 4.18):

$$q_{MKm}(t) = p_{MK} \cdot t \qquad t = 0 \div MK\_prolif \qquad (4.15)$$

Assuming the population of MKs at the generic time t to be distributed, with respect to the state of maturation, without any privileged state (see Figure 4.16), we can state that the number of MKs in the single maturation state (N_MKs_per_s) is a constant given by:

$$N\_MKs\_per\_s = \frac{N\_MKs}{MK\_life} \tag{4.16}$$

where N_MKs is the total number of MKs in bone marrow.

Given the hypothesis of uniform distribution, we can also state that, in steady-state, COX distribution of maturing MKs over the maturation state *s* at the generic instant *t* (Q_{10}) coincides with the time-course of q_{MKm} (equation (4.15)) multiplied by the number of MKs in each maturation state:

$$\boldsymbol{Q_{10}} = \frac{N\_MKs}{MK\_life} \cdot p_{MK} \cdot s, \qquad s = 0 \div MK\_life \tag{4.17}$$

COX total amount in all the maturing MKs is given by the sum of COX of each single maturing MK, thus is computed integrating equation (4.17) over *s*, and, in steady-state, is given by:

$$\begin{aligned} \boldsymbol{COX_{10}} = & \int_0^{MK\_matur} \frac{N\_MKs}{MK\_life} \cdot p_{MK} \cdot s \; ds = \\ & = \frac{1}{2} \cdot \frac{N\_MKs}{MK\_life} \cdot p_{MK} \cdot MK\_matur^2 \end{aligned} \tag{4.18}$$

The overall synthesis of new COX in the generic state *s* is a flux given by the single MK production p_{MK} multiplied by the number of MKs in the maturation state *s*:

$$\boldsymbol{P(s)} = \frac{N\_MKs}{MK\_life} \cdot p_{MK} \qquad s = 0 \div MK\_matur \tag{4.19}$$

P(s) is a constant function defined in the interval *0 ÷ MK_matur*, since production is supposed to take place in maturing MKs only.

- Proliferating megakaryocytes

In the single proliferating MK, a simple constant flux takes place, this flux representing COX amount transferred to ProPLTs in the unit time, until the MK is completely devoid

of COX. Thus the COX time-course in the single proliferating MK ($q_{MKp}(t)$) is a linear function (see Figure 4.18):

$$q_{MKp}(t) = p_{MK} \cdot \frac{MK\_matur}{MK\_prolif} \cdot (MK\_prolif - t), \qquad t = 0 \div MK\_prolif \tag{4.20}$$

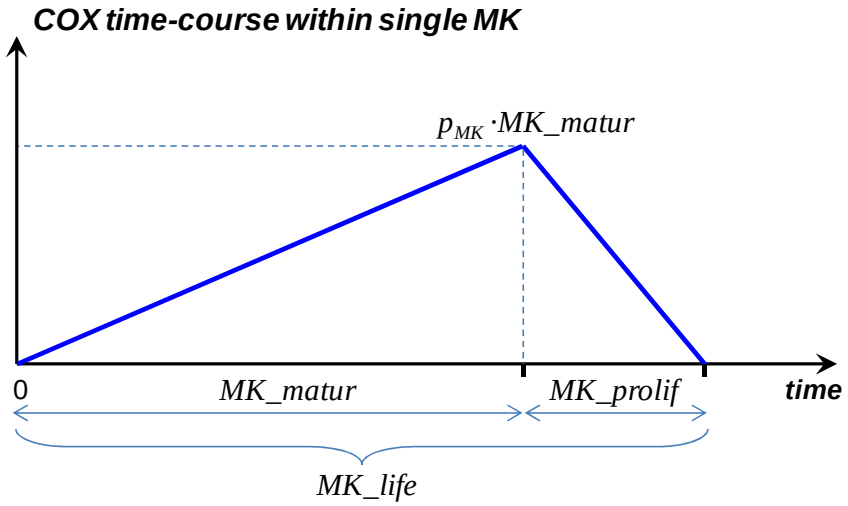


Figure 4.18: *COX time-course within the single megakaryocyte.*

Given the usual hypothesis of uniform distribution, we can state that, in steady-state, the distribution of proliferating MKs over the maturation state *s* at the generic time *t* (Q_{20}) is represented by the same linear function of the COX time-course in the single proliferating MK (equation (4.20)) multiplied by the number of MKs in each maturation state:

$$\boldsymbol{Q_{20}} = p_{MK} \cdot \frac{MK\_matur}{MK\_prolif} \cdot (MK\_prolif - s) \cdot \frac{N\_MKs}{MK\_life}, \qquad s = 0 \div MK\_prolif \tag{4.21}$$

Since the COX distribution of the single MK over its maturation state is a continuous function, denoting with s_1 the maturation state of the maturing MK and with s_2 the maturation state of the proliferating MK, for each generic time *t* it necessarily will be:

$$Q_1(s_1 = MK\_matur, t) = Q_2(s_2 = 0, t) \tag{4.22}$$

that is, the initial condition for Q_2 it is a function of Q_1, and, in particular, coincides with the final value of the Q_1 distribution:

$$Q_2(0,t) = \boldsymbol{f_2}[\boldsymbol{Q_1}(\boldsymbol{s},\boldsymbol{t})] = Q_1(s = MK\_matur, t) \tag{4.23}$$

COX total amount in all the proliferating MKs is given by the sum of COX of each single proliferating MK, thus is computed integrating equation (4.21) over *s*, and, in steady-state, is given by:

$$\boldsymbol{COX_{20}} = \int_0^{MK\_prolif} \frac{N\_MKs}{MK\_life} \cdot p_{MK} \cdot \frac{MK\_matur}{MK\_prolif} \cdot (MK\_prolif - s)\ ds = \\ = \frac{1}{2} \cdot \frac{N\_MKs}{MK\_life} \cdot p_{MK} \cdot MK\_matur \cdot MK\_prolif \tag{4.24}$$

COX total amount in all the MKs (maturing and proliferating), can be computed as the sum of equation (4.18) and equation (4.24):

$$COX_{MKs} = COX_{10} + COX_{20} = \frac{1}{2} \cdot N\_MKs \cdot p_{MK} \cdot MK\_matur \tag{4.25}$$

To derive the differential equation expressing the evolution of COX distribution in the proliferating MKs (Q_2(s,t)) in function of the maturation state s, the following assumptions were made:

- ProPLTs are constantly generated during the proliferation period of a MK, thus, the number of ProPLTs generating in each maturation state of the proliferating MK (N_new_ProPLTs_per_s) is supposed to be constant and equal to:

$$N\_new\_ProPLTs\_per\_s = \frac{N\_ProPLTs\_per\_MK}{MK\_prolif} \tag{4.26}$$

- COX within the single proliferating MK is considered as a whole amount which, at each maturation state *s*, will be uniformly distributed among the ProPLTs "to be generated yet" at sate *s*.

From equation (4.26), the number of ProPLTs "to be generated yet" (N_ProPLTs_tbg) is given by:

$$N\_ProPLTs\_tbg(t) = \frac{\left(MK_{prolif} - t\right)}{MK_{prolif}} \cdot N\_ProPLTs\_per\_MK \quad (4.27)$$

Given the usual hypothesis of uniform distribution for the MKs population, the distribution over the maturation state *s* coincides with the time-course within the single MK multiplied by the number of MKs per state, thus, at the generic state *s*, the COX amount leaving the proliferating MK is given by the COX total amount in *s*, divided by the number of ProPLTs "to be generated yet" in *s*, multiplied by the number of new ProPLTs generating in *s*:

$$\frac{Q_2(s,t)}{N\_ProPLTs\_tbg(s)} \cdot N\_new\_ProPLTs\_per\_s = \frac{Q_2(s,t)}{MK\_prolif - s}$$

and the differential equation expressing the dependency of Q_2(s,t) on the maturation state *s*, is given by:

$$\frac{\partial Q_2(s,t)}{\partial s} = \boldsymbol{f}[\boldsymbol{Q_2}(\boldsymbol{s},\boldsymbol{t}),\boldsymbol{s}] = \frac{Q_2(s,t)}{MK\_prolif - s} \quad (4.28)$$

In steady-state, Q_2(s,t) is given by equation (4.21), and equation (4.28) becomes:

$$\frac{\partial Q_2(s,t)}{\partial s} = p_{MK} \cdot \frac{N\_MKs}{MK\_life} \cdot \frac{MK\_matur}{MK\_prolif} \quad (4.29)$$

which yields to a linear equation for Q_2(s,t) over the maturation state s, as confirmed by equation (4.21).

In figure Figure 4.19, 2 examples are shown: in the left column, active-COX time-course within the single proliferating MK in steady-state; in the right column, active-COX time-course within the single proliferating MK in the case of an instant and partial inactivation of COX at $t = t^*$. For each column, top panel represents active-COX time-course within the single proliferating MK, middle panel the number of ProPLTs to be generated as a function of the time, bottom panel active-COX amount leaving the proliferating MK. While in steady-state, active-COX amount transferred to ProPLTs is a constant amount

(panel C), in the other case, one can see how active-COX amount transferred to ProPLTs decreases after the inactivation at $t = t^*$ (panel F).

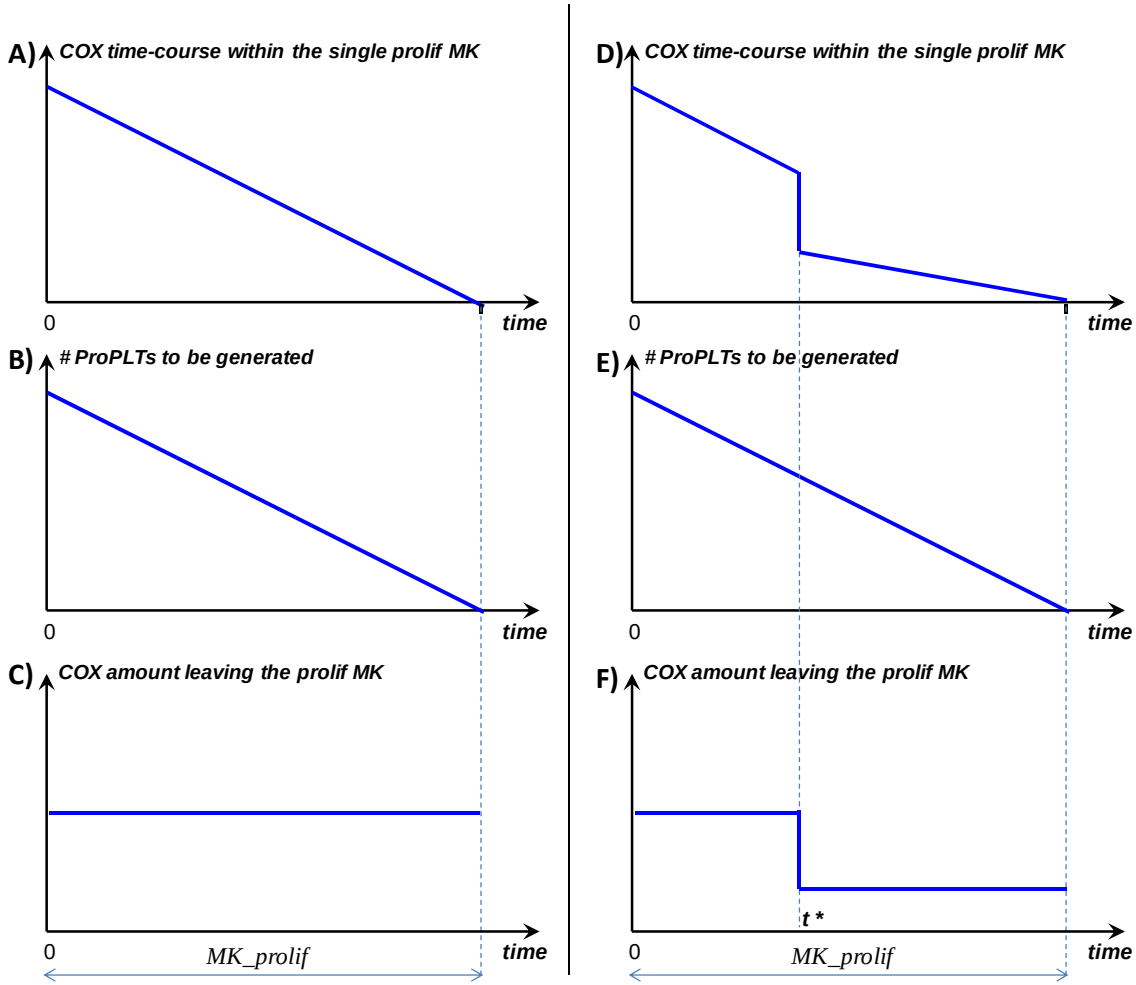


Figure 4.19: *Example of time-course of COX amount leaving the proliferating MK, in steady-state (left column), and in the case of an instant inactivation of COX at t = t* (right column). For each column, top panel represents active-COX time-course within the single proliferating MK, middle panel the number of ProPLTs to be generated as a function of the time, bottom panel active-COX amount leaving the proliferating MK.*

- Proplatelets

The single ProPLT inherits a certain amount of COX when it is generated by its father MK, and then stores that COX amount for all its life (ProPLT_life), until it splits into 2 PLTs.

At the generic time *t*, the single newborn ProPLT inherits a COX amount ($q_{ProPLT}(t)$) given by the total COX amount in its father proliferating MK at time *t* ($q_{MKp}(t)$) divided by the number of ProPLTs "to be generated yet" at time *t*:

$$q_{ProPLT}(t) = \frac{q_{MKp}(t)}{N\_ProPLTs\_tbg(t)} \tag{4.30}$$

The newborn ProPLTs are the ProPLTs whose state of maturation is $s = 0$.

The total number of newborn ProPLTs at each instant t (N_new_ProPLTs_s0) is a constant given by the number of proliferating MKs at time t multiplied by the number of ProPLTs generated in the unit time (which is the constant given by equation (4.26)):

$$N\_new\_ProPLTs\_s0 = \frac{N_{MKs}}{MK_{life}} \cdot N\_ProPLTs\_per\_MK \tag{4.31}$$

Since the ProPLTs population is uniformly distributed, equation (4.31) represents also the constant number of ProPLTs in each maturation state.
The COX amount in all the newborn ProPLTs ($Q_3(0,t)$) is given by the COX-amount inherited by the single ProPLT (equation (4.30)) multiplied by the total number of newborn ProPLTs (equation (4.31)), which yields to:

$$Q_3(0,t) = \frac{q_{MKp}(s,t)}{MK\_prolif - s} \cdot \frac{N\_MKs \cdot MK\_prolif}{MK\_life} \tag{4.32}$$

where $q_{MKp}(s)$ is the COX amount in the single proliferating MK, and it is simply given by the distribution of COX amount in the proliferating MKs over the maturation state s ($Q_2(s,t)$), divided by the number of MKs in each maturation state (equation (4.16)), which yields to the following expression:

$$Q_3(0,t) = \boldsymbol{f_3}[\boldsymbol{Q_2}(\boldsymbol{s},\boldsymbol{t})] = Q_2(s,t) \cdot \frac{MK\_prolif}{MK\_prolif - s} \tag{4.33}$$

In steady-state, $Q_2(s,t)$ is Q_{20}. Considering that ProPLTs population is uniformly distributed and that the distribution of COX in ProPLTs over the maturation state ($Q_3(s,t)$) is constant (see equation (4.7)), $Q_3(s,t)$ in steady-state (Q_{30}) is obtained by using the expression for Q_{20}, given by equation (4.21), in equation (4.33):

$$\boldsymbol{Q_{30}} = \frac{N\_MKs}{MK\_life} \cdot p_{MK} \cdot MK\_matur, \qquad s = 0 \div ProPLT\_life \tag{4.34}$$

COX total amount in all the ProPLTs is given by the sum of COX of each single ProPLT, and, in steady-state, is computed integrating equation (4.34) over *s*:

$$\begin{aligned} \boldsymbol{COX_{30}} = & \int_0^{ProPLT\_life} \frac{N\_MKs}{MK\_life} \cdot p_{MK} \cdot MK\_matur \; ds = \\ & = \frac{N\_MKs}{MK\_life} \cdot p_{MK} \cdot MK\_matur \cdot ProPLT\_life \end{aligned} \tag{4.35}$$

- Platelets

The single PLTs inherits half of the COX amount of its mother ProPLT. COX within PLTs is supposed to undergo a degradation process regulated by the degradation rate coefficient *k*. As for MKs and ProPLTs, also the total population of PLTs is uniformly distributed over the maturation state, meaning that the number of PLTs in each state of maturation is the same. Considering that the whole COX amount stored in all the ProPLTs at the end of their life (i.e. for *s* = ProPLT_life) is transferred to newborn PLTs (i.e. the set of all PLTs whose maturation state is *s* = 0), it is easy to compute the initial condition for the distribution of COX in PLTs over s ($Q_4(s,t)$) as a function of the distribution of the previous compartment:

$$Q_4(0,t) = \boldsymbol{f_4}[\boldsymbol{Q_3}(\boldsymbol{s},\boldsymbol{t})] = Q_3(s = ProPLT\_life, t) \tag{4.36}$$

The differential equation expressing the kinetics of $Q_4(s,t)$ over s is the partial differential equation (4.10), whose solution yields to the expression for the distribution of COX in PLTs over the maturation state s:

$$Q_4(s,t) = Q_4(0,t) \cdot e^{-k \cdot s} \qquad s = 0 \div PLT\_life \tag{4.37}$$

Using equation (4.34), the expression for $Q_4(s,t)$ in steady-state is:

$$\boldsymbol{Q_{40}} = \frac{N\_MKs}{MK\_life} \cdot p_{MK} \cdot MK\_matur \cdot e^{-k \cdot s} \qquad s = 0 \div PLT\_life \tag{4.38}$$

COX total amount in all the PLTs is given by the sum of COX of each single ProPLT, and, in steady-state, is computed integrating equation (4.38) over s:

$$\begin{aligned} \boldsymbol{COX_{40}} = & \int_0^{PLT\_life} \frac{N\_MKs}{MK\_life} \cdot p_{MK} \cdot MK\_matur \cdot e^{-k \cdot s} \, ds = \\ & = \frac{N\_MKs}{MK\_life} \cdot p_{MK} \cdot MK\_matur \cdot PLT\_life \cdot \frac{1 - e^{-k \cdot PLT\_life}}{k} \end{aligned} \tag{4.39}$$

Figure 4.20 shows a summary picture of COX kinetics within each single cell type: in the upper panel, COX kinetics within the single MK; in the middle panel, COX kinetics within the single ProPLT; in the bottom panel, COX kinetics within the single PLT.

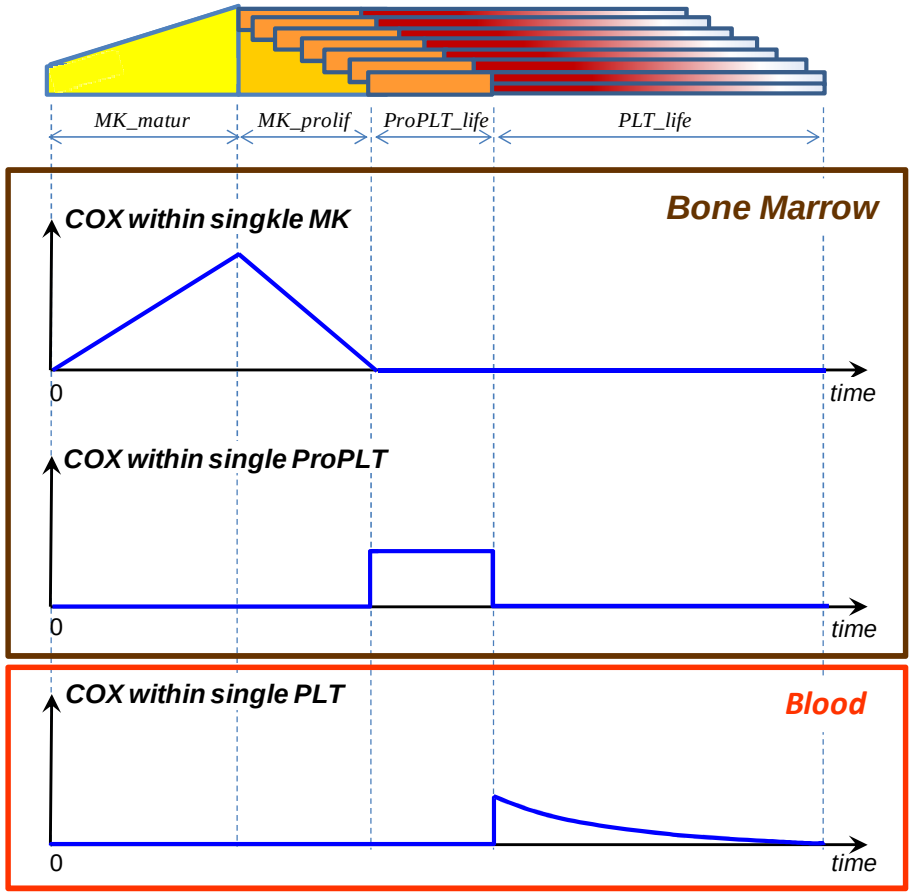


Figure 4.20: *COX kinetics within the single MK (top), the single ProPLT (middle) and the single PLT (bottom)*

4.4.2.2 Why a distributed model?

The model is distributed meaning that the state variables, i.e. COX amounts, depend not only on the time but also on the state of maturation of the cells in which COX is stored: the state variables are, in fact, time-dependent distributions.

The reason why a distributed model was implemented is that a lumped compartmental linear model misses the information on how COX is distributed among cells at different maturation states, thus not properly describing the timing of COX kinetics. For example, a lumped model is not able to explain a pure delay in the time-course recovery of active-COX in the case of a complete shooting-down of the COX in platelets precursors: in fact, if all the COX in the platelets and proplatelets is inactivated, there won't be any COX re-appearance in platelets before a $\Delta t = ProPLT\_life$, because new active COX, produced in megakaryocytes, needs to be transferred from megakaryocytes to proplatelets, and then has to wait a period equal to the life of a proplatelet before moving to platelets. A lumped model cannot reproduce these behavior. In Figure 4.22, an example in which COX is completely and instantly inactivated in proplatelets and platelets at *t = 1 day* is shown for the distributed model (blue curve) and for a lumped version of the model (red curve) developed on the same literature knowledge (Figure 4.21) and described by equations (4.40)-(4.43). In the example of Figure 4.22, *ProPLT_life* is set to 1 day.

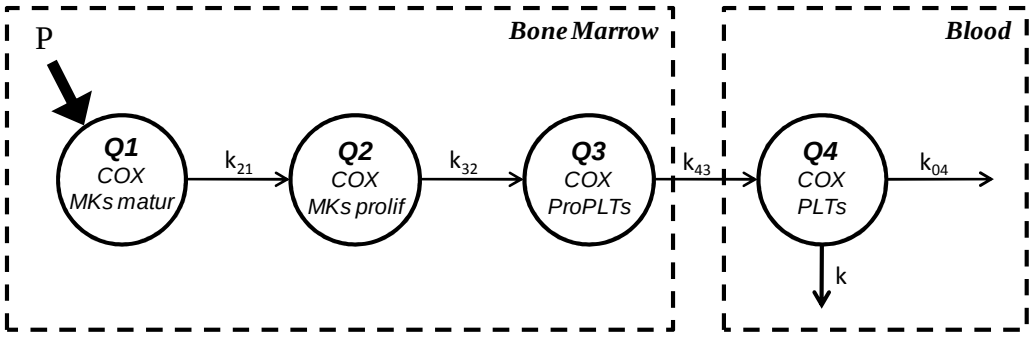


Figure 4.21: *Lumped version of the model for COX kinetics.*

$$\begin{cases} \dfrac{d(Q_1(t))}{dt} = P_1(t) - k_{21} \cdot Q_1(t), \quad Q_1(0) = Q_{10} & (4.40) \\ \dfrac{d(Q_2(t))}{dt} = k_{21} \cdot Q_1 - k_{32} \cdot Q_2(t), \quad Q_2(0) = Q_{20} & (4.41) \\ \dfrac{d(Q_3(t))}{dt} = k_{32} \cdot Q_2 - k_{43} \cdot Q_3(t), \quad Q_3(0) = Q_{30} & (4.42) \\ \dfrac{d(Q_4(t))}{dt} = k_{43} \cdot Q_3 - k_{04} \cdot Q_4(t), \quad Q_4(0) = Q_{40} & (4.43) \end{cases}$$

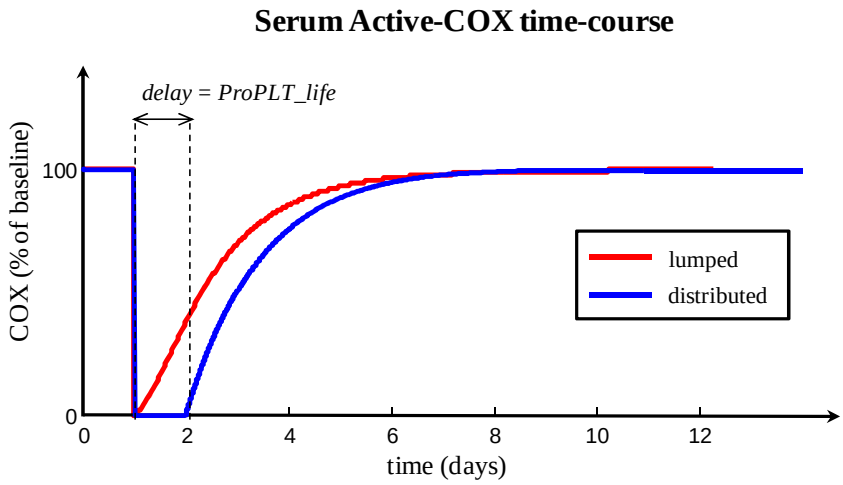


Figure 4.22: *comparison between a lumped model and a distributed one for a complete inactivation of COX in both ProPLTs and PLTs.*

4.4.3 Aspirin PD

Aspirin acts on COX permanently inactivating it, trough an irreversible acetylation process: a single molecule of Aspirin (ASA) reacts with a single molecule of COX producing a single molecule of salicylic acid (SA) and a single molecule of acetylated COX (COX^A) [140]. The reaction follows a first order kinetics [130] and it is regulated by the constant λ (equation (4.44)). λ's unit of measure is [$mass^{-1}time^{-1}$].

$$ASA + COX \xrightarrow{\lambda} SA + COX^A \tag{4.44}$$

Reaction (4.44) can be described by the mass-action law, which states that the speed of a reaction is proportional on the product of the molar concentrations of the reagents:

$$\begin{cases} \frac{d[ASA(t)]}{dt} = -\lambda \cdot [ASA(t)] \cdot [COX(t)] & (4.45) \\ \frac{d[COX(t)]}{dt} = -\lambda \cdot [ASA(t)] \cdot [COX(t)] & (4.46) \\ \frac{d[SA(t)]}{dt} = \lambda \cdot A[SA(t)] \cdot [COX(t)] & (4.47) \\ \frac{d[COX^A(t)]}{dt} = \lambda \cdot [ASA(t)] \cdot [COX(t)] & (4.48) \end{cases}$$

In the model of COX kinetics (Figure 4.17), a new flux appears from each compartment, representing COX acetylation by aspirin, as shown in Figure 4.23.

The model is still described by the old equations (4.1)-(4.13), with the only difference that, now, in the derivative of COX with respect to time in each compartment (equations

(4.2), (4.5), (4.8) and (4.11), respectively) a new term appears, representing degradation by aspirin, modeled as in equation (4.46). Equations (4.2), (4.5), (4.8) and (4.11) are then replaced by the followings:

$$\begin{cases} \dfrac{\partial Q_1(s,t)}{\partial t} = -\lambda \cdot [ASA(t)]_{BM} \cdot Q_1(s,t), & Q_1(s,0) = Q_{10} \quad (4.49) \\ \dfrac{\partial Q_2(s,t)}{\partial t} = -\lambda \cdot [ASA(t)]_{BM} \cdot Q_2(s,t), & Q_2(s,0) = Q_{20} \quad (4.50) \\ \dfrac{\partial Q_3(s,t)}{\partial t} = -\lambda \cdot [ASA(t)]_{BM} \cdot Q_3(s,t), & Q_3(s,0) = Q_{30} \quad (4.51) \\ \dfrac{\partial Q_4(s,t)}{\partial t} = -\lambda \cdot [ASA(t)]_{Bl} \cdot Q_4(s,t), & Q_4(s,0) = Q_{40} \quad (4.52) \end{cases}$$

where $[ASA(t)]_{BM}$ and $[ASA(t)]_{Bl}$ represent the time-course of aspirin molar concentration in bone marrow and blood, respectively.

According to equations (4.49)-(4.52), aspirin is supposed to act uniformly on COX, i.e. aspirin in blood will uniformly acetylate COX in all the PLTs, and aspirin in bone marrow will uniformly acetylate COX in all the MKs and the ProPLTs.

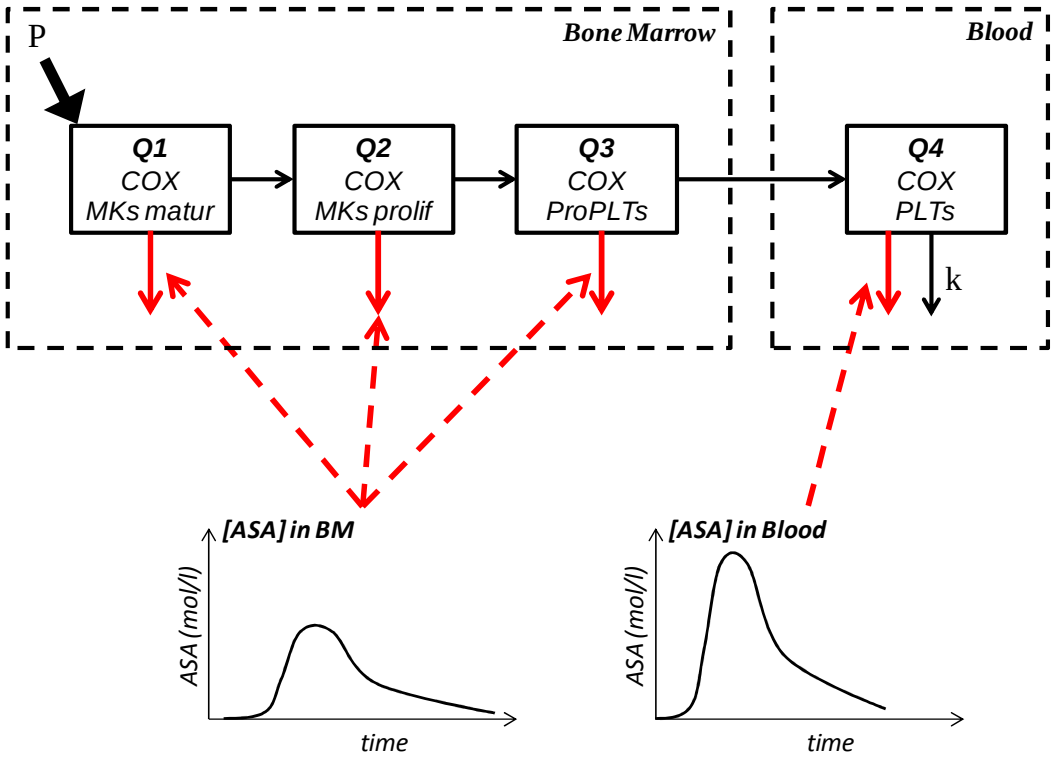


Figure 4.23: *Aspirin effect on COX kinetics: Aspirin contributes to COX elimination with a rate coefficient equal to λ·ASA(t). Aspirin concentration time-course in bone marrow acts on the compartments of maturing megakaryocytes, proliferating megakaryocytes and proplatelets; aspirin concentration time-course in blood acts on the compartment of platelets. The regulation is represented by dashed red arrows.*

4.4.4 Aspirin PK

Aspirin represents the intervention input of the model. As for the thrombopoiesis mechanism, information on the time-course of aspirin was derived from the literature.

A first issue to be faced when approaching aspirin PK is that aspirin can present 2 different formulations: simple compressed tablets or enteric-coated tablets. Enteric-coating of drug tablets is achieved by covering tablets with a polymer layer (usually a polyacid soluble in aqueous media), and is used to prevent the release of drugs in the stomach, either to reduce the risk of gastrointestinal side effects or/and to maintain the stability of drugs which are subject to degradation in the gastric environment [126]. Aspirin is enteric coated to protect gastric mucosa from corrosion, and this is particularly important for patients on chronic aspirin therapy.

These 2 different formulation of aspirin (to which we will refer as *ASA* for the non-coated formulation and *ecASA* for the coated one) lead to different kinetics. One of the first work conducted with the objective to compare different aspirin formulation is the one of Ali and others [93], where the authors enrolled six healthy subjects and measured, for each subject, ASA levels in plasma following a first ingestion of compressed tablets 650 mg aspirin, and a second ingestion, after a wash-out interval of 5 weeks, of enteric-coated tablets 650 mg aspirin. Figure 4.24 shows the mean time-course of the 2 different formulations: as one can see, ASA concentration (in μg/ml) reached its peak in plasma about 45 minutes after compressed tablet administration (upper left panel), and about 4 hours after enteric-coated administration (upper right panel), thus exhibiting a very delayed and slower kinetics in the enteric-coated formulation compared to the non-coated one. Moreover, the authors observed that both aspirin formulations resulted in widely variable ASA levels (as one can see by the error bars of Figure 4.24) and, most notably, ASA was undetectable in plasma during 3 experiments involving enteric-coated formulation [126]. These results are relevant since they highlight the issue of the interindividual variability in response to aspirin. The work provides more complete information, since the authors measured also platelets COX activity, as presented in Figure 4.24, where the mean COX time-course is shown for the enteric coated formulation (bottom right panel) and for the non-coated one (bottom left panel). The delayed kinetics of enteric-coated formulation is reflected in COX kinetics too, since COX recovery is delayed in response to the enteric-coated administration, even if, apart from the delay, COX time-courses appears similar after 24 hours in both cases.

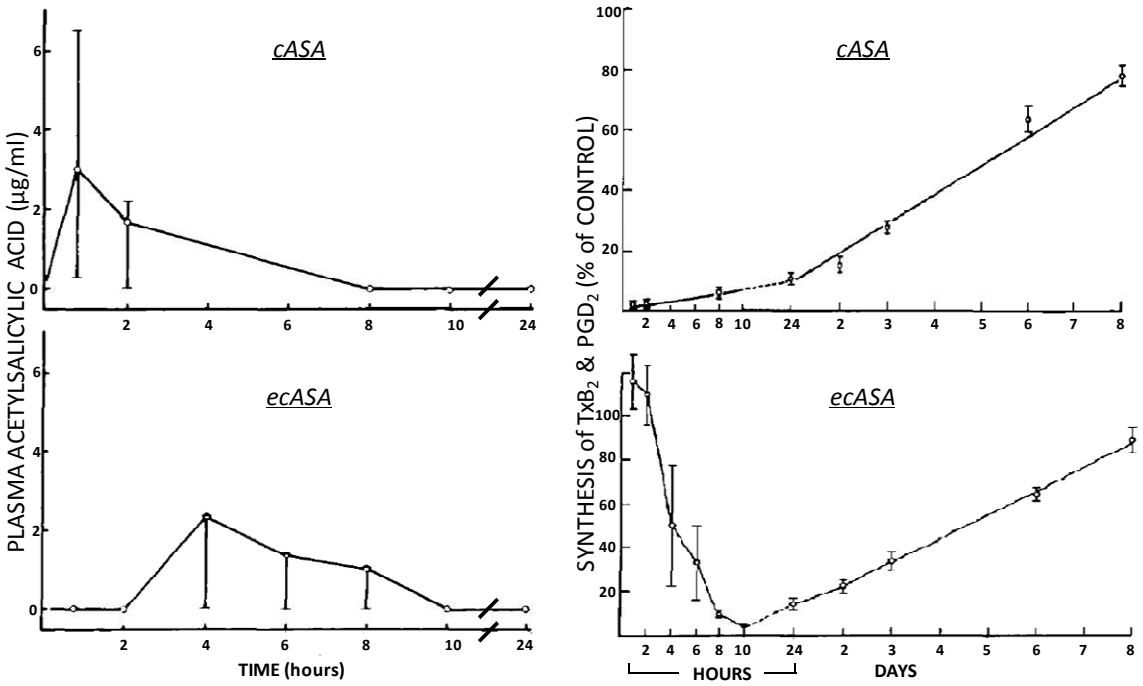


Figure 4.24: *time-course (mean ± SEM) of ASA (left column) and COX (right column) after ingestion of aspirin 650 mg of compressed tablets (upper panels) and enteric-coated tablets (bottom panels) in six healthy volunteers. Figure from [126].*

Similar results regarding the kinetics of different formulation of aspirin were obtained by Sai and others [132], who enrolled twelve healthy volunteers to receive four separated 100 mg oral aspirin administration: intact and chewed non-coated tablets, and intact and chewed enteric-coated tablets. Figure shows the four resulting time-courses of plasma ASA concentration. ASA was detectable in serum within 20 minutes after the ingestion of intact non-coated aspirin, although the authors pointed out how significant variability was observed. After ingestion of intact coated aspirin, ASA was not detectable in serum until 4 hours, while, when coated aspirin was chewed, ASA was detectable within 20 minutes after ingestion. Moreover, levels of ecASA were significantly lower than the others (even if the peak of ecASA concentration could not be determined since there were not measurements after 8 hours and ecASA appeared to keep on rising after 8 hours). From the experiment, the authors concluded that enteric-coated formulation results in a slower kinetics of aspirin and emphasized that a significant interindividual variability was observed [132].

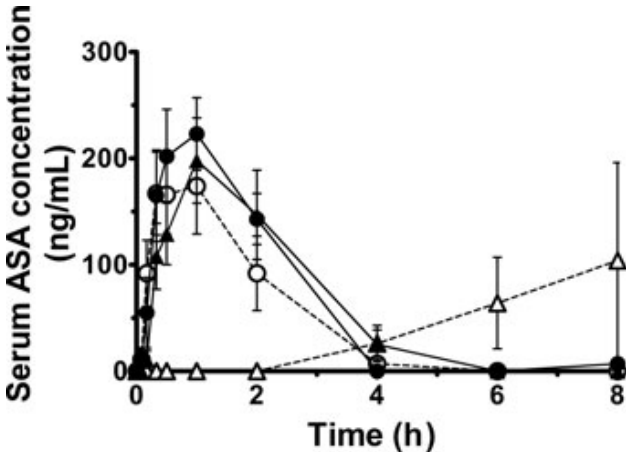


Figure 4.25: *Serum ASA concentration after ingestion of intact (open) and chewed (closed) non-coated (circle) and coated (triangle) tablets was measured for 8 h. Each point represents the mean ± SEM of 10–12 patients. Figure from [132]*.

This wide variability was confirmed by many other studies, and extreme results were obtained by Ross-Lee and others [129], who studied fourteen healthy volunteers randomized to receive 1200 mg of ASA or 1300 mg of ecASA. For all the volunteers administered with ecASA, ASA levels in serum were below the detection limit of the assay (< 0.5 mg/l) at all times of measurements. Only repeating the experiment with 650 mg and measuring ASA levels with a more sensitive technique, peak concentrations were observed (mean 0.24 mg/l ± 11) 4 hours after dosing (mean 0.24 mg/l ± 11) in 3 subjects, and at 6 hours in one subject.

This brief overview on the current literature knowledge on aspirin PK, makes it rather clear that:

- aspirin PK strongly depends on tablet formulation: enteric-coating results in a delayed and slower kinetics with respect to the non-coated preparation;
- interindividual variability plays a major role in the appearance of ASA in serum.

An *in silico* model, which aims to investigate the adequacy of different aspirin regimens, necessarily needs to mathematically model not only COX but also aspirin kinetics. Moreover, as it's clear from section 4.4.3, not only ASA time-course in blood is needed but in bone marrow too, and, due to limited access to bone marrow megakaryocytes, only a model allows to simulate this kinetics without invasive and expensive test.

Thus, a compartmental model of ASA PK was developed (Figure 4.26):

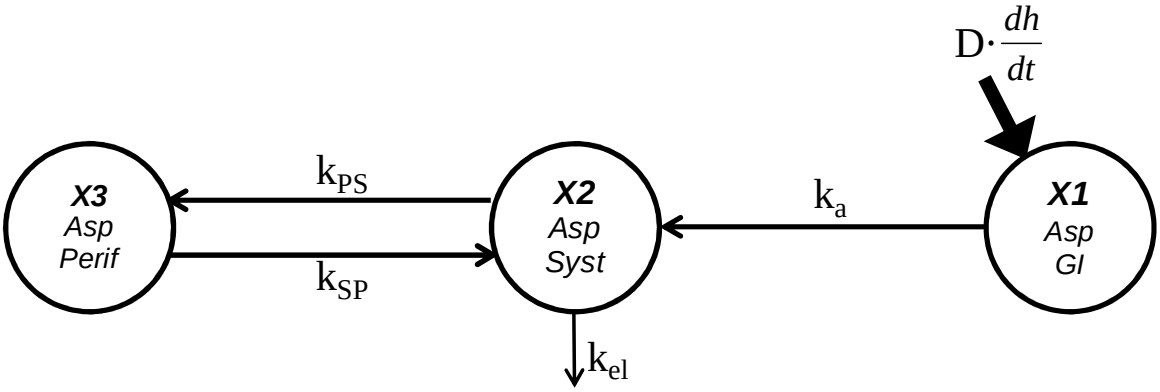


Figure 4.26: *The compartmental model for ASA kinetics.*

The model is described by the following linear differential equations:

$$\begin{cases} \dot{X_1} = D \cdot \dfrac{d\big(f(t)\big)}{dt} - k_a \cdot X_1, & X_{10} = 0 \quad (4.53) \\ \dot{X_2} = k_a \cdot X_1 - \big(k_a + k_{pc}\big) \cdot X_2 + k_{cp} \cdot X_3, & X_{20} = 0 \quad (4.54) \\ \dot{X_3} = k_{pc} \cdot X_2 - k_{cp} \cdot X_3, & X_{30} = 0 \quad (4.55) \end{cases}$$

where:

- X_1, X_2 and X_3 represent ASA amounts in gastro-intestinal tract, systemic compartment (i.e. systemic circulation) and bone marrow compartment (peripheral compartment), respectively. The unit of measure for X_1, X_2 and X_3 is [*mass*];
- k_a represents the rate of appearance of ASA from gastro-intestinal compartment to systemic compartment; k_{el} represents the rate ASA elimination from systemic compartment and ASA distribution to the whole body; k_{ps} and k_{sp} represent the rates of exchange from systemic to peripheral compartments and vice versa, respectively. The unit of measure for k_a, k_{el}, k_{ps} and k_{sp} is [$time^{-1}$];
- the input of the model is given by the product of the oral dose D (unit of measure: [*mass*]) and the derivative of a suitable function $h(t)$, used to simulated different kinetics of ASA, depending on the formulation. $h(t)$ is given by the following equation of Hill:

$$h(t) = \frac{t^m}{K + t^m} \quad (4.56)$$

where K is given by:

$$K = \frac{m+1}{m-1} \cdot t\_flex^{m} \tag{4.57}$$

and *t_flex* is the time in which *h(t)* has its point of inflection (Figure 4.27).

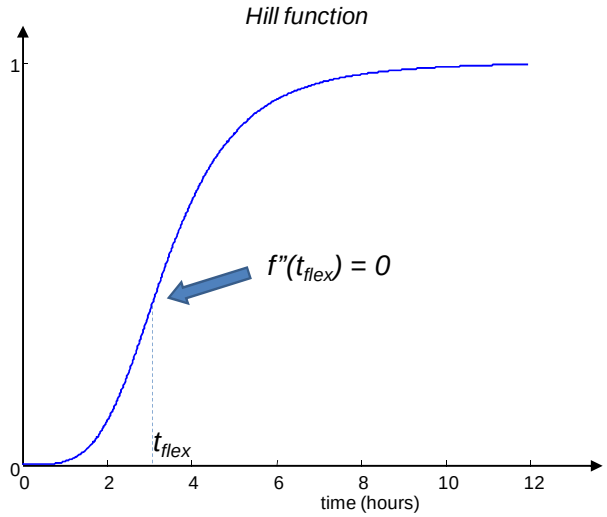


Figure 4.27: *Example of h(t) used to simulate different kinetics of release of ASA.*

h(t) represents the drug availability, thus:

$$\lim_{t\to\infty} h(t) = 1 \tag{4.58}$$

meaning that all the initial dose is released in the systemic compartment. The derivative of *h(t)* is given by:

$$\frac{dh(t)}{dt} = \frac{m \cdot t^{m-1} \cdot K}{(K + t^{m})^{2}} \tag{4.59}$$

and it simulates the release of the oral dose *D*.

Figure 4.28 shows the effect on *h(t)* (left column) and on the consequent flux of appearance in the systemic compartment $k_a \cdot X_1(t)$ (right column) caused by a variation of the parameter *t_flex* (upper row) and *m* (bottom row), following an ingestion of 100 mg aspirin (the parameter k_a was set to a nominal value of 0.1 min^{-1} in the example). As one can see, *t_flex* is responsible for the delay in the release of the drug and for the speed of the kinetics (the grater *t_flex* the grater the delay and the slower

the kinetics), while *m* mainly controls the speed of the release (the grater *m*, the faster the kinetics). Both the parameters have a direct effect on the peak of the flux of appearance, since the faster and earlier is the release, the grater is the peak.

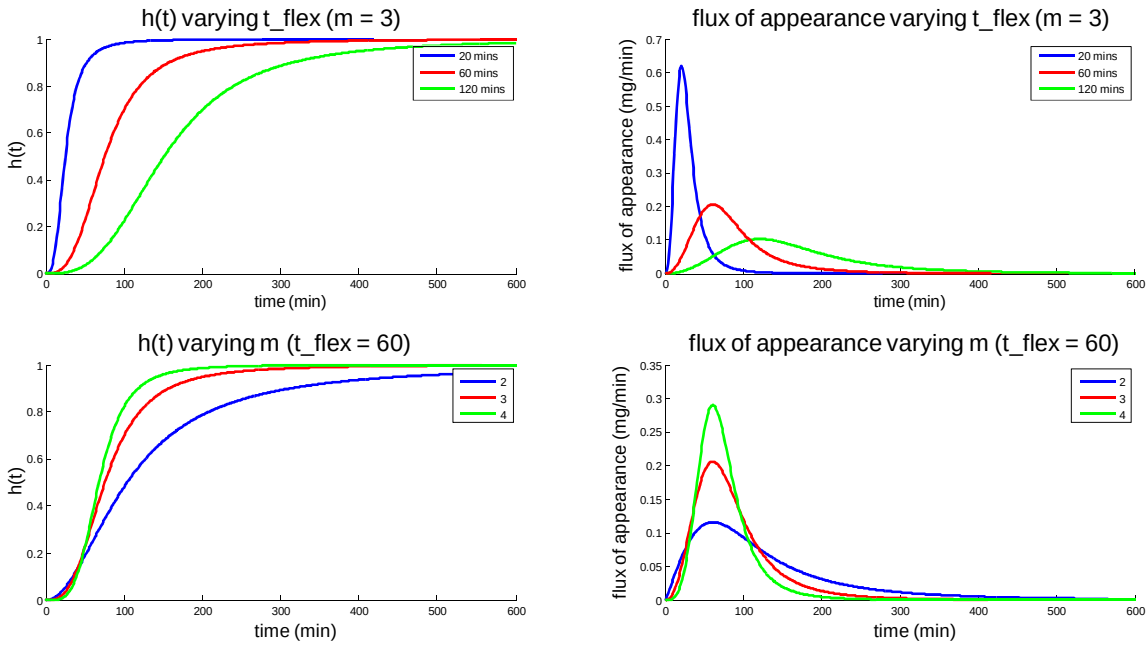


Figure 4.28: *Variation of h(t) (left panels) and $k_a \cdot X_1(t)$ (right panels), for different values of t_flex (upper panels) and m (bottom panels), following a dose of 100 mg ASA. Values used for t_flex are 20, 60 and 120 minutes, values used for m are 2, 3 and 4. k_a is set to 0.1 min^{-1}.*

While COX is physically contained into the cells and its kinetics is dependent on the different steps of the thrombopoiesis process, aspirin is free to move through cell membranes by diffusion, not depending on the processes of cell maturation. This is why aspirin kinetics is described by a traditional lumped model, i.e. the only independent variable is time.

Given mutual interaction between aspirin and COX (see ASA PD in section 4.4.3), the model of aspirin PK cannot be modeled separately from the model of COX kinetics, thus a unique aggregated model is needed. In the following section the final model will be then presented.

4.4.5 *Final model*

The connection of the 2 sub-models for COX kinetics (section 4.4.2) and for aspirin kinetics (section 4.4.4), combined by aspirin pharmacodynamics described in section 4.4.3, results in the final model of Figure 4.29, described by equations (4.60)-(4.75).

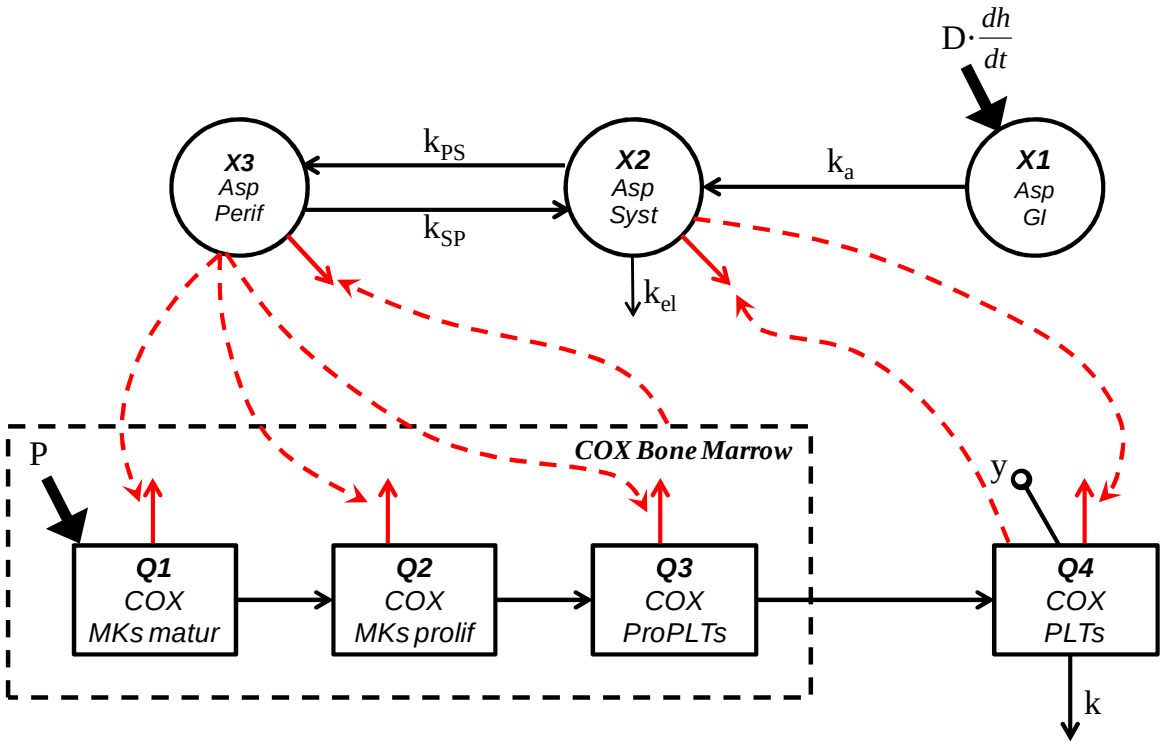


Figure 4.29: *final model, partially lumped and partially distributed, of aspirin action. Red dashed lines represents interaction between aspirin and COX.*

In summary, the input is given by the oral dose of aspirin multiplied by the derivative of the hill function *h(t)* used to simulated enteric-coated formulation. Aspirin enters the gastro-intestinal compartment X_1 and then appears, with a constant rate *ka*, in the systemic compartment. Aspirin in the systemic compartment is partially eliminated and utilized by all the other tissues, with a constant rate k_{el}, and partially transferred to the bone marrow compartment X_3 with a constant rate k_{ps}. Aspirin in bone marrow can move back to the systemic compartment with a constant rate k_{sp}. Aspirin in the systemic compartment acts on COX in the compartment of circulating platelets Q_4, while aspirin in bone marrow acts on COX in bone marrow, i.e. on the compartments of maturing and proliferating megakaryocytes (Q_1 and Q_2 respectively) and on the compartment of proplatelets Q_3. The measure is represented by COX in circulating platelets.

$$
\begin{cases}
\dot{X_1} = D - k_a \cdot X_1, & X_{10} = 0 \quad (4.60)\\
\dot{X_2} = k_a \cdot X_1 - \left[k_{el} + k_{ps} + \lambda \cdot \dfrac{Q_4}{V_{Q4}}\right] \cdot X_2 + k_{sp} \cdot X_3, & X_{20} = 0 \quad (4.61)\\
\dot{X_3} = k_{ps} \cdot X_2 - \left[k_{ps} + \lambda \cdot \dfrac{Q_1 + Q_2 + Q_3}{V_{QBM}}\right] \cdot X_3, & X_{30} = 0 \quad (4.62)\\
\dfrac{\partial Q_1(s,t)}{\partial s} = P(s), & Q_1(0,t) = 0 \quad (4.63)\\
\dfrac{\partial Q_1(s,t)}{\partial t} = -\lambda \cdot \dfrac{X_3}{V_{X3}} \cdot Q_1(s,t), & Q_1(s,0) = Q_{10} \quad (4.64)\\
COX_1(t) = \displaystyle\int_0^{MK\_matur} Q_1(s,t)ds, & COX_1(0) = COX_{10} \quad (4.65)\\
\dfrac{\partial Q_2(s,t)}{\partial s} = f[Q_2(s,t), s], & Q_2(0,t) = f_2[Q_1(s,t)] \quad (4.66)\\
\dfrac{\partial Q_2(s,t)}{\partial t} = -\lambda \cdot \dfrac{X_3}{V_{X3}} \cdot Q_2(s,t), & Q_2(s,0) = Q_{20} \quad (4.67)\\
COX_2(t) = \displaystyle\int_0^{MK\_prolif} Q_2(s,t)ds, & COX_2(0) = COX_{20} \quad (4.68)\\
\dfrac{\partial Q_3(s,t)}{\partial s} = 0, & Q_3(0,t) = f_3[Q_2(s,t)] \quad (4.69)\\
\dfrac{\partial Q_3(s,t)}{\partial t} = -\lambda \cdot \dfrac{X_3}{V_{X3}} \cdot Q_3(s,t), & Q_3(s,0) = Q_{30} \quad (4.70)\\
COX_3(t) = \displaystyle\int_0^{ProPLT\_life} Q_3(s,t)ds, & COX_3(0) = COX_{30} \quad (4.71)\\
\dfrac{\partial Q_4(s,t)}{\partial s} = -k \cdot Q_4(s,t), & Q_4(0,t) = f_4[Q_3(s,t)] \quad (4.72)\\
\dfrac{\partial Q_4(s,t)}{\partial t} = -\lambda \cdot \dfrac{X_2}{V_{X2}} \cdot Q_4(s,t), & Q_4(s,0) = Q_{40} \quad (4.73)\\
COX_4(t) = \displaystyle\int_0^{PLT\_life} Q_4(s,t)ds, & COX_4(0) = COX_{40} \quad (4.74)
\end{cases}
$$

$$y = COX_4(t) \quad (4.75)$$

Since modeling aspirin PD by the mass-action law needs the concentrations of the reagents, the volumes of the different compartments have now to be accounted for.
The volumes appearing now in the equations are: V_{QBM}, V_{Q4}, V_{X2} and V_{X3}, representing the total volume of megakaryocytes and proplatelets together, the total volume of circulating platelets, the total volume of the systemic circulation compartment and the total volume of bone marrow, respectively. Indicative values for V_{X2} and V_{X3} were taken directly from [108]. Value for V_{QBM} was derived combining information on megakaryocyte dimensions, from [114], and on megakaryocytes numerosity, from [106]. Value for V_{Q4} was derived from [109], as product between the mean number of platelets per liter of blood and the mean platelet volume.

4.4.6 Model Parameters

Given the final model of Figure 4.29, described by equations (4.60)-(4.75), the complete list of parameters of the model, together with the respective literature ranges or nominal values, is reported in Table 4.2.

The total number of parameters is 20, of which:

- 13 parameters (*N_MKs_per_Kg*, *MK_matur*, *MK_prolif*, *ProPLT_life*, *PLT_life*, *N_ProPLTs_per_MK*, *N_PLTs_per_MK*, k, V_{QBM}, V_{Q4}, ka, V_{X2}, V_{X3}) can be considered known directly or derived from the literature. It is important to make clear that nominal values of the parameters are to be meant for healthy subjects.
- 7 parameters (p_{MK}, *t_flex*, m, k_{ps}, k_{sp}, k_{el} and λ) are unknown.

An a-priori identifiability analysis was performed, using the DAISY (Differential Algebra for Identifiability of SYstems) software by Bellu and others [95], which implements a differential algebra algorithm to perform parameter identifiability analysis for linear and nonlinear dynamic models described by polynomial or rational equations. The model turned out to be neither globally nor locally identifiable.
However, since the model has an explorative aim and its main objective is to qualitatively describe experimental evidence, unknown parameter estimation by fitting real data did not appear to be the more suitable strategy, especially considering that: *i)* only one output was available (i.e. measurements in one compartment only) *ii)* data are characterized by a high variability, which inevitably leads to inaccurate parameters estimates.
Thus, the model was not simplified and no data fitting was performed.

Unknown parameters setting is described in the following section.

	Parameter	Description	Literature value
THROMBOPOIESIS	*N_MKs_per_Kg*	number of megakaryocytes per kilogram of subject weight	~ $6.1e^{6}$ [106]
	MK_matur	maturation period of megakaryocytes	~ 2-4 days [117]
	MK_prolif	proliferation period of megakaryocytes	~ 4-10 hours [106]-[116]
	ProPLT_life	lifespan of proplatelets	~ 12-24 hours [117]
	PLT_life	lifespan of platelets	~ 7-10 days [135]
	N_ProPLTs_per_MK	number of proplatelets generated from each megakaryocyte	~ 1000 ÷ 1500 [116]
	N_PLTs_per_ProPLT	number of platelets generated from each proplatelet	2 [116]
COX KINETICS	p_{MK}	COX new synthesis within the single maturing megakaryocyte	?
	k	COX degradation rate in platelets	~ $4.8e^{-4}$ min^{-1} [103]
	V_{QBM}	volume of all megakaryocytes and all proplatelets together	~ 1000 ml [114]
	V_{Q4}	volume of all circulating platelets	~ 15 ml [109]
ASA PK	*t_flex*	time of inflection of the hill function	?
	m	degree of the hill function	?
	k_a	rate of ASA appearance from gastro-intestinal tract	~ 0.1875 min^{-1} [123]
	k_{el}	rate of ASA elimination and distribution to all tissues from systemic circulation	?
	k_{ps}	rate of ASA exchange from systemic circulation to bone marrow	?
	k_{sp}	rate of ASA exchange from bone marrow to systemic circulation	?
	V_{X2}	volume of systemic circulation	~ 5600 ml [108]
	V_{X3}	volume of bone marrow tissue	~ 1177 ml [108]
ASA PD	λ	rate constant of the reaction between ASA and COX	?

Table 4.2: *Parameters of the final model, with literature ranges and nominal values.*

4.5 *Parameters setting*

With the objective of setting unknown parameters to reasonable values, a set of simulations was performed, where, for each parameter, different values from a search interval have been tested, fixing all the others to nominal values. The known parameters were set to the mean value of the respective range or to the nominal value reported in Table 4.2. For each unknown parameter, Table 4.3 reports the nominal value used and the search interval.

Parameter	nominal value used in simulations	Search interval
p_{MK}	$1e^{-15}$ g/min	$1e^{-18} \div 1e^{-11}$
λ	$2e^{3}$ $mol^{-1}min^{-1}$	$10^{2} \div 10^{5}$
t_flex	180 mins	20 ÷ 480
m	4	2 ÷ 8
k_{el}	0.2 min^{-1}	0.01 ÷ 1
k_{ps}	0.01 min^{-1}	0.001 ÷ 1
k_{sp}	0.01 min^{-1}	0.001 ÷ 1

Table 4.3: *Nominal values and search intervals for the unknown parameters.*

A sensitivity analysis was carried out by computing the sensitivity of two main output variables to each unknown parameter, performing a simulation of one week therapy 100 mg ecASA once a day.

The two output variables are:

- *lag-time*: the delay in the recovery of platelets COX, formally defined as the time required to reach 10% of steady-state.
- *rise-time*: the time required for platelets COX to go from 10% to 90% of its steady-state level.

The sensitivity of the output variable *out* to the parameter *p* was computed as:

$$S(p) = \frac{dout(p)}{dp} \cdot \frac{p}{out(p)} \tag{4.76}$$

In the following, the variation of each single parameter is discussed and, in section 4.5.1.8, the mean sensitivities are summarized in Table 4.4.

4.5.1.1 COX production within the single MK: p_{MK}

There's no direct information on p_{MK} in the literature. However, since TxB_2 is supposed to be proportional to COX activity [133], an indicative value for p_{MK} can be computed with the hypothesis of a proportion 1:1 between TXb_2 and COX. Using the expression for platelets COX at steady-state (equation (4.39)) and the baseline value for serum TxB_2 in healthy subjects (retrieved from [133]), a value of 1.5^{e-15} g/min cab be computed for p_{MK}. Rather than the absolute value of p_{MK}, the amplification effect on pMK has been investigated: since the proportion 1:1 between TxB_2 and COX is not confirmed in the literature, the real production within the single MK is supposed to be *N* times p_{MK}, and the effect of a variation of *N* has been studied. In particular, an increase of *N* results in an increase of COX levels in each compartment (equations (4.18), (4.24), (4.35) and (4.39)). Considering equations (4.45) and (4.46), which describe the interaction between COX and ASA, one can see that, if COX increases by *N* times, COX kinetics (equation (4.45)) does not change, while ASA kinetics (equation (4.46)) changes and the effect of COX on ASA is amplified by *N* times. Thus, ASA is consumed much faster if COX increases, and this indirectly affects COX too, since if ASA decreases very fast, the effect on COX is lower. This is confirmed by simulations, in particular, by simulating a single 100 mg aspirin intake, one can see that, as *N* increases, ASA concentration peak in serum decreases (Figure 4.30.A) and, consequently, COX maximal acetylation in PLTs decreases too (Figure 4.30.B).

Figure 4.31 shows the output variables *lag-time* (panel A) and *rise-time* (panel B) as functions of *N*, for a simulation of one week therapy 100 mg ecASA once a day.

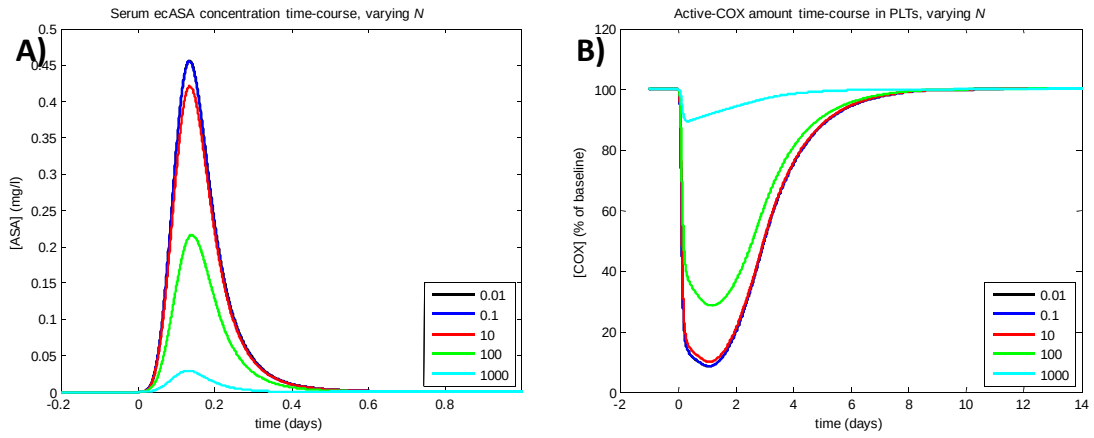


Figure 4.30: *ASA concentration in serum (A) and COX time-course in PLTs (B), increasing the parameter N (0.01, 0.1, 10, 100, 1000). Single aspirin intake at t = 0.*

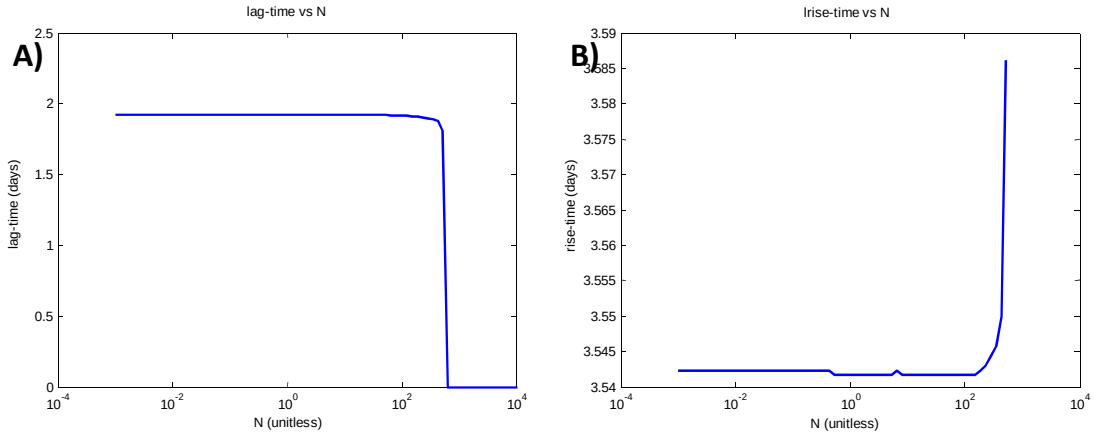


Figure 4.31: *lag-time (A) and rise-time (B) as function of the parameter N.*

The output lag-time is more influenced than the rise-time by a variation of the parameter *N*. It is interest to note how values greater than ~ 10^2 result in lag-time equal to zero and an absence of rise-time. This is because the maximal acetylation reached is too small, as one can see in Figure 4.30.B. Values smaller than ~ 10^2 seem to result in quite constant values of both lag-time and rise-time. In particular, the lag-time has a value of about 2 days for *N* values between 10^{-3} and 10^2. A good choice for *N* is likely to be within this interval.

4.5.1.2 Reaction constant: λ

Considering again equations (4.45) and (4.46), which describe the interaction between COX and ASA, one can see that a variation on the value of λ affects both COX and ASA kinetics. In particular, the grater is λ, the greater is the mutual effect between COX and ASA, thus the maximal acetylation of PLTs COX is grater and the acetylation reaction is faster (Figure 4.32.A shows this behavior for a simulation of a single intake of 100 mg aspirin). This is particular relevant in the case of repeated doses: Figure 4.32.B reports the results of a simulation of one week 100 mg ecASA once a day; one can see that, a small value of λ allows to obtain an effect of the duration of the therapy, i.e. we need more than one intake to achieve the maximal effect. However, a small value for λ does not result in a complete acetylation of COX in PLTs, thus not reflecting reality.

To observe a sensible effect of λ on ASA concentrations, COX levels have to be increased (by increasing p_{MK} as explain in the previous section), otherwise no effect of λ can be detected (Figure 4.33.A). This is because COX concentration is much smaller than ASA concentration, thus, to observe an effect on ASA, COX concentration needs to be grater (equation (4.46)). Amplifying p_{MK} for example by 1000 times, an effect of λ on ASA

concentration can be observed (Figure 4.33.B): as λ increases, ASA peak decreases, meaning that the effect of COX on ASA in greater.

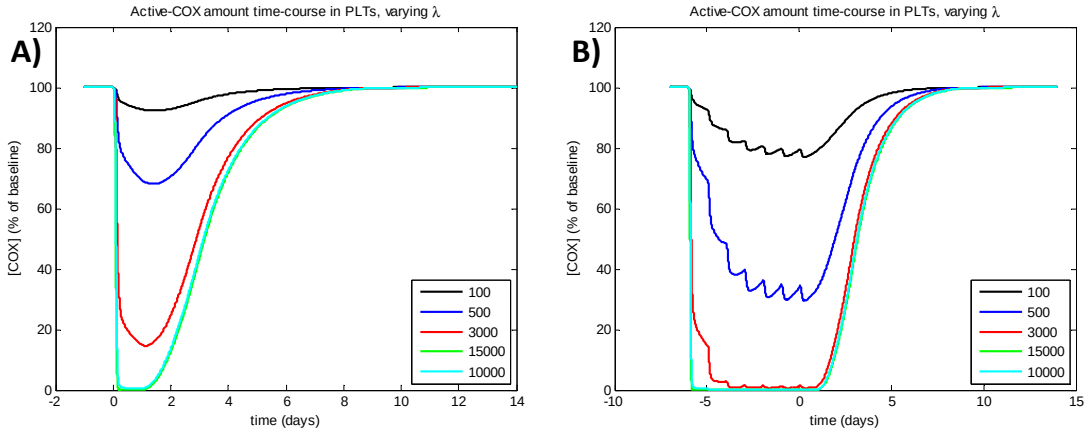


Figure 4.32: *COX time-course in PLTs after a single dose (A) and after one week treatment with 24 hours intakes (B), increasing the parameter λ (100, 500, 3000, 15000, 10000). Aspirin last intake at t = 0.*

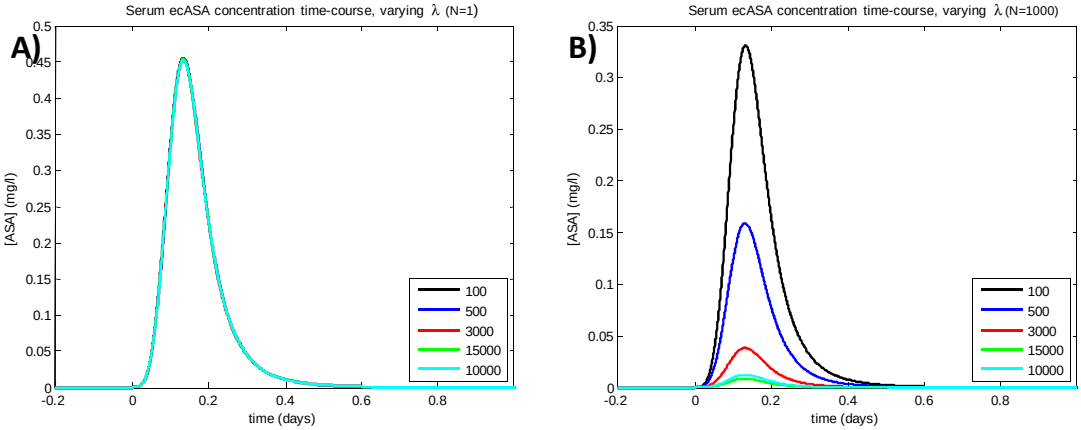


Figure 4.33: *ASA concentration in serum varying λ (100, 500, 3000, 15000 and 10000) with no amplification of COX production (N=1, panel A) and with a one thousands amplification of COX production (N = 1000, panel B). Single aspirin intake at t = 0.*

Figure 4.34 shows the output variables *lag-time* (panel A) and *rise-time* (panel B) as functions of λ, for a simulation of one week therapy 100 mg ecASA once a day.

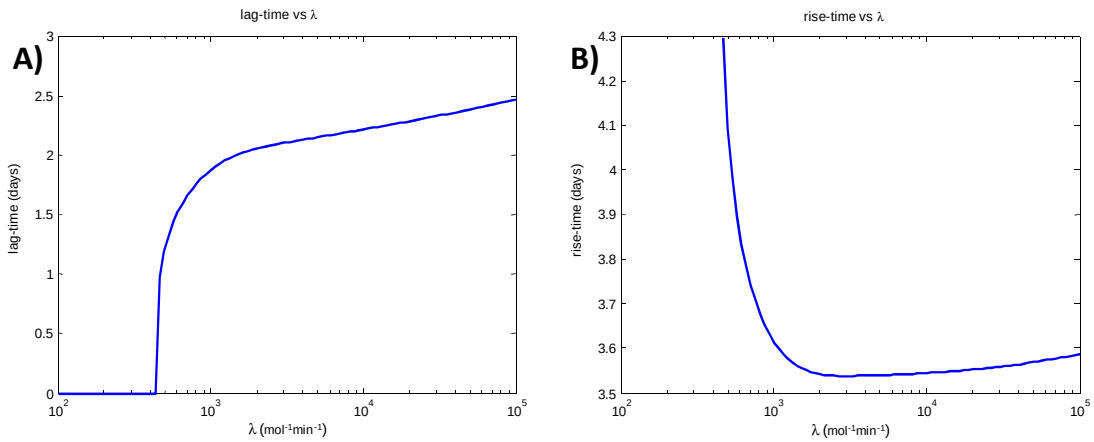


Figure 4.34: *lag-time (A) and rise-time (B) as function of the parameter λ.*

It is interesting to note how small values of λ result in lag-time equal to zero, meaning that there is not enough acetylation of COX in PLTs (see in Figure 4.32). As a consequence, no rise-time can be computed for small λ values. For greater values, one can see that, as λ increases, the lag-time increases, while the rise-time decreases, becoming quite constant for λ values greater than ~ 10^3, even if the rise-time appears to exhibits a minimum for λ values between 10^3 and 10^4. A lag-time of about 2 days is obtained for λ values between 10^2 and 10^4, thus a good choice for N is likely to be within this interval.

4.5.1.3 Time of inflection of h(t): t_flex

As already described in section 4.4.4, *t_flex* is the parameter representing the time at which the hill function, used o simulate the enteric-coated formulation of aspirin, has its point of inflection. The parameter is responsible for the delay in the release of the drug and for the speed of ASA kinetics. Figure 4.35 shows results of a simulation of a single 100 mg aspirin intake: as one can see, the grater *t_flex* the grater the delay and the slower the kinetics of ASA (panel A). COX kinetics is affected in the same way even if the overall effect is not so strong (panel B).

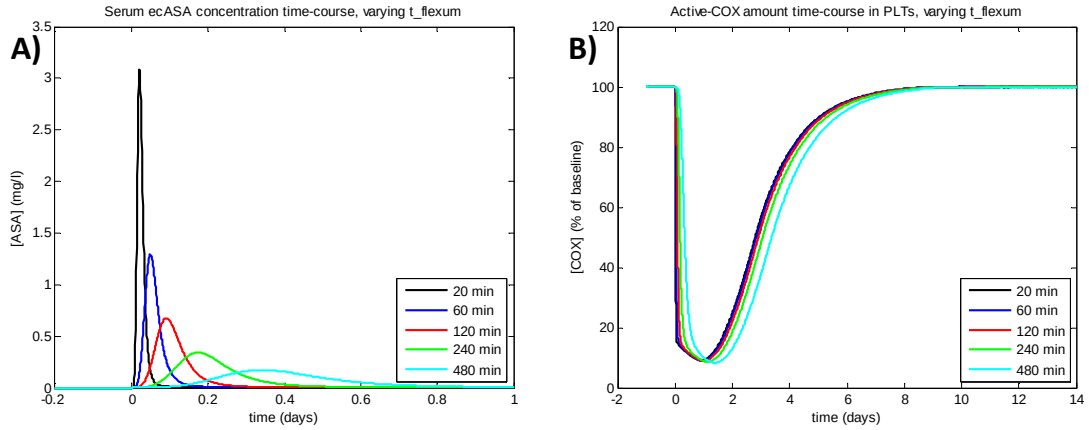


Figure 4.35: *ASA concentration in serum (A) and COX time-course in PLTs (B), increasing the parameter t_flex (20, 60, 120, 240 and 480 minutes). Single aspirin intake at t = 0.*

Figure 4.36 shows the output variables *lag-time* (panel A) and *rise-time* (panel B) as functions of *t_flex*, for a simulation of one week therapy 100 mg ecASA once a day.

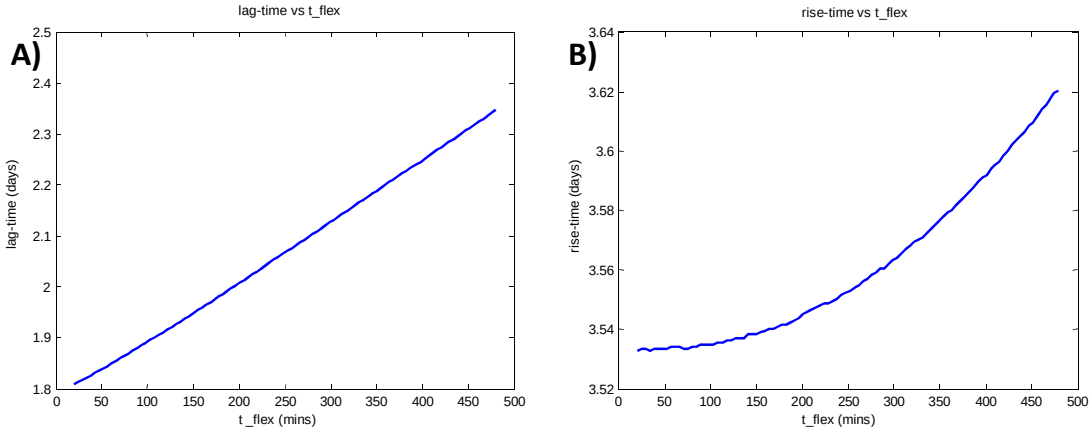


Figure 4.36: *lag-time (A) and rise-time (B) as function of the parameter t_flex.*

The parameter *t_flex* mainly affects the lag-time: as it increases, the lag-time increases too. *t_flex* exhibits the same effect on the rise-time, even if much weaker. In particular, a lag-time of about 2 days is obtained for *t_flex* values around 200 minutes, thus a good choice for *t_flex* is likely to be near this value.

4.5.1.4 Degree of h(t): m

As already described in section 4.4.4, *m* is the degree of the hill function *h(t)* and it mainly controls the speed of ASA kinetics. Figure 4.37 shows results of a simulation of a single 100 mg aspirin intake: as one can see, the grater *m*, the faster the kinetics and the higher the peak of ASA (panel A). The kinetics of COX recovery is affected too, but with definitely smaller effect (panel B).

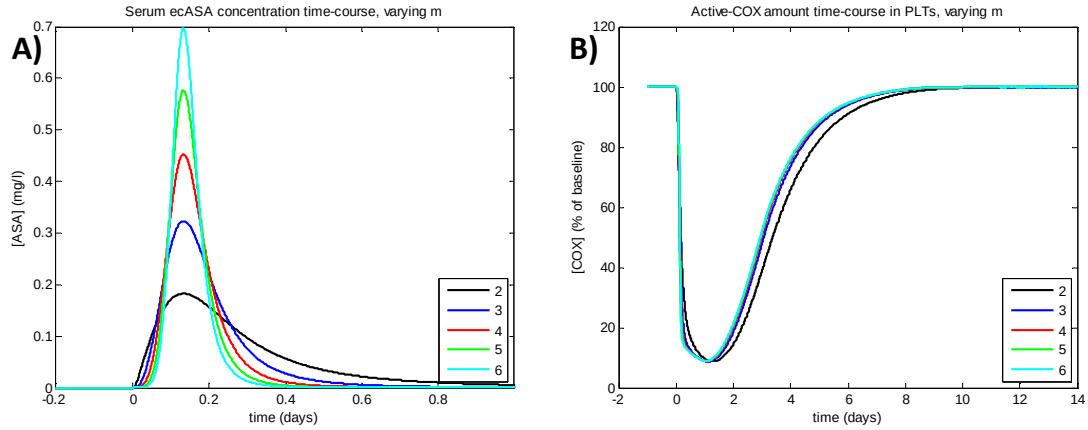


Figure 4.37: *ASA concentration in serum (A) and COX time-course in PLTs (B), increasing the parameter m (2, 3, 4, 5 and 6). Aspirin intake at t = 0.*

Figure 4.38 shows the output variables *lag-time* (panel A) and *rise-time* (panel B) as functions of *m*, for a simulation of one week therapy 100 mg ecASA once a day.

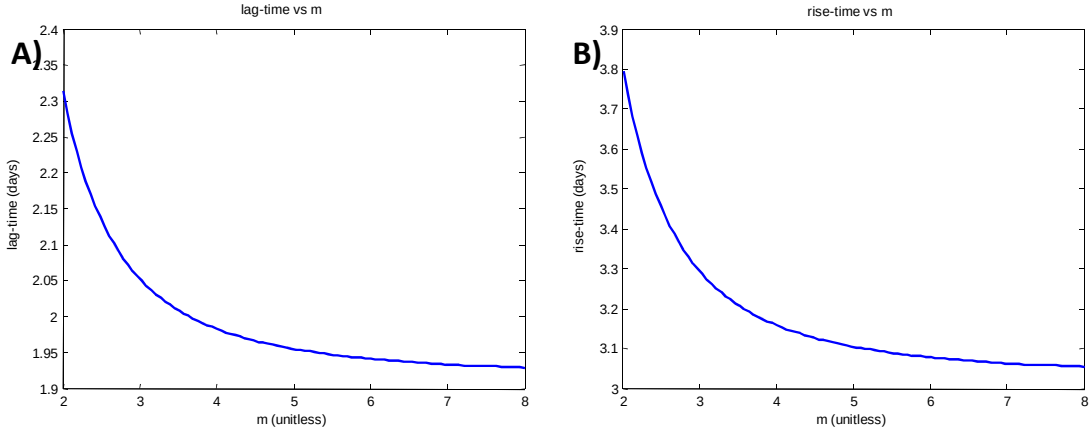


Figure 4.38: *lag-time (A) and rise-time (B) as function of the parameter m.*

Both the lag-time and the rise-time are decreasing function of *m*, even if the effect of this parameter is not so great. $m \sim 4$ appears a good choice to obtain a lag-time of about 2 days.

4.5.1.5 Elimination rate from the systemic compartment: k_{el}

k_{el} represents the overall ASA elimination from the central compartment, sum of several mechanisms: pre-systemic uptake from the liver, elimination with urine, and utilization by other tissues. All these mechanisms have been modeled as a single flux, which is supposed to be proportional to ASA concentration in the systemic compartment, via the rate coefficient k_{el}. Thus, it is clear how k_{el} plays a major role in the kinetics of ASA in the systemic compartment, and, as a consequence, on all the other compartments. Figure

4.39 shows results of a simulation of a single 100 mg aspirin intake: as k_{el} increases, ASA elimination from the central compartment is faster, thus the peak of ASA concentration decreases and the kinetics if faster (panel A); consequently, COX acetylation decreases (panel B).

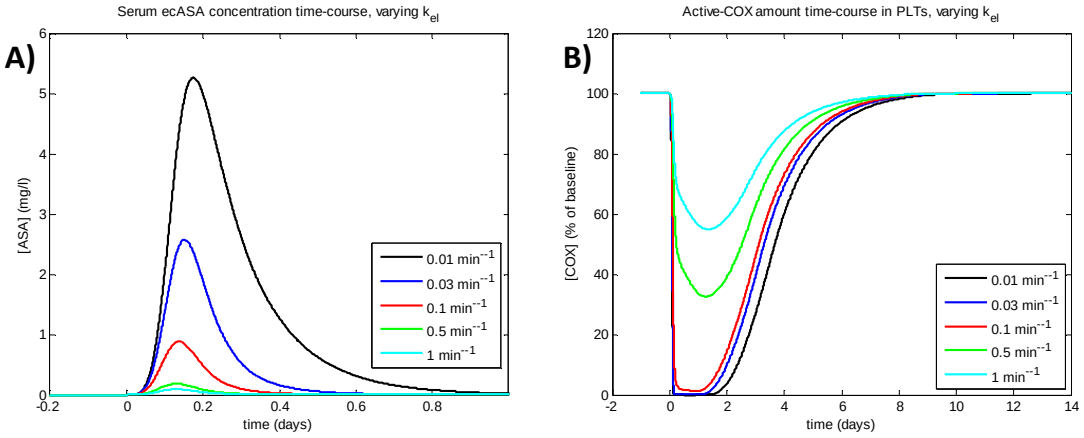


Figure 4.39: *ASA concentration in serum (A) and COX time-course in PLTs (B), increasing the parameter k_{el} (0.001, 0.01, 0.1, 0.5 and 1 min^{-1}). Aspirin intake at t = 0.*

Figure 4.40 shows the output variables *lag-time* (panel A) and *rise-time* (panel B) as functions of k_{el}, for a simulation of one week therapy 100 mg ecASA once a day.

As one can see, as k_{el} increases the lag-time decreases, since ASA elimination is greater. For k_{el} near to 1, the lag-time becomes almost null, meaning that ASA elimination is so great that COX acetylation is too small. The rise-time is quite constant for small values of k_{el}, (even if there seems to be a minimum for k_{el} values around 10^{-1}) then it increases as k_{el} increases. For k_{el} near to 1, no rise-time can be computed, since COX acetylation is too small. Since k_{el} models several mechanism, including uptake from the liver which is known to be relevant [123], reasonable values could be in the range of 0.1÷0.5 min^{-1}. This appears to be confirmed by Figure 4.40.A, since a correct lag-time of about 2 days is obtained for k_{el} values near to 10^{-1}.

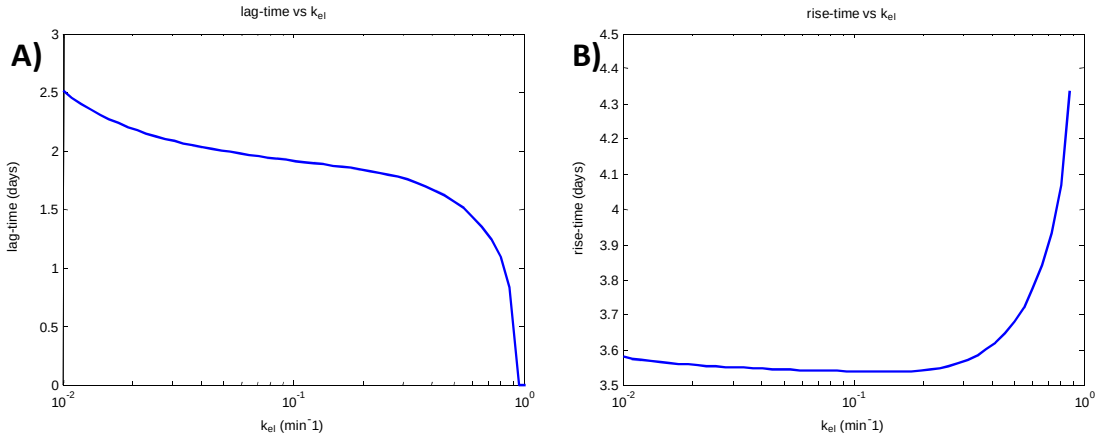


Figure 4.40: *lag-time (A) and rise-time (B) as function of the parameter m.*

4.5.1.6 Transfer rate from systemic to peripheral compartment: k_{ps}

This parameters represents the flux of ASA from the central compartment to the bone marrow compartment. Figure 4.41 to Figure 4.43 show results of a simulation of a single 100 mg aspirin intake: as k_{ps} increases, a greater amount of ASA is transferred from the central compartment to the peripheral one, thus, the peak of ASA concentration in the systemic compartment decreases (Figure 4.41.A) and the one in the peripheral compartment increases (Figure 4.41.B). This results in a greater acetylation of COX in MKs, both maturing and proliferating (Figure 4.42) and ProPLTs (Figure 4.43.A) which produces a more delayed recovery of COX in PLTs, as shown in Figure 4.43.B.

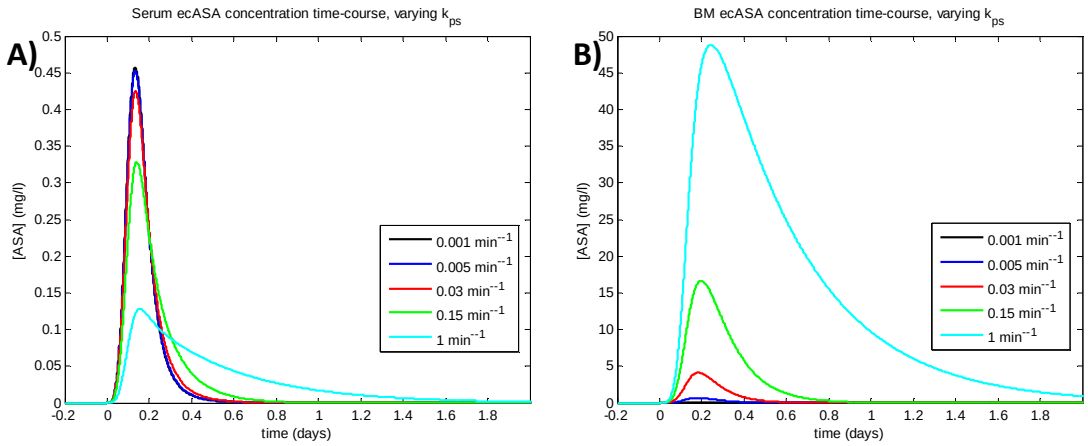


Figure 4.41: *ASA concentration in serum (A) and in bone marrow (B), increasing the parameter k_{ps} (0.001, 0.005, 0.03, 0.15 and 1 min^{-1}). Single aspirin intake at t = 0.*

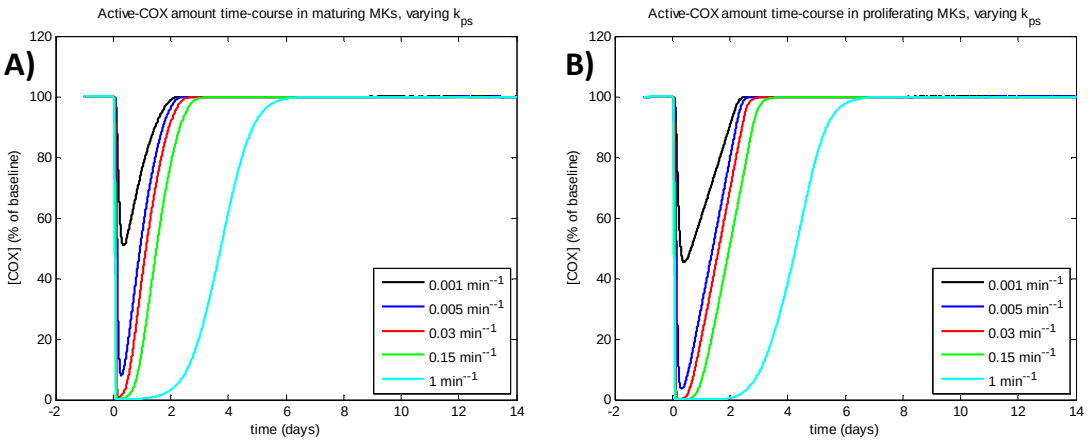


Figure 4.42: *COX time-course in maturing (A) and proliferating (B) MKs, increasing the parameter k_{ps} (0.001, 0.005, 0.03, 0.15 and 1 min^{-1}). Single aspirin intake at t = 0.*

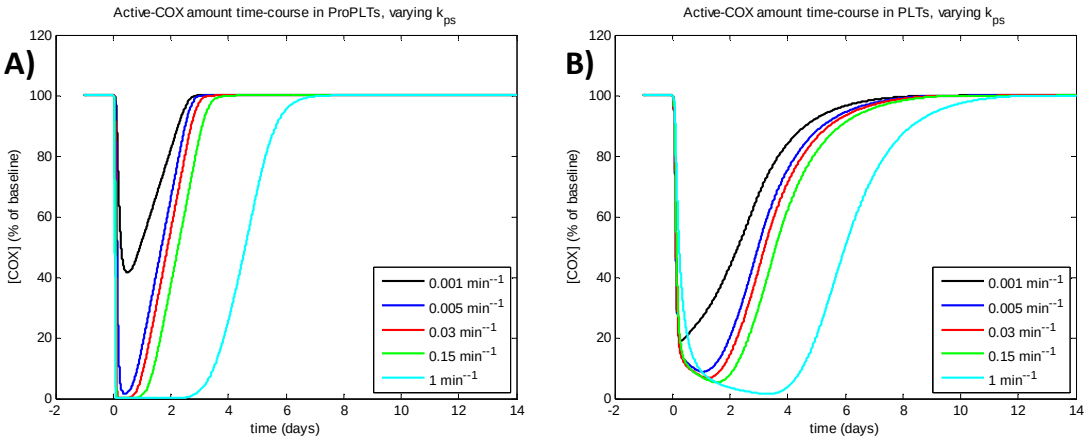


Figure 4.43: *COX time-course in ProPLTs (A) and PLTs (B), increasing the parameter k_{ps} (0.001, 0.005, 0.03, 0.15 and 1 min^{-1}). Single aspirin intake at t = 0*

Figure 4.44 shows the output variables *lag-time* (panel A) and *rise-time* (panel B) as functions of k_{ps}, for a simulation of one week therapy 100 mg ecASA once a day.

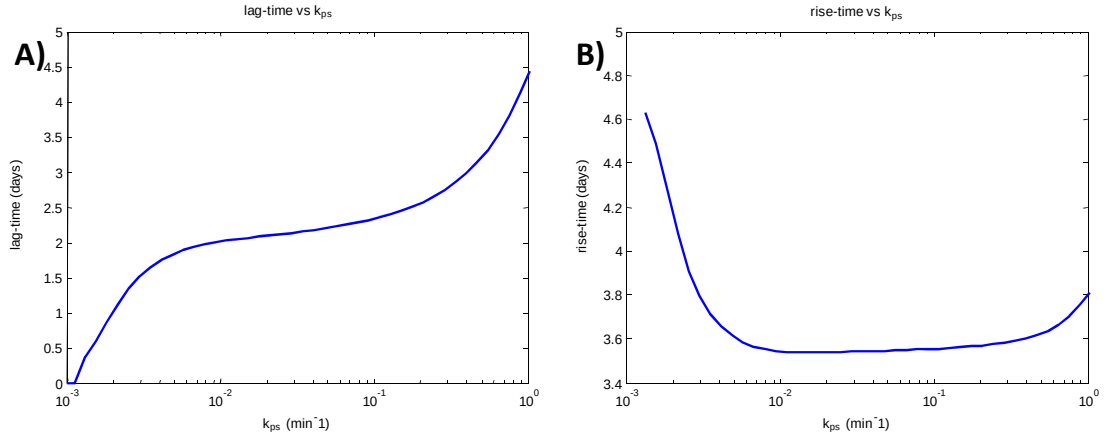


Figure 4.44: *lag-time (A) and rise-time (B) as function of the parameter k_{ps}.*

As one can see, small values of k_{ps} result in a small lag-time, while great values of k_{ps} result in a too high lag-time. This is because great values of k_{ps} cause ASA to accumulate in bone marrow, requiring several days before a complete elimination. This causes a slower COX recovery in PLTs. The rise-time seems to exhibits a minimum for a k_{ps} values of about 10^{-2}. A lag-time of about 2 days is obtained for k_{ps} values between 10^{-2} and 10^{-1}, thus a good choice for k_{ps} is likely to be within this interval.

4.5.1.7 Transfer rate from systemic to peripheral compartment: k_{sp}

This parameters represents the flux of ASA from the bone marrow compartment to the systemic one. Figure 4.45 to Figure 4.47 show results of a simulation of a single 100 mg aspirin intake: as k_{sp} decreases, the kinetics of ASA in bone marrow is slower and the peak increases, as shown in Figure 4.45.B. ASA in systemic compartment is not sensitively affected by k_{sp} (Figure 4.45.A). This is because the flux from the peripheral compartment is dominated by the flux from the gastro-intestinal compartment (which is grater), thus the overall appearance of ASA in the systemic compartment is little influenced by the former flux. As a consequence, as k_{sp} decreases, the acetylation of COX in MKs, both maturing and proliferating (Figure 4.46), and ProPLTs (Figure 4.47.A) increase, resulting in a more delayed recovery of COX in the PLTs compartment, as shown in Figure 4.47.B.

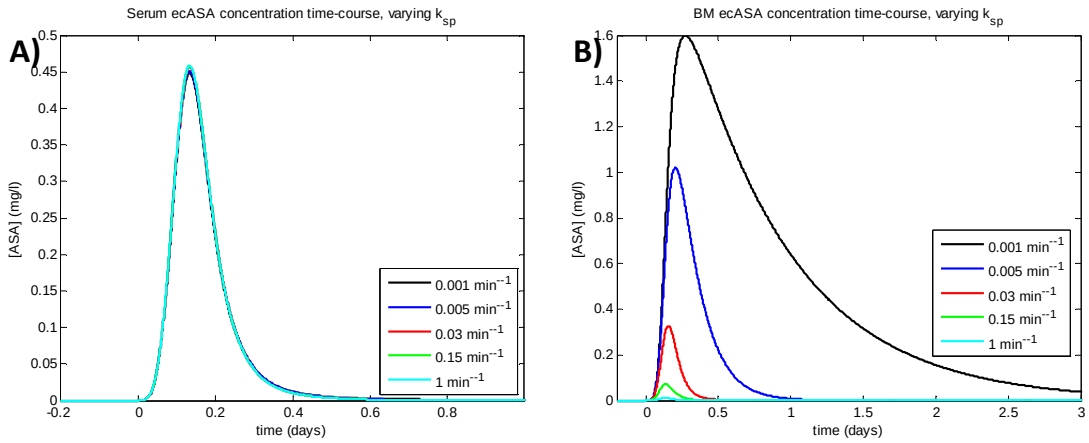


Figure 4.45: *ASA concentration in serum (A) and in bone marrow (B), increasing the parameter k_{sp} (0.001, 0.005, 0.03, 0.15 and 1 min^{-1}). Single aspirin intake at t = 0.*

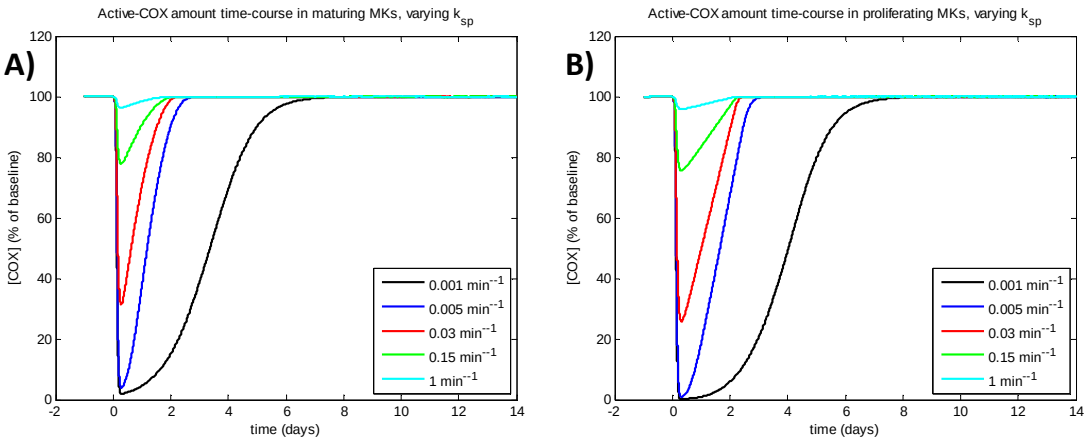


Figure 4.46: *COX time-course in maturing (A) and proliferating (B) MKs, increasing the parameter k_{sp} (0.001, 0.005, 0.03, 0.15 and 1 min^{-1}). single aspirin intake at t = 0.*

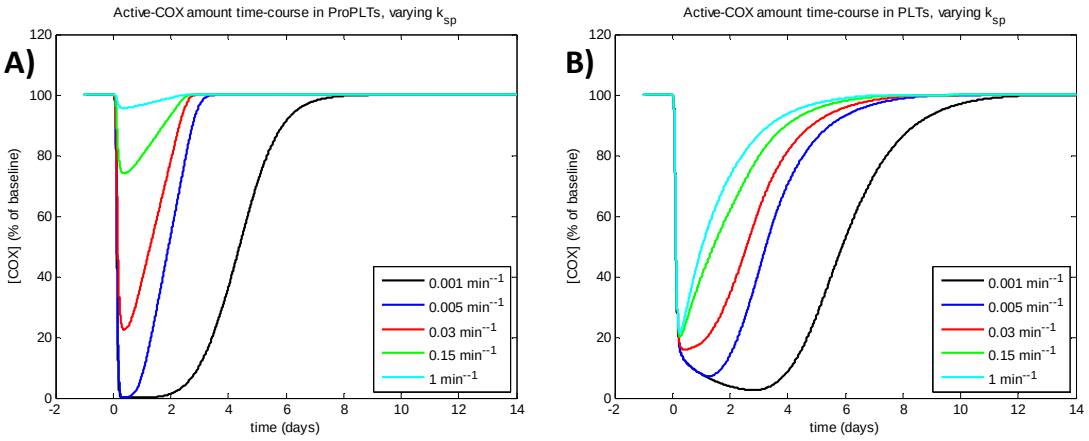


Figure 4.47: *COX time-course in ProPLTs (A) and PLTs (B), increasing the parameter k_{sp} (0.001, 0.005, 0.03, 0.15 and 1 min^{-1}). Single aspirin intake at t = 0.*

Figure 4.48 shows the output variables *lag-time* (panel A) and *rise-time* (panel B) as functions of k_{sp}, for a simulation of one week therapy 100 mg ecASA once a day.

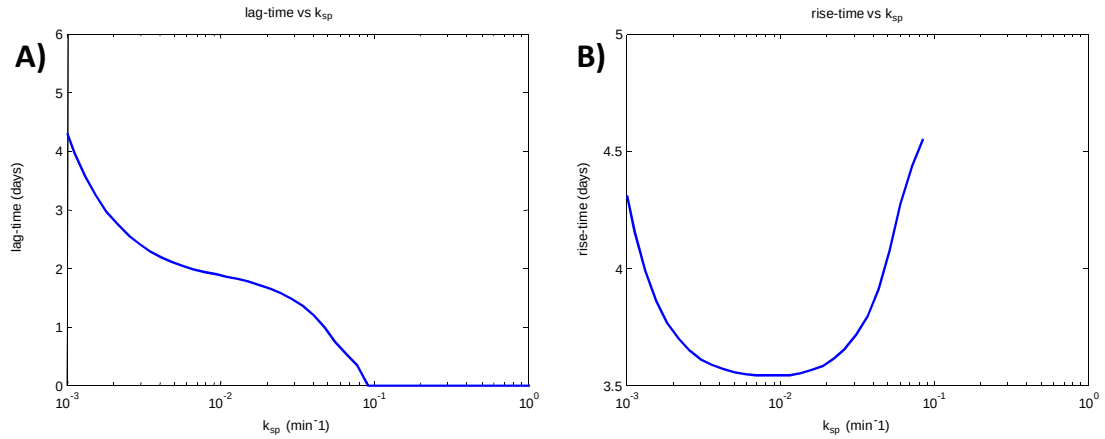


Figure 4.48: *lag-time (A) and rise-time (B) as function of the parameter k_{sp}.*

In the opposite way respect to k_{ps}, small values of k_{sp} result in a too high lag-time, while great values of k_{sp} result in no lag-time (i.e. no sufficient acetylation in bone marrow). This is because small values of k_{sp} cause ASA to accumulate in bone marrow, requiring several days before a complete elimination. As one can see, for k_{sp} values greater than ~ 10^{-1} the lag-time becomes null and, as a consequence, the rise-time cannot be computed since the maximal ecetylation of COX in PLTs is too small. The rise-time exhibits a minimum for a k_{sp} values of about 10^{-2}. A lag-time of about 2 days is obtained for k_{sp} values of about 10-^{2} results in a lag-time of about 2 days, thus a good choice for k_{sp} is likely to be near this value.

4.5.1.8 Sensitivity results

Table 4.4 reports the mean sensitivity of each output variable to each unknown parameter.

Output variable	**Parameter**						
	p_{MK}	λ	t_flex	m	k_{el}	k_{ps}	k_{sp}
lag-time	-0.005	0.360	0.129	-0.085	-0.274	0.503	-0.516
rise-time	0.001	-0.035	0.018	0.083	0.044	-0.032	-0.011

Table 4.4: *mean sensitivities of lag-time and rise-time to each unknown parameter.*

The parameters regulating aspirin exchanges between systemic and peripheral compartment, i.e. k_{ps} and k_{sp}, have a major effect on the output variable lag-time. λ and *t_flex* have a non-negligible effect on lag-time too, even if lower, and the weakest effect is the one of p_{MK}. On the other hand, *m* is the parameter exhibiting the greater effect on the output variable rise-time, followed by k_{el} and by λ and k_{ps}. Even for the rise-time p_{MK} is the parameter with the weakest effect.

4.5.1.9 Final parameter values

Given the indicative results of simulations performed to investigate the effect of each unknown parameter on the output, the final list of values for all the parameters is reported in Table 4.5. Each known parameter has been set to its mean literature value (see Table 4.2), while for each unknown parameter a reasonable value has been set on the basis of indications described in the previous sections.

	Parameter	final value
THROMBOPIESIS	*M_MKs_per_Kg*	1.6e6
	N_ProPLTs_per_MK	1000
	N_PLTs_per_ProPLT	2
	MK_matur	3 days
	MK_prolif	7 hours
	ProPLT_life	18 hours
	PLT_life	8.5 days
COX KINETICS	V_{QBM}	1000 ml
	V_{Q4}	15 ml
	k	$4.8e^{-4}$ min^{-1}
	p_{MK}	$1e^{-15}$ g/min
ASA PK	*t_flex*	180 mins
	m	4
	k_a	0.1875 min^{-1}
	k_{ps}	0.01 min^{-1}
	k_{sp}	0.01 min^{-1}
	k_{el}	0.2 min^{-1}
	V_{X2}	5600 ml
	V_{X3}	1177 ml
ASA PD	λ	$2e^{3}$ $mol^{-1}min^{-1}$

Table 4.5: *Final parameter values.*

4.6 *Data*

Data available to evaluate the model performances come from the clinical trial performed on healthy subjects described in section 4.2.2.1. Data are relative to 48 healthy Caucasian subjects randomized to 1 to 8 groups, according to treatment duration, ranging from 1 to 8 weeks. Each patient received enteric-coated aspirin 100 mg once a day and was instructed to take tablets at the same time of the day. Serum TxB_2 (in ng/ml) was measured at the end of each week of aspirin, and at days 1, 2, 3 and 7 after withdrawal [133].

In Figure 4.49 the mean curves (as percentage of baseline) for each group are presented, where also the steady-state value of TxB_2 during aspirin treatment was added as initial value at time $t = 0$.

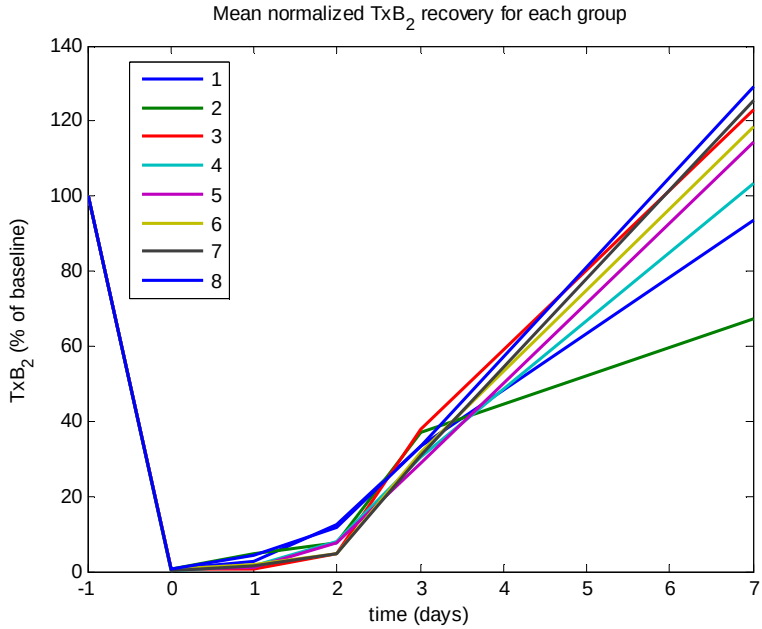


Figure 4.49: *Mean curves of TxB2 baseline and recovery after aspirin therapy for each group of treatment.*

Authors observed that the overall kinetics of TxB_2 recovery showed a complex sigmoidal pattern and that initial recovery of serumTxB_2 levels seem to differ among groups. Since, by visual inspection of the data, it is difficult to observe differences among groups, the mean curve of all the data was computed (Figure 4.50) and the model was tested on it.

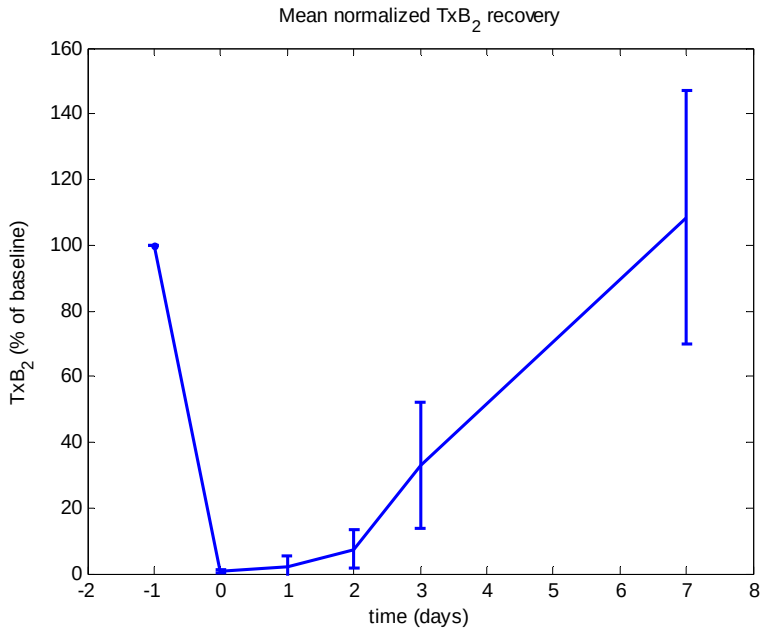


Figure 4.50: *Mean curve (± SD) over all the 48 subjects of TxB2 recovery (as percentage of baseline) after aspirin therapy.*

4.7 *Simulations*

4.7.1 Healthy subjects

A qualitative description of the data presented in the previous section was performed, running a simulation of one week reproducing the same aspirin regimen of experimental data (100 mg ecASA every 24 hours), using parameters values reported in Table 4.5.

Figure 4.51 shows the results of the simulation against real data.

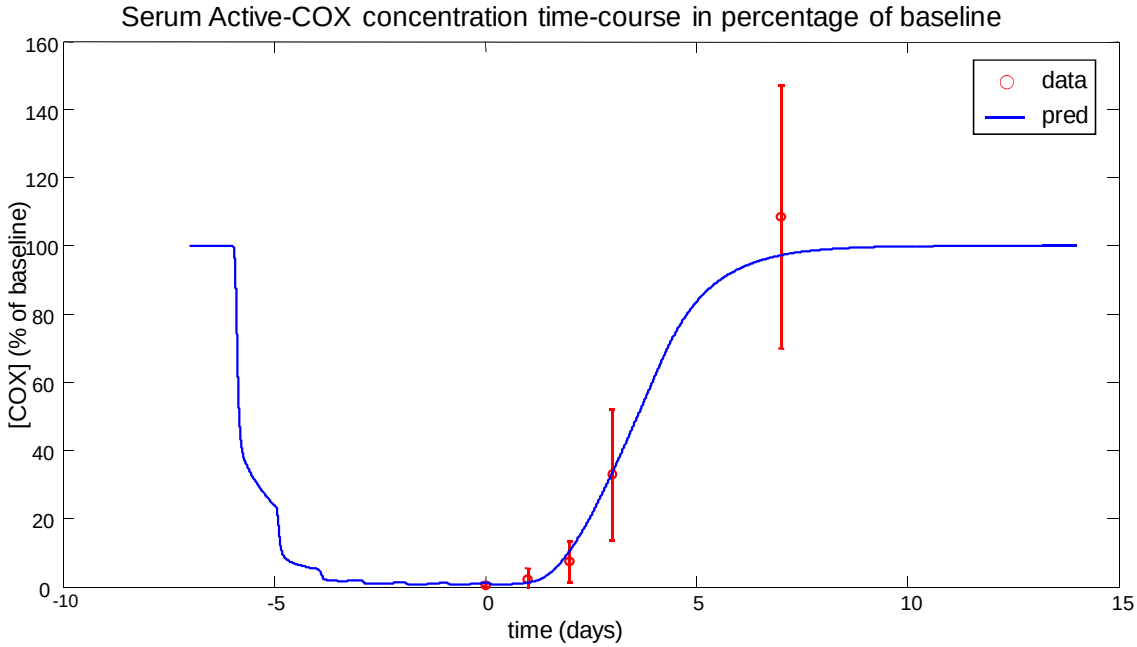


Figure 4.51: *Serum active-COX time-course prediction (in percentage of baseline value) against real data, after a week of 100 mg ASA every 24 hours. First aspirin intake is at t = -6 days. Last intake is at t = 0.*

As one can see, the prediction is quite good, since the COX time-course in serum exhibits the ~2 days delay after aspirin withdrawal and the sigmoidal shape with a complete recovery about one week after withdrawal.
Moreover, it can be seen how 3-4 intakes are needed in order to obtain the maximal effect.

In particular, the curve of Figure 4.51 is characterized by the following output parameters:

Simulation on healthy subjects	
maximal acetylation of serum COX	99.2 % of baseline
lag-time of COX recovery	1.9 days
rise-time of COX recovery	3.6 days
COX recovery at 7 days	95 % of baseline
slope of COX recovery between 12 and 24 hours	0.042 %/h

Table 4.6: *Output parameters of the simulated COX recovery for healthy subjects.*

4.7.2 Diabetic subjects

The model can be used also as a mean to test potential mechanisms for the diminished response to aspirin in diabetic patients, leading to treatment failure in most cases (the so-called 'aspirin resistance'). In the literature it has been hypothesized that the faster TxB_2 recovery after an aspirin intake characterizing diabetic patients may be caused by an enhanced COX biosynthesis due to faster platelet turnover.

In order to investigate how variations in the parameters of thrombopoiesis and COX kinetics affect the output, a sensitivity analysis, similar to the one conducted for the unknown parameter in section 4.5, has been performed. The tested parameters, together with nominal values and search intervals, are reported in Table 4.7.

Parameter	nominal value	Search interval
MK_matur	3 days	0.25 ÷ 4
MK_prolif	7 hours	1 ÷ 48
ProPLT_life	18 hours	1 ÷ 48
PLT_life	8.5 days	2 ÷ 14
k	$4.8e^{-4}$ min^{-1}	$6.8e^{-5} \div 4.8e^{-3}$

Table 4.7: *Nominal values and search intervals for the parameters of thrombopoiesis and COX kinetics.*

The mean sensitivities of the lag-time and rise-time to each tested parameter are reported in Table 4.8.

Output variable	Parameter				
	MK_matur	MK_prolif	ProPLT_life	PLT_life	k
lag-time	0.212	0.159	0.388	0.012	-0.152
rise-time	0.201	0.053	-0.001	0.077	-0.292

Table 4.8: *mean sensitivities of lag-time and rise-time to each tested parameter.*

As one can see, the parameter which most affect the lag-time is the life of the proplatelets (*ProPLT_life*), in particular, as *ProPLT_life* increases the lag-time increases. This result confirms the hypothesis made in section 4.4.2.2 about the major role of *ProPLT_life* in determining the delay in platelets COX recovery, thus it is quite encouraging. On the other hand, *ProPLT_life* minimally affects the rise-time, meaning that this parameter is responsible only for the delay but not for the slope in the recovery of platelets COX after aspirin intakes. The rise-time is most influenced by the rate of COX degradation (*k*): as *k* decreases the overall COX kinetics becomes slower, thus the rise-time increases. It is interesting to note how the parameter *MK_matur*, representing the maturation period of megakaryocytes, has a strong effect on both the lag-time and the rise-time. This means that a longer maturation period of megakaryocyte results in a slower recovery of platelets COX, producing both a longer delay and a lower slope of the recovery.

The main difference between healthy and diabetic subjects, depicted in results from clinical trials described in section 4.2.2, is the speed of COX recovery after aspirin intakes. From Table 4.8, the parameters which most affect the recovery of COX after an aspirin intake are the maturation time of megakaryocytes (*MK_matur*) and the COX degradation rate (*k*). The parameter *ProPLT_life* instead simply results in a shift of COX recovery, but not in a faster recovery. Thus, a possible mechanism to explain the faster COX recovery in diabetic patients can be represented by an enhanced megakaryocyte turnover (in particular a faster maturation) and an increased utilization of COX by platelets. For example, using all the values reported in Table 4.5, except for *MK_matur* and *k* which were decreased from 3 days to 1 day and from $4.8e^{-5}$ min^{-1} to $1e^{-5}$ min^{-1} respectively, the recovery of serum COX in a diabetic patient was simulated, in response to the same therapy of one week 100 mg ecASA once a day undergone by healthy subjects. Results are shown in Figure 4.52, where the simulated time-course of serum COX for the diabetic patient (green curve) is compared to the one for healthy subjects (blue curve). By visual inspection, COX recovery results markedly faster in case of diabetes, though the maximal acetylation is comparable to the one of healthy subjects, as confirmed by results from clinical trials (see section 4.2.2.2). Table 4.9 reports the output parameters of the diabetic curve, compared to the healthy ones. As one can see, maximal acetylation of serum COX is almost complete in both cases, and the main difference is represented by the speed of the recovery, in particular both the lag-time and the rise-time are smaller and the slope of the recovery between 12 and 24 hours is doubled.

Simulation on subjects:	healthy	diabetic
maximal acetylation	99.2 % of baseline	98 % of baseline
lag-time	1.9 days	1.2 days
rise-time	3.6 days	2.5 days
recovery at 7 days	95 % of baseline	100 % of baseline
slope of COX recovery between 12 and 24 hours	0.042 %/h	0.098 %/h

Table 4.9: *Output parameters of the simulated COX recovery for a diabetic subject, compared to healthy subjects*

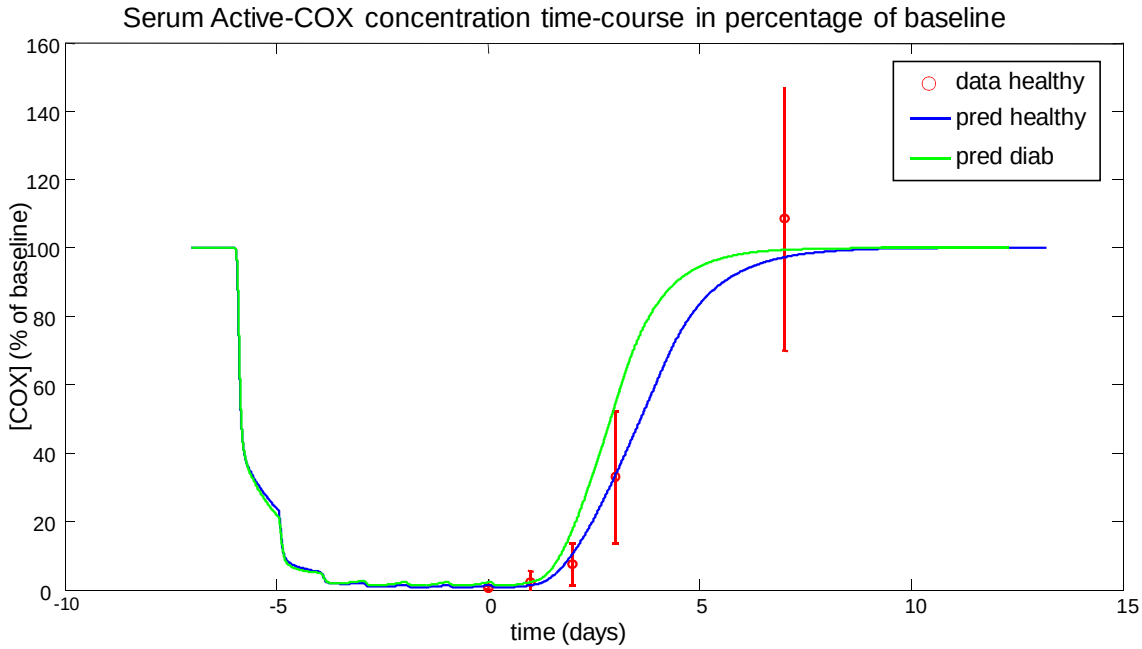


Figure 4.52: *Simulation of COX time-course in diabetic patients (green curve) versus healthy patients (blue curve) after a week of aspirin therapy 100 mg once a day. Last aspirin intake is at t = 0.*

4.7.3 Different aspirin regimens

The model was also tested to explain the effect of different aspirin regimens. Following the experimental protocol of [128], the slope of COX recovery between 12 and 24 hours after aspirin intake was computed for three different aspirin therapies: *i)* one week of 100 mg ecASA once a day every 24 hours (100od), *ii)* one week of 200 mg ecASA once a day every 24 hours (200od), *iii)* one week of 100mg ecASA twice a day every 12 hours (100bd). Results are reported in and in Figure 4.53 and Table 4.10. The model predicted a stronger effect for both the 200od and the 100bd therapy, compared to the 100od therapy, but was not able to correctly predict a stronger effect of the 100bd therapy with respect to the 200od therapy. The most probable explanation is that, the stronger effect of the 100bd

therapy is due to some non-linearity in the kinetics of aspirin, which the model is not able to reproduce.

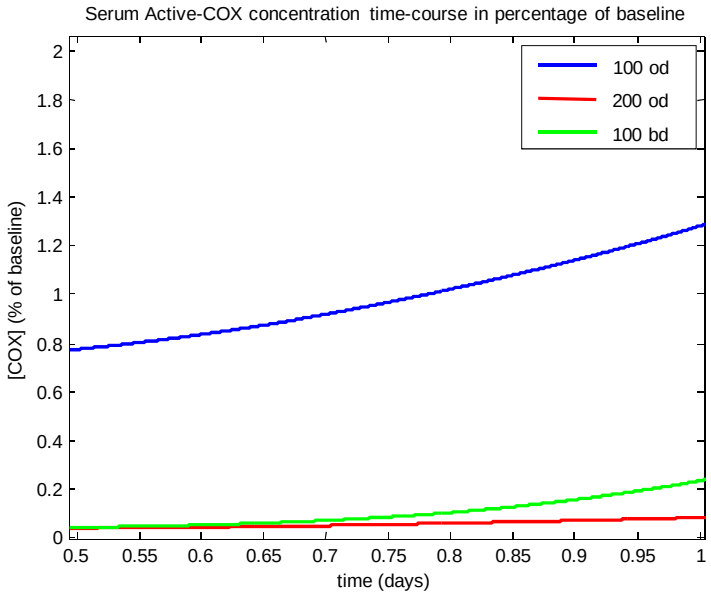


Figure 4.53: *Serum COX recovery between 12 and 24 hours after intake (at t = 0) for different aspirin regimens: 100 mg once a day (blue), 200 mg once a day (red) and 100 mg twice a day (green).*

aspirin regimen	slope of COX recovery between 12 and 24 hours
100od	0.042 %/h
200od	0.011 %/h
100bd	0.015 %/h

Table 4.10: *Slope of COX recovery between 12 and 24 hours (as percentage of baseline per hour) for different aspirin regimens: 100 mg once a day (100od), 200 mg once a day (200od) and 100 mg twice a day (100bd).*

4.8 *Discussion*

In this chapter the problem of modeling the effect of the treatment on the phenotype was discussed, focusing on the preventive effect of aspirin against atherothrombosis and cardiovascular episodes.

The effect of aspirin in the prevention of cardiovascular complications has been widely studied and reported in the literature. In particular, the comparison between healthy subjects and diabetic ones results in a quite different picture: while for the formers aspirin

has been proofed to have a significant effect, a clear benefit of aspirin in the prevention of major cardiovascular events in people with diabetes remains unproved.

From a biological point of view, the best-characterized mechanism of action of aspirin is the inhibition of thromboxane-dependent platelet function, through permanent inactivation of the COX-1, an enzyme which leads to the final formation of thromboxane TxB_2, the major promoter of platelets activation and aggregation.
Although widely studied in many clinical trials, a characterization of this mechanisms from a modeling point of view is still missing in the literature.

A compartmental model of COX kinetics and aspirin pharmacokinetics-pharmacodynamics has been developed with the objective of qualitatively describing and simulating the process of COX inhibition and reappearance in platelets in response to aspirin exposure.
The model consists of four key-elements (interconnected each other), describing: *i)* the timing of the thrombopoiesis mechanism, *ii)* COX kinetics, *iii)* aspirin PK and *iv)* aspirin PD, respectively.
The two main innovative features of the work are represented by: *i)* the distributed description adopted for COX kinetics, which makes the model capable to correctly simulate COX time-course in the different compartments according to the timing of the thrombopoiesis mechanism, *ii)* the interconnection between the aspirin PK model and the COX kinetics model (which differs from the classical approach of using the drug concentration as a forcing input of the model), which allows to correctly model the not separable interaction between aspirin and COX.

The model has been tested on data of serum thromboxane TxB_2 (which is proportional to platelets COX activity) recovery levels after aspirin withdrawal in 48 healthy subjects, treated with aspirin 100 mg daily for 1 to 8 weeks. Data are taken from [133]. Given the explorative aim of the model and the available data, the evaluation of the model was performed from a qualitative point of view, obtaining a good prediction for the time-course of COX recovering in serum.
The model, however, predicts the need for 3-4 aspirin intakes only before reaching the maximal effect of the treatment, while the authors concluded from [133] that at least 1-2 weeks of treatment are required to achieve maximal effect. The authors in fact observed

that initial recovery of serumTxB_2 levels seem to differ among groups: they performed a 2-factor repeated measurements analysis of variance with the post hoc Holm-Sidak test for pairwise comparison, and found that at days 1 and 2 following aspirin withdrawal, TxB_2 values were similar in the subjects treated for 1 and 2 weeks and significantly higher than the corresponding values of longer treatment groups (*p-value* < 0.05). Since, by visual inspection of the data (see Figure 4.49), it is difficult to observe differences among groups (also because of the high interindividual variability), a complete statistical analysis was performed, in order to confirm or reject the hypothesis of significant differences.

In particular, a one-side t-test for each sample time *t* and on each couple of consecutive groups (*i,i+1*) was performed, in order to test whether TxB_2 levels of group *i* were significantly higher than TxB_2 levels of group *i+1* at the same times: $TxB_2^{i}(t) > TxB_2^{i+1}(t)$, i.e. whether treating one more week with aspirin had a significant effect in the decrease of TxB_2 recovery. Also clustering group 1 and 2 vs other groups was tested. Results are shown in Table 4.11.

one-sided t-test: group i > group i+1					
	p-value				
test	**0 d**	**1 d**	**2 d**	**3 d**	**7 d**
1 > 2 ?	0.428	0.373	0.067	0.361	0.065
2 > 3 ?	0.284	0.057	0.214	0.428	0.001
3 > 4 ?	0.329	0.152	0.123	0.256	0.146
4 > 5 ?	0.140	0.471	0.408	0.454	0.305
5 > 6 ?	0.400	0.246	0.068	0.447	0.447
6 > 7 ?	0.226	0.277	0.480	0.437	0.366
7 > 8 ?	0.187	0.040	0.061	0.378	0.416
(1,2) > (3,4,5,6,7,8) ?	0.083	0.048	0.051	0.210	0.015

Table 4.11: *one-sided t-test for each groups couple for each time.*

From Table 4.11, we observe only one p-value < 0.05, coming from testing group 7 (7 weeks treatment) versus group 8 (8 weeks treatment) at time t = 1 day, all the other tests being not significant. Testing group 1 and 2 together (1-2 weeks treatment) versus all the other groups (3 to 8 week treatment) resulted in only one weak significant difference at time t = 1 day.

Then, a test for trend was performed, in order to test whether a trend of TxB_2 levels exists along groups, i.e. whether increasing the duration of aspirin treatment induced a

significant decrease of TxB_2 levels. A COX-Stuart test for trend for continuous data was performed, for each sample time *t*. All the six p-values were greater than the significant threshold 0.05, thus no significant effect of the treatment duration on the decrease of TxB_2 levels could be found.
Finally, the same analysis carried out by the authors was re-implemented. A 2-factor mixed ANOVA with the post hoc Holm-Sidak test for pairwise comparison was performed. The overall ANOVA found a weak significant difference among groups (p-value = 0.046), but the post hoc Holm-Sidak test was not able to significantly cluster the groups (all the p-value were greater than the significant threshold).
From this statistical analysis it is difficult to conclude about significant differences among groups and the effect of therapy duration on platelets COX recovery probably needs future insights.

The model was also tested for a potential mechanism to explain the diminished response to aspirin in diabetic patients and for explaining the effect of different aspirin regimens.
By modifying two key parameters, describing the maturation period of megakaryocytes and the COX degradation rate constant respectively, the model was able to simulate a faster COX recovery in the 12-24 hours interval after aspirin intake for diabetic subjects, thus reproducing literature findings.
The model, however, was not able to explain the greater effect (resulting in a slower COX recovery) of a therapy with intakes of 100 mg ecASA every 12 hours, compared to a therapy with intakes of 200 mg ecASA every 24 hours, thus underlining the need for future refinements in particular regarding aspirin pharmacokinetics.

In conclusion, though future improvements are needed, the actual model represents a good starting point for further refinements and investigations. Future experiments with multiple measurements (i.e. simultaneous measurements in different compartments) could help to obtain a deeper understanding of the involved phenomena, providing the model with additional information, which could help designing personalized antiplatelet regimens in diabetes mellitus.

Conclusions

In this thesis, the problem of investigating long-term complications of diabetes mellitus has been faced with a multi-level approach. Given the complex-nature of such a disease, the multi-level approach allows to characterize the phenomena of interest at different levels of detail, according to data availability. In the present work of thesis, three main levels of study have been discussed and, for each one, novel investigation methodologies have been proposed.

The context of investigation of the first level of study is the one of Genome Wide Association Studies, in which the objective is, on one hand, to detect correlation between one or more SNPs and a discrete trait (diabetes, in this case) and, on the other, to learn a rule to perform subject classification. The multivariate analysis approaches, developed so far, still suffer for the lack of precision and stability of the lists of biomarkers selected, mainly due to linkage disequilibrium, i.e. the non-random association between the true genetic causes and the SNPs in genomic regions close to them, which confounds the search for genetic biomarkers. A new algorithm, Bag of Naïve Bayes, was developed to effectively tackle this problem. BoNB is based on Naïve Bayes classification enriched by three main features to tailor the Naïve Bayes framework to Genome Wide SNP data analysis: (a) a bagging of Naïve Bayes classifiers, to improve the robustness of the predictions, (b) a novel strategy for ranking and selecting the attributes used by each bagged classifier, to enforce attribute independence, and (c) a permutation-based procedure for selecting significant biomarkers, based on their marginal utility in the classification process. The effectiveness of BoNB was demonstrated by applying it to the WTCCC case-control study on Type 1 Diabetes: BoNB outperforms two algorithms from the state of the art (a Naïve Bayes Classifier and HyperLASSO) in terms of classification performance, and all the genetic biomarkers identified by BoNB are meaningful for Type 1 Diabetes, thus confirming the good performance also in terms of precision of the selected biomarkers.

The second level of study deals with the *in-silico* modeling of complex diseases. Recently, due to alarming increasing of world's diabetic incidence, a requirement for diabetes

simulation models has been identified in the medical and healthcare policy community to facilitate the simultaneous evaluation of long-term clinical end economic effects of treatment, and, as a result, a number of models have been developed, mainly based on Markov Models. In this thesis, the progression of two vascular diabetes complications (Cardiovascular disease and Nephropathy) was modeled using Dynamic Bayesian Networks, which, differently from Markov Models, are more powerful since they allow a more easy handle of information. The model was developed on the DCCT dataset, integrating both phenotypic information and information on treatment. Results regarding the simulated progression of complications show very good performances, exhibiting a prediction accuracy greater than 95 % for the considered outcomes, , thus proving the effectiveness of the model. Moreover, the flexible structure of the DBN makes the model suitable for future developments, such as the introduction of diabetic Retinopathy, as an additional outcome, and the genotypic information, as a potential mean to improve predictions. Based on the DBN model, a web Java application, which will implement also cost-effectiveness and cost-utility analyses, is currently under development.

The last level of study focuses on the *in-silico* modeling of drug action, in particular regarding the effect of aspirin against atherothrombosis and cardiovascular episodes. A compartmental model of aspirin PKPD was developed from literature information, in order to simulate the inhibition of COX enzyme (the major promoter of platelets activation and aggregation, which leads to the formation of thrombi) by aspirin. The model was built on four interconnected key-elements, describing thrombopoiesis mechanism, COX kinetics, aspirin pharmacokinetics and aspirin pharmacodynamics, respectively. Innovative features of the work are represented by the distributed description adopted for COX kinetics and by the not separable interconnection between aspirin PK and COX kinetics, which allow to potentially simulate response to any drug exposure, without using any forcing input. Given the explorative aim of the work, the model was used to qualitatively describe data of healthy subjects, as well to test potential mechanisms for the diminished response, exhibited by diabetic patients, to aspirin therapy (the so-called ‘aspirin resistance’). Although representing a good starting point, the model needs further refinements and investigations: future experiments and additional data will make the model suitable to help designing personalized antiplatelet regimens in diabetes mellitus.

Bibliography

[1] Diabetes - Overview. NHS. Retrieved 2013-07-14

[2] Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. (2010) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26(3): 392-398.

[3] Armitage P. (1955) Tests for linear trends in proportions and frequencies. Biometrics 11(3): 375-386.

[4] Ayers KL, Cordell HJ. (2010) SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genet Epidemiol 34(8): 879-891.

[5] Balding DJ, Bishop M, Cannings C. (2008) Handbook of statistical genetics. : Wiley. com.

[6] Barrett JC, Cardon LR. (2006) Evaluating coverage of genome-wide association studies. Nat Genet 38(6): 659-662.

[7] Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, et al. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41(6): 703-707.

[8] Benjamini Y, Hochberg Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society.Series B (Methodological) : 289-300.

[9] Bonferroni CE. (1936) Teoria statistica delle classi e calcolo delle probabilita. .

[10] Boulesteix A, Slawski M. (2009) Stability and aggregation of ranked gene lists. Briefings in Bioinformatics 10(5): 556-568.

[11] Breiman L. (2001) Random forests. Mach Learning 45(1): 5-32.

[12] Breiman L. (1996) Bagging predictors. Mach Learning 24(2): 123-140.

[13] Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, et al. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145): 661-678.

[14] Caramori M, Fioretto P, Mauer M. (2000) The need for early predictors of diabetic nephropathy risk: Is albumin excretion rate sufficient? Diabetes 49(9): 1399-1408.

[15] Chuang H, Lee E, Liu Y, Lee D, Ideker T. (2007) Network-based classification of breast cancer metastasis. Molecular Systems Biology 3(1).

[16] Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, et al. (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37(11): 1243-1246.

[17] Davis CA, Gerick F, Hintermair V, Friedel CC, Fundel K, et al. (2006) Reliable gene signatures for microarray classification: Assessment of stability and performance. Bioinformatics 22(19): 2356-2363.

[18] Di Camillo B, Sanavia T, Martini M, Jurman G, Sambo F, et al. (2012) Effect of size and heterogeneity of samples on biomarker discovery: Synthetic and real data assessment. PloS One 7(3): e32200.

[19] Dudbridge F, Gusnanto A. (2008) Estimation of significance thresholds for genomewide association scans. Genet Epidemiol 32(3): 227-234.

[20] Efron B, Tibshirani R. An introduction to the bootstrap. 1993 chapman & hall new york. .

[21] Ein-Dor L, Kela I, Getz G, Givol D, Domany E. (2005) Outcome signature genes in breast cancer: Is there a unique set? Bioinformatics 21(2): 171-178.

[22] Ein-Dor L, Zuk O, Domany E. (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proceedings of the National Academy of Sciences 103(15): 5923-5928.

[23] Finne P, Reunanen A, Stenman S, Groop P, Grönhagen-Riska C. (2005) Incidence of end-stage renal disease in patients with type 1 diabetes. JAMA: The Journal of the American Medical Association 294(14): 1782-1787.

[24] Groop P, Forsblom C. (2003) Diabetic nephropathy—an acquired or inherited disease? 1253: 149-161.

[25] Haffner SM, Lehto S, Rönnemaa T, Pyörälä K, Laakso M. (1998) Mortality from coronary heart disease in subjects with type 2 diabetes and in nondiabetic subjects with and without prior myocardial infarction. N Engl J Med 339(4): 229-234.

[26] Hall N. (2007) Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol 210(9): 1518-1525.

[27] Henriksson F, Agardh C, Berne C, Bolinder J, Lönnqvist F, et al. (2000) Direct medical costs for patients with type 2 diabetes in sweden. J Intern Med 248(5): 387-396.

[28] Hoerl AE, Kennard RW. (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1): 55-67.

[29] Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genetics 4(7): e1000130.

[30] Hoh J, Ott J. (2004) Genetic dissection of diseases: Design and methods. Curr Opin Genet Dev 14(3): 229-232.

[31] Hysi PG, Young TL, Mackey DA, Andrew T, Fernández-Medarde A, et al. (2010) A genome-wide association study for myopia and refractive error identifies a susceptibility locus at 15q25. Nat Genet 42(10): 902-905.

[32] Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, et al. (2008) Repeatability of published microarray gene expression analyses. Nat Genet 41(2): 149-155.

[33] Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, et al. (2005) Multiple-laboratory comparison of microarray platforms. Nature Methods 2(5): 345-350.

[34] Isomaa B, Almgren P, Tuomi T, Forsén B, Lahti K, et al. (2001) Cardiovascular morbidity and mortality associated with the metabolic syndrome. Diabetes Care 24(4): 683-689.

[35] Kim S. (2009) Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. BMC Bioinformatics 10(1): 147.

[36] Klein R, Klein BE, Moss SE, Cruickshanks KJ. (1994) The wisconsin epidemiologic study of diabetic retinopathy: XIV. ten-year incidence and progression of diabetic retinopathy. Arch Ophthalmol 112(9): 1217.

[37] Kollar D, Friedman N. (2009) Probabilistic graphical models: Principles and techniques. : The MIT Press.

[38] Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J. (2005) Independence and reproducibility across microarray platforms. Nature Methods 2(5): 337-344.

[39] Lewis CM. (2002) Genetic association studies: Design, analysis and interpretation. Briefings in Bioinformatics 3(2): 146-153.

[40] Manly KF, Nettleton D, Hwang JG. (2004) Genomics, prior probability, and statistical tests of multiple hypotheses. Genome Res 14(6): 997-1001.

[41] Mitchell TM. (1997) Machine learning. 1997. Burr Ridge, IL: McGraw Hill 45.

[42] Nepom, MD, Ph. D, Gerald T. (1995) Class II antigens and disease susceptibility. Annu Rev Med 46(1): 17-25.

[43] Prentice RL, Pyke R. (1979) Logistic disease incidence models and case-control studies. Biometrika 66(3): 403-411.

[44] Sebastiani P, Abad-Grau M. (2007) Bayesian networks for genetic analysis. Systems Bioinformatics: An Engineering Case-Based Approach : 205-227.

[45] Sebastiani P, Solovieff N, Puca A, Hartley SW, Melista E, et al. (2010) Genetic signatures of exceptional longevity in humans. Science 10: 1126.

[46] Sebastiani P, Timofeev N, Dworkis DA, Perls TT, Steinberg MH. (2009) Genome-wide association studies and the genetic dissection of complex traits. Am J Hematol 84(8): 504-515.

[47] Shi L, Campbell G, Jones WD, Campagne F, Wen Z, et al. (2010) The MicroArray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 28(8): 827-838.

[48] Soedamah-Muthu SS, Fuller JH, Mulnier HE, Raleigh VS, Lawrenson RA, et al. (2006) High risk of cardiovascular disease in patients with type 1 diabetes in the UK A cohort study using the general practice research database. Diabetes Care 29(4): 798-804.

[49] Solé X, Bonifaci N, López-Bigas N, Berenguer A, Hernández P, et al. (2009) Biological convergence of cancer signatures. PLoS One 4(2): e4544.

[50] Tibshirani R. (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society.Series B (Methodological) : 267-288.

[51] Tisch R, McDevitt H. (1996) Insulin-dependent diabetes mellitus. Cell 85(3): 291-297.

[52] Weir BS. (1997) Genetic data analysis. .

[53] Wigginton JE, Cutler DJ, Abecasis GR. (2005) A note on exact tests of hardy-weinberg equilibrium. The American Journal of Human Genetics 76(5): 887-893.

[54] Wu TT, Chen YF, Hastie T, Sobel E, Lange K. (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25(6): 714-721.

[55] Zhang C. (2010) Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics 38(2): 894-942.

[56] Zimmet P, Alberti K, Shaw J. (2001) Global and societal implications of the diabetes epidemic. Nature 414(6865): 782-787.

[57] Zou H, Hastie T. (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(2): 301-320.

[58] Zucknick M, Richardson S, Stronach EA. (2008) Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. Stat Appl Genet Mol Biol 7(1): 7.

[59] American Diabetes Association. (2013) Standards of medical care in diabetes—2013. Diabetes Care 36(Supplement 1): S11-S66.

[60] Byers T, Levin B, Rothenberger D, Dodd GD, Smith RA. (1997) American cancer society guidelines for screening and surveillance for early detection of colorectal polyps and cancer: Update 1997. CA: A Cancer Journal for Clinicians 47(3): 154-160.

[61] Diabetes Control and Complications Trial Research Group. (1986) The diabetes control and complications trial (DCCT): Design and methodologic considerations for the feasibility phase. Diabetes 35(5): 530-545.

[62] Eastman RC, Javitt JC, Herman WH, Dasbach EJ, Copley-Merriman C, et al. (1997) Model of complications of NIDDM: II. analysis of the health benefits and cost-effectiveness of treating NIDDM with the goal of normoglycemia. Diabetes Care 20(5): 735-744.

[63] Eastman RC, Javitt JC, Herman WH, Dasbach EJ, Zbrozek AS, et al. (1997) Model of complications of NIDDM: I. model construction and assumptions. Diabetes Care 20(5): 725-734.

[64] Friedman N, Linial M, Nachman I, Pe'er D. (2000) Using bayesian networks to analyze expression data. Journal of Computational Biology 7(3-4): 601-620.

[65] Gelfand AE, Smith AF. (1990) Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85(410): 398-409.

[66] Glover F, Laguna M. (1997) Tabu search. : Springer.

[67] Heckerman D, Geiger D, Chickering DM. (1995) Learning bayesian networks: The combination of knowledge and statistical data. Mach Learning 20(3): 197-243.

[68] Jensen FV, Nielsen TD. (2007) Bayesian networks and decision graphs. : Springer.

[69] Johannesson M, Jönsson B, Kjekshus J, Olsson AG, Pedersen TR, et al. (1997) Cost effectiveness of simvastatin treatment to lower cholesterol levels in patients with coronary heart disease. N Engl J Med 336(5): 332-336.

[70] Klein R, Klein BE, Moss SE, Davis MD, DeMets DL. (1989) The wisconsin epidemiologic study of diabetic retinopathy: X. four-year incidence and progression of diabetic retinopathy when age at diagnosis is 30 years or more. Arch Ophthalmol 107(2): 244.

[71] Kuntz KM, Weinstein MC. (2001) Modelling in economic evaluation. Economic Evaluation in Health Care: Merging Theory with Practice : 141-171.

[72] Larsen ML, Hørder M, Mogensen EF. (1990) Effect of long-term monitoring of glycosylated hemoglobin levels in insulin-dependent diabetes mellitus. N Engl J Med 323(15): 1021-1025.

[73] Mandel JS, Bond JH, Church TR, Snover DC, Bradley GM, et al. (1993) Reducing mortality from colorectal cancer by screening for fecal occult blood. N Engl J Med 328(19): 1365-1371.

[74] Mueller E, Maxion-Bergemann S, Gultyaev D, Walzer S, Freemantle N, et al. (2006) Development and validation of the economic assessment of glycemic control and long-term effects of diabetes (EAGLE) model. Diabetes Technology & Therapeutics 8(2): 219-236.

[75] Murphy KP. (2002) Dynamic Bayesian Networks: Representation, Inference and Learning .

[76] Nishida C, Ko G, Kumanyika S. (2009) Body fat distribution and noncommunicable diseases in populations: Overview of the 2008 WHO expert consultation on waist circumference and Waist–Hip ratio. Eur J Clin Nutr 64(1): 2-5.

[77] Oppenheimer GM. (2010) Framingham heart study: The first 20 years. Prog Cardiovasc Dis 53(1): 55-61.

[78] Palmer AJ. (2013) Computer modeling of diabetes and its complications: A report on the fifth mount hood challenge meeting. Value in Health .

[79] Palmer AJ, Roze S, Valentine WJ, Minshall ME, Foos V, et al. (2004) The CORE diabetes model: Projecting long-term clinical outcomes, costs and costeffectiveness of interventions in diabetes mellitus (types 1 and 2) to support clinical and reimbursement decision-making. Current Medical Research and Opinion® 20(S1): S5-S26.

[80] Palmer AJ, Roze S, Valentine WJ, Minshallb ME, Hayes C, et al. (2004) Impact of changes in HbA1c, lipids and blood pressure on long-term outcomes in type 2 diabetes patients: An analysis using the CORE diabetes model. Current Medical Research and Opinion® 20(S1): S53-S58.

[81] Palmer AJ, Rozea S, Valentinea WJ, Minshallb ME, Foosa V, et al. (2004) Validation of the CORE diabetes model against epidemiological and clinical studies. Current Medical Research and Opinion® 20(S1): S27-S40.

[82] Ramsey SD, McIntosh M, Etzioni R, Urban N. (2000) Simulation modeling of outcomes and cost effectiveness. Hematol Oncol Clin North Am 14(4): 925-938.

[83] Sebastiani P, Abad M, Ramoni MF. (2005) Bayesian networks for genomic analysis. Genomic Signal Processing and Statistics : 281-320.

[84] Solano MP, Goldberg RB. (2006) Lipid management in type 2 diabetes. Clinical Diabetes 24(1): 27-32.

[85] Sonnenberg FA, Beck JR. (1993) Markov models in medical decision making a practical guide. Medical Decision Making 13(4): 322-338.

[86] Spruance SL, Reid JE, Grace M, Samore M. (2004) Hazard ratio in clinical trials. Antimicrob Agents Chemother 48(8): 2787-2792.

[87] Streufert S, Satish U, Barach P. (2001) Improving medical care: The use of simulation technology. Simulation & Gaming 32(2): 164-174.

[88] Tsamardinos I, Brown LE, Aliferis CF. (2006) The max-min hill-climbing bayesian network structure learning algorithm. Mach Learning 65(1): 31-78.

[89] Turner R, Holman R, Cull C, Stratton I, Matthews D, et al. (1998) Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet 352(9131): 837-853.

[90] Weinstein MC, Toy EL, Sandberg EA, Neumann PJ, Evans JS, et al. (2001) Modeling for health care and other policy decisions: Uses, roles, and validity. Value in Health 4(5): 348-361.

[91] World Health Organization. Global Database on Body Mass Index.Bmi Classification.2006 .

[92] Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED. (2004) Advances to bayesian network inference for generating causal networks from observational biological data. Bioinformatics 20(18): 3594-3603.

[93] Ali M, McDonald J, Thiessen J, Coates P. (1980) Plasma acetylsalicylate and salicylate and platelet cyclooxygenase activity following plain and enteric-coated aspirin. Stroke 11(1): 9-13.

[94] Belch J, MacCuish A, Campbell I, Cobbe S, Taylor R, et al. (2008) The prevention of progression of arterial disease and diabetes (POPADAD) trial: Factorial randomised placebo controlled trial of aspirin and antioxidants in patients with diabetes and asymptomatic peripheral arterial disease. BMJ: British Medical Journal 337.

[95] Bellu G, Saccomani MP, Audoly S, D'Angiò L. (2007) DAISY: A new software tool to test global identifiability of biological and physiological systems. Comput Methods Programs Biomed 88(1): 52-61.

[96] Burch JW, Stanford N, Majerus PW. (1978) Inhibition of platelet prostaglandin synthetase by oral aspirin. J Clin Invest 61(2): 314.

[97] Davì G, Catalano I, Averna M, Notarbartolo A, Strano A, et al. (1990) Thromboxane biosynthesis and platelet function in type II diabetes mellitus. N Engl J Med 322(25): 1769-1774.

[98] De Berardis G, Sacco M, Strippoli GF, Pellegrini F, Graziano G, et al. (2009) Aspirin for primary prevention of cardiovascular events in people with diabetes: Meta-analysis of randomised controlled trials. BMJ: British Medical Journal 339.

[99] De Gaetano G. (2001) Low-dose aspirin and vitamin E in people at cardiovascular risk: A randomised trial in general practice. collaborative group of the primary prevention project. Lancet 357(9250): 89-95.

[100] DiChiara J, Bliden KP, Tantry US, Hamed MS, Antonino MJ, et al. (2007) The effect of aspirin dosing on platelet function in diabetic and nondiabetic patients an analysis from the aspirin-induced platelet effect (ASPECT) study. Diabetes 56(12): 3014-3019.

[101] Evangelista V, De Berardis G, Totani L, Avanzini F, Giorda C, et al. (2007) Persistent platelet activation in patients with type 2 diabetes treated with low doses of aspirin. Journal of Thrombosis and Haemostasis 5(11): 2197-2203.

[102] Freedman JE. (2006) The aspirin resistance controversy clinical entity or platelet heterogeneity? Circulation 113(25): 2865-2867.

[103] Goltsov A, Lebedeva G, Humphery-Smith I, Goltsov G, Demin O, et al. (2010) In silico screening of nonsteroidal anti-inflammatory drugs and their combined action on prostaglandin H synthase-1. Pharmaceuticals 3(7): 2059-2081.

[104] Hansson GK. (2005) Inflammation, atherosclerosis, and coronary artery disease. N Engl J Med 352(16): 1685-1695.

[105] Hansson L, Zanchetti A, Carruthers SG, Dahlöf B, Elmfeldt D, et al. (1998) Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: Principal results of the hypertension optimal treatment (HOT) randomised trial. The Lancet 351(9118): 1755-1762.

[106] Harker LA. (1968) Megakaryocyte quantitation. J Clin Invest 47(3): 452.

[107] Harker LA. (1968) Kinetics of thrombopoiesis. J Clin Invest 47(3): 458.

[108] Harris PA, Gross J. (1974) Preliminary pharmacokinetic model for adriamycin (NSC-123127). Cancer Chemotherapy Reports.Part 1 59(4): 819-825.

[109] Hekimsoy Z, Payzin B, Örnek T, Kandoğan G. (2004) Mean platelet volume in type 2 diabetic patients. J Diabetes Complications 18(3): 173-176.

[110] Hennekens C. (1989) Final report on the aspirin component of the ongoing physicians' health study. N Engl J Med 321(3): 129-135.

[111] Hong Y, Gengo FM, Rainka MM, Bates VE, Mager DE. (2008) Population pharmacodynamic modelling of aspirin-and ibuprofen-induced inhibition of platelet aggregation in healthy subjects. Clin Pharmacokinet 47(2): 129-137.

[112] Juul-Moller S, Edvardsson N, Sorensen S, Jahnmatz B, Rosen A, et al. (1992) Double-blind trial of aspirin in primary prevention of myocardial infarction in patients with stable chronic angina pectoris. The Lancet 340(8833): 1421-1425.

[113] Kassoff A, Buzney SM, McMeel JW, Weiter JJ, Doyle GJ, et al. (1992) Aspirin effects on mortality and morbidity in patients with diabetes MellitusEarly treatment diabetic retinopathy study report 14. JAMA: The Journal of the American Medical Association 268(10): 1292-1300.

[114] Mathur A, Hong Y, Wang G, Erusalimsky JD. (2004) Assays of megakaryocyte development. In: Anonymous Platelets and Megakaryocytes. : Springer. pp. 309-322.

[115] Ogawa H, Nakayama M, Morimoto T, Uemura S, Kanauchi M, et al. (2008) Low-dose aspirin for primary prevention of atherosclerotic events in patients with type

2 diabetes. JAMA: The Journal of the American Medical Association 300(18): 2134-2141.

[116] Patel SR, Hartwig JH, Italiano JE. (2005) The biogenesis of platelets from megakaryocyte proplatelets. J Clin Invest 115(12): 3348.

[117] Patel SR, Richardson JL, Schulze H, Kahle E, Galjart N, et al. (2005) Differential roles of microtubule assembly and sliding in proplatelet formation by megakaryocytes. Blood 106(13): 4076-4085.

[118] Patrono C. (2003) Aspirin resistance: Definition, mechanisms and clinical read-outs. Journal of Thrombosis and Haemostasis 1(8): 1710-1713.

[119] Patrono C, Ciabattoni G, Patrignani P, Pugliese F, Filabozzi P, et al. (1985) Clinical pharmacology of platelet cyclooxygenase inhibition. Circulation 72(6): 1177-1184.

[120] Patrono C, Baigent C, Hirsh J, Roth G. (2008) Antiplatelet DrugsAmerican college of chest physicians evidence-based clinical practice guidelines. CHEST Journal 133(6_suppl): 199S-233S.

[121] Patrono C, Coller B, FitzGerald GA, Hirsh J, Roth G. (2004) Platelet-active drugs: The relationships among dose, effectiveness, and side effects the seventh ACCP conference on antithrombotic and thrombolytic therapy. CHEST Journal 126(3_suppl): 234S-264S.

[122] Patrono C, García Rodríguez LA, Landolfi R, Baigent C. (2005) Low-dose aspirin for the prevention of atherothrombosis. N Engl J Med 353(22): 2373-2383.

[123] Pedersen AK, FitzGerald GA. (1984) Dose-related kinetics of aspirin. presystemic acetylation of platelet cyclooxygenase. N Engl J Med 311(19): 1206.

[124] Peto R, Gray R, Collins R, Wheatley K, Hennekens C, et al. (1988) Randomised trial of prophylactic daily aspirin in british male doctors. Br Med J (Clin Res Ed) 296(6618): 313.

[125] Pignone M, Alberts MJ, Colwell JA, Cushman M, Inzucchi SE, et al. (2010) Aspirin for primary prevention of cardiovascular events in people with diabetes. J Am Coll Cardiol 55(25): 2878-2886.

[126] Porter S, Ridgway K. (1982) The permeability of enteric coatings and the dissolution rates of coated tablets. J Pharm Pharmacol 34(1): 5-8.

[127] Pulcinelli FM, Biasucci LM, Riondino S, Giubilato S, Leo A, et al. (2009) COX-1 sensitivity and thromboxane A2 production in type 1 and type 2 diabetic patients under chronic aspirin treatment. Eur Heart J 30(10): 1279-1286.

[128] Rocca B, Santilli F, Pitocco D, Mucci L, Petrucci G, et al. (2012) The recovery of platelet cyclooxygenase activity explains interindividual variability in responsiveness to low-dose aspirin in patients with and without diabetes. Journal of Thrombosis and Haemostasis 10(7): 1220-1230.

[129] Ross-Lee L, Elms M, Cham B, Bochner F, Bunce I, et al. (1982) Plasma levels of aspirin following effervescent and enteric coated tablets, and their effect on platelet function. Eur J Clin Pharmacol 23(6): 545-551.

[130] Roth GJ, Stanford N, Majerus PW. (1975) Acetylation of prostaglandin synthase by aspirin. Proceedings of the National Academy of Sciences 72(8): 3073-3076.

[131] Ruggeri ZM. (2002) Platelets in atherothrombosis. Nat Med 8(11): 1227-1234.

[132] Sai Y, Kusaka A, Imanishi K, Matsumoto M, Takahashi R, et al. (2011) A randomized, quadruple crossover single-blind study on immediate action of chewed and unchewed low-dose acetylsalicylic acid tablets in healthy volunteers. J Pharm Sci 100(9): 3884-3891.

[133] Santilli F, Rocca B, De Cristofaro R, Lattanzio S, Pietrangelo L, et al. (2009) Platelet cyclooxygenase inhibition by low-dose aspirin is not reflected consistently by platelet function assays: Implications for aspirin “resistance”. J Am Coll Cardiol 53(8): 667-677.

[134] Thon JN, Italiano JE. (2012) Does size matter in platelet production? Blood 120(8): 1552-1561.

[135] Thon JN, Italiano JE. (2010) Platelet formation. 47(3): 220-226.

[136] Trial TP. (1998) Randomised trial of low-intensity oral anticoagulation with warfarin and low-dose aspirin in the primary prevention of ischaemic heart disease in men at increased risk. the medical research Council's general practice research framework. Lancet 351(9098): 233-241.

[137] Trialists'Collaboration A. (2002) Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. Bmj 324(7329): 71-86.

[138] Vainchenker W, Bouguet J, Guichard J, Breton-Gorius J. (1979) Megakaryocyte colony formation from human bone marrow precursors. Blood 54(4): 940-945.

[139] Warner TD, Mitchell JA. (2002) Cyclooxygenase-3 (COX-3): Filling in the gaps toward a COX continuum? Proceedings of the National Academy of Sciences 99(21): 13371-13373.

[140] Wood AJ, Patrono C. (1994) Aspirin as an antiplatelet drug. N Engl J Med 330(18): 1287-1294.