



HAL
open science

Exploring Risk Factors and Prediction Models for Sudden Cardiac Death with Machine Learning

Younès Youssfi

► **To cite this version:**

Younès Youssfi. Exploring Risk Factors and Prediction Models for Sudden Cardiac Death with Machine Learning. Statistics [math.ST]. Institut Polytechnique de Paris, 2023. English. NNT : 2023IP-PAG006 . tel-04231416

HAL Id: tel-04231416

<https://theses.hal.science/tel-04231416>

Submitted on 6 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT
POLYTECHNIQUE
DE PARIS

NNT : 2023IPPAG006

Thèse de doctorat



Exploring Risk Factors and Prediction Models for Sudden Cardiac Death with Machine Learning

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Nationale de la Statistique et de l'Administration Économique

École doctorale n°574 École Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau le 12/09/2023

YOUNÈS YOUSSEFI

Composition du Jury :

Emmanuel Bacry Directeur de Recherche, Université Paris-Dauphine (CEREMADE)	Président du jury
Philippe Ravaut Professeur des Universités, Université Paris-Cité (CRESS)	Rapporteur
Robin Ryder Maître de Conférences, Université Paris-Dauphine (CEREMADE)	Rapporteur
Salima El Kolei Maître de Conférences, ENSAI (CREST)	Examineur
Sophie Donnet Directrice de Recherche, INRAE (SOLSTIS)	Examineur
Nicolas Chopin Professeur, ENSAE (CREST)	Directeur de thèse
Xavier Jouven Professeur des Universités, Université Paris-Cité (PARCC)	Directeur de thèse

Abstract

Sudden cardiac death (SCD) is defined as a sudden natural death presumed to be of cardiac cause, heralded by abrupt loss of consciousness in the presence of witness, or in the absence of witness occurring within an hour after the onset of symptoms. Despite progress in clinical profiling and interventions, it remains a major public health problem, accounting for 10 to 20% of deaths in industrialised countries, with survival after SCD below 10%. The annual incidence is estimated 350,000 in Europe, and 300,000 in the United States. Efficient treatments for SCD management are available. One of the most effective options is the use of implantable cardioverter defibrillators (ICD). However, identifying the best candidates for ICD implantation remains a difficult challenge, with disappointing results so far.

This thesis aims to address this problem, and to provide a better understanding of SCD in the general population, using statistical modeling. We analyze data from the Paris Sudden Death Expertise Center and the French National Healthcare System Database to develop three main works:

1. The first part of the thesis aims to identify new subgroups of SCD to improve current stratification guidelines, which are mainly based on cardiovascular variables. To this end, we use natural language processing methods and clustering analysis to build a meaningful representation of medical history of patients. This work is described in Chapter 3.
2. The second part aims to build a prediction model of SCD in order to propose a personalized and explainable risk score for each patient, and accurately identify very-high risk subjects in the general population. To this end, we train a supervised classification algorithm, combined with the SHapley Additive exPlanation method, to analyze all medical events that occurred up to 5 years prior to the event. This work is described in Chapter 4.
3. The last part of the thesis aims to identify the most relevant information to select in large medical history of patients. We propose a bi-level variable selection algorithm for generalized linear models, in order to identify both individual and group effects from predictors. Our algorithm is based on a Bayesian approach and uses a Sequential Monte Carlo method to estimate the posterior distribution of variables inclusion. This work is described in Chapter 5.

Résumé

La mort subite de l'adulte est définie comme une mort inattendue sans cause extracardiaque évidente, survenant avec un effondrement rapide en présence d'un témoin, ou en l'absence de témoin dans l'heure après le début des symptômes. Son incidence est estimée à 350,000 personnes par an en Europe et 300,000 personnes aux Etats-Unis, ce qui représente 10 à 20% des décès dans les pays industrialisés. Malgré les progrès réalisés dans la prise en charge, le pronostic demeure extrêmement sombre. Moins de 10% des patients sortent vivants de l'hôpital après la survenue d'une mort subite. Les défibrillateurs automatiques implantables offrent une solution thérapeutique efficace chez les patients identifiés à haut risque de mort subite. Leur identification en population générale demeure donc un enjeu de santé publique majeur, avec des résultats jusqu'à présent décevants.

Cette thèse propose des outils statistiques pour répondre à ce problème, et améliorer notre compréhension de la mort subite en population générale. Nous analysons les données du Centre d'Expertise de la Mort Subite et les bases médico-administratives de l'Assurance Maladie, pour développer trois travaux principaux :

1. La première partie de la thèse vise à identifier de nouveaux sous-groupes de mort subite pour améliorer les modèles actuels de stratification du risque, qui reposent essentiellement sur des variables cardiovasculaires. Nous utilisons des modèles d'analyse du langage naturel et de clustering pour construire une nouvelle représentation pertinente de l'historique médical des patients. Ce travail est décrit dans le Chapitre 3.
2. La deuxième partie vise à construire un modèle de prédiction de la mort subite, capable de proposer un score de risque personnalisé et explicable pour chaque patient, et d'identifier avec précision les individus à très haut risque en population générale. Nous entraînons pour cela un algorithme de classification supervisée, combiné avec l'algorithme SHapley Additive exPlanations, pour analyser l'ensemble des consommations de soin survenus jusqu'à 5 ans avant l'événement. Ce travail est décrit dans le Chapitre 4.
3. La dernière partie de la thèse vise à identifier le niveau optimal d'information à sélectionner dans des bases médico-administratives de grande dimension. Nous proposons un algorithme de sélection de variables bi-niveaux pour des modèles linéaires généralisés, permettant de distinguer les effets de groupe des effets individuels pour chaque variable. Cet algorithme repose sur une approche bayésienne et utilise une méthode de Monte Carlo séquentiel pour estimer la loi *a posteriori* de sélection des variables. Ce travail est décrit dans le Chapitre 5.

Contents

1	General Introduction, Motivations and Contributions	17
1.1	Sudden Cardiac Death	17
1.2	Description of the Data	40
1.3	Objective of the Thesis	46
1.4	Summary of Contributions	46
2	Introduction (en français)	61
2.1	La Mort Subite de l'Adulte	61
2.2	Description des Données	62
2.3	Résumé des Contributions	64
3	Identifying Subgroups of SCD with Clustering Analysis	75
3.1	Introduction	75
3.2	Methods	76
3.3	Results	80
3.4	Discussion	92
4	Personalized Prediction Model of SCD in the General Population	93
4.1	Introduction	93
4.2	Methods	94
4.3	Results	97
4.4	Discussion	108
4.5	Conclusion	109
5	Scalable Bayesian Bi-Level Variable Selection in Generalized Linear Models	111
5.1	Introduction	111
5.2	Model	112
5.3	The proposed algorithm	114
5.4	Numerical experiments	117
5.5	Conclusion	123
	Appendices	125
	Appendix A Prediction Model	127
	Bibliography	143
	List of talks	151

List of Tables

4.1	Baseline characteristics of the populations	100
A.1	Medical codes used for the CVD model	128
A.2	Baseline characteristics of the populations	138
A.3	Comparison of the predictive performances	141
A.4	Sensitivity analyses	142

List of Figures

1.1	Scheme of the circulatory system	18
1.2	Circulation of blood through the heart	19
1.3	Electrical conduction system of the heart	20
1.4	Ventricular fibrillation	22
1.5	Asystole	22
1.6	Sudden cardiac death vs heart attack	24
1.7	Average incidence rates of SCD estimated in each European country	26
1.8	Risk factors for SCD	29
1.9	The role of ejection fraction	30
1.10	Implantable cardioverter-defibrillator	31
1.11	The chain of survival	32
1.12	Risk stratification of SCD	33
1.13	AI in health	36
1.14	Various tasks covered by NLP in medicine	38
1.15	Data sources	40
1.16	The SDEC registry	42
1.17	HECM algorithm	44
1.18	Data extracted from the SNDS database	45
1.19	architectures of Word2Vec	48
2.1	Schéma méthodologique du modèle de clustering	66
2.2	Schéma méthodologique du modèle de prédiction	70
3.1	Methodological overview of the clustering study	82
3.2	Visualization of the clusters (derivation cohort)	84
3.3	Main outpatient drugs in each cluster (derivation cohort)	87
3.4	Main hospital diagnoses in each cluster (derivation cohort)	87
3.5	Socio-demographics, SCD characteristics and outcomes of patients in the total derivation cohort and in each cluster	88
3.6	Visualization of the clusters (validation cohort)	89
3.7	Main outpatient drugs in each cluster (validation cohort)	90
3.8	Main hospital diagnoses in each cluster (validation cohort)	90
3.9	Socio-demographics, SCD characteristics and outcomes of patients in the total validation cohort and in each cluster	91
4.1	Methodological overview of the study	99
4.2	AUC curves	102
4.3	Histogram of predicted risks	103
4.4	Importance of the variables	105
4.5	Individual explanations of the prediction model	107
5.1	Comparison of ALA and LA for posterior inclusion probabilities of groups and predictors when n varies	118

5.2	Comparison of ALA and LA for posterior inclusion probabilities of groups and predictors when p varies	119
5.3	Comparison of ALA and LA for run time of waste-free SMC	119
5.4	Groups and predictors selected by the ALA-based SMC sampler	121
5.5	Bi-level variable selection scheme proposed by the ALA-based SMC sampler	122
5.6	Kernel density estimate of the interquartile range (log scale) of the marginal posterior inclusion probabilities (variables) for the ALA-based SMC sampler.	123
A.1	Flow chart of the populations	127
A.2	Medical codes used for baseline characteristics of the populations	131
A.3	Flow chart of the variables	139
A.4	Calibration plots	140

List of Algorithms

1	K-Means algorithm	51
2	CatBoost classification model	54
3	SMC sampler	60
4	Tempering Waste-free SMC	115
5	Independent Metropolis kernel used to move the particles within Algorithm 3 at time t	116

List of Abbreviations

ACLS	Advanced Cardiac Life Support
AI	Artificial Intelligence
ALA	Approximate Laplace Approximation
ARIC	Atherosclerosis Risk in Communities
ARVC	Arrhythmogenic Right Ventricular Cardiomyopathy
ATC	Anatomical Therapeutic Chemical
AUC	Area Under the receiver operating characteristic Curve
BSPP	Brigade des Sapeurs Pompiers de Paris
BrS	Brugada Syndrome
CAD	Coronary Artery Disease
CEMS	Centre d'Expertise de la Mort Subite
CEPIDC	Centre d'Épidémiologie sur les Causes Médicales de Décès
CEREES	Committee of Expertise for Research, Studies and Evaluations in the field of Health
CHD	Coronary Heart Disease
CHS	Cardiovascular Health Study
CNAM	Caisse Nationale d'Assurance Maladie
CNIL	French National Data Protection Agency
CPR	Cardiopulmonary Resuscitation
CPVT	Catecholaminergic Polymorphic Ventricular Tachycardia
CVD	Cardiovascular Diseases
DCIR	Datamart de Consommation Inter Régime
DCM	Dilated Cardiomyopathy
DNA	Deoxyribonucleic Acid
ECG	Electrocardiogram
EHR	Electronic Health Records
EU	European Union
ESS	Effective Sample Size
FHS	Framingham Heart Study
GEHRS	Global Electric Heterogeneity Risk Score
HCM	Hypertrophic Cardiomyopathy
HECM	Healthcare Expenditures and Conditions Mapping
ICD	Implantable Cardioverter-Defibrillator
ICD-10	International Classification of Diseases, 10th revision
LQTS	Long QT Syndrome
LVEF	Left Ventricular Ejection Fraction
MCMC	Monte-Carlo par Chaîne de Markov
NLP	Natural Language Processing
OHCA	Out-Of-Hospital Cardiac Arrests
PMSI	Programme de Médicalisation des Systèmes d'Information
PPV	Positive Predictive Value
PY	Person-Year
RSA	Résumé de Sortie Anonyme

SA	Sinoatrial
SCD	Sudden Cardiac Death
SDEC	Paris Sudden Death Expertise Center
SHAP	SHapley Additive Explanations
SMC	Sequential Monte Carlo
SNDS	Système National des Données de Santé
SNIIRAM	Système National d'Informations Inter-Régimes de l'Assurance Maladie
T-SNE	t-Distributed Stochastic Neighbor Embedding
VF	Ventricular Fibrillation

Chapter 1

General Introduction, Motivations and Contributions

Contents

1.1 Sudden Cardiac Death	17
1.2 Description of the Data	40
1.3 Objective of the Thesis	46
1.4 Summary of Contributions	46

This Chapter provides an overview of the main concepts covered in the thesis. First, Section 1.1 explains the basics of sudden cardiac death, including the physio-pathology and epidemiology of SCD, as well as the current research challenges in this field. Then, Section 1.2 describes the data used in this work. Finally, Section 1.4 gives a summary of the main contributions of this thesis.

1.1 Sudden Cardiac Death

The Cardiovascular System

A broad introduction to the basic anatomy and physiology of the heart, as well as the key functions of the cardiovascular system is first needed for non-medical experts to understand the underlying mechanisms of sudden cardiac death. The human cardiovascular system is a closed circuit network, primarily responsible for distributing oxygen, nutrients, and hormones to the body's tissues and organs, while removing metabolic waste products. It plays a critical role in maintaining a constant internal state in our body, such as blood pressure, pH levels and body temperature. The central component of the cardiovascular system is the heart, a muscular organ located within the thoracic cavity. The heart pumps blood throughout the vascular network, composed of arteries, arterioles, capillaries, venules, and veins (see Figure Figure 1.1):

- Arteries are thick-walled, large vessels that are part of the systemic circuit, and carry blood away from the heart. They subdivide into smaller arterioles that ultimately lead to capillaries. Capillaries are the smallest of blood vessels where gas exchange occurs between the blood and the surrounding tissues. Exchange of nutrients, electrolytes, and metabolic waste products also takes place at the capillary level.
- Venules and veins are part of the pulmonary circuit. They transport deoxygenated blood from the body organs and tissues to the lungs, where it is oxygenated and carbon dioxide is eliminated.

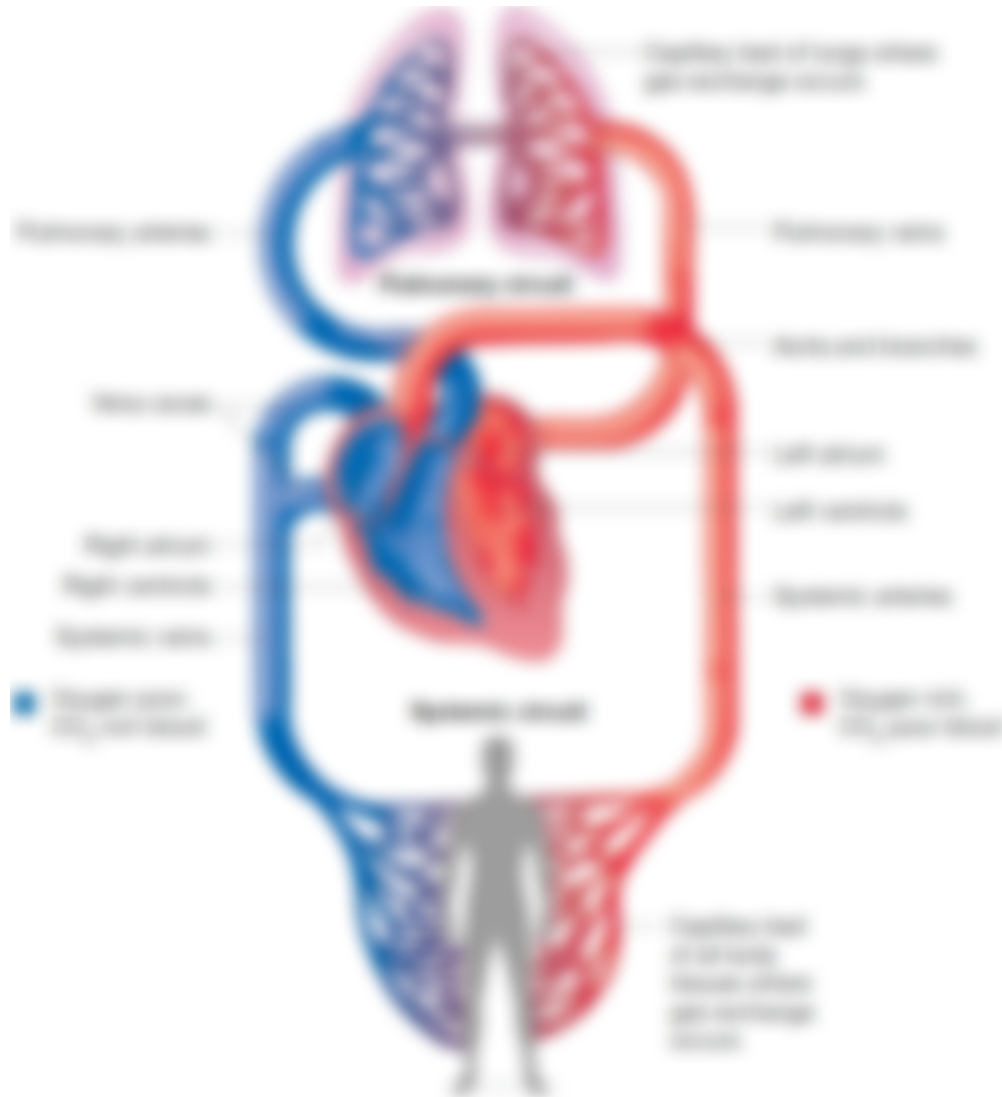


FIGURE 1.1: Scheme of the circulatory system

The pulmonary circulation picks up oxygen from the lungs, and the systemic circulation delivers oxygen to the body

Source: *The Cardiovascular System, Pearson Education*

The heart is composed of four chambers, two atria and two ventricles, which play a unique role in the heart's function (see Figure 1.2). The atria are the two upper chambers and receive blood from the veins. The ventricles are the two lower chambers and pump blood out of the heart. This unique arrangement allows for efficient blood circulation throughout the body, which is a complex but synchronized event:

1. Deoxygenated blood from all the tissues in the body enters the right atrium via two large veins known as the superior vena cava and inferior vena cava.
2. Upon entry, the right atrium contracts, and blood flows through the tricuspid valve into the relaxed right ventricle.
3. Subsequently, the right ventricle contracts, and blood is pumped through the pulmonary valve into the pulmonary artery, which carries it to the lungs for oxygenation.

4. Once the blood is oxygenated, it returns to the heart, entering the left atrium from the pulmonary veins.
5. The left atrium contracts, and the oxygenated blood flows through the mitral valve into the relaxed left ventricle.
6. Finally, when the left ventricle contracts, the blood is pumped through the aortic valve and into the aorta, which carries blood to all parts of the body.

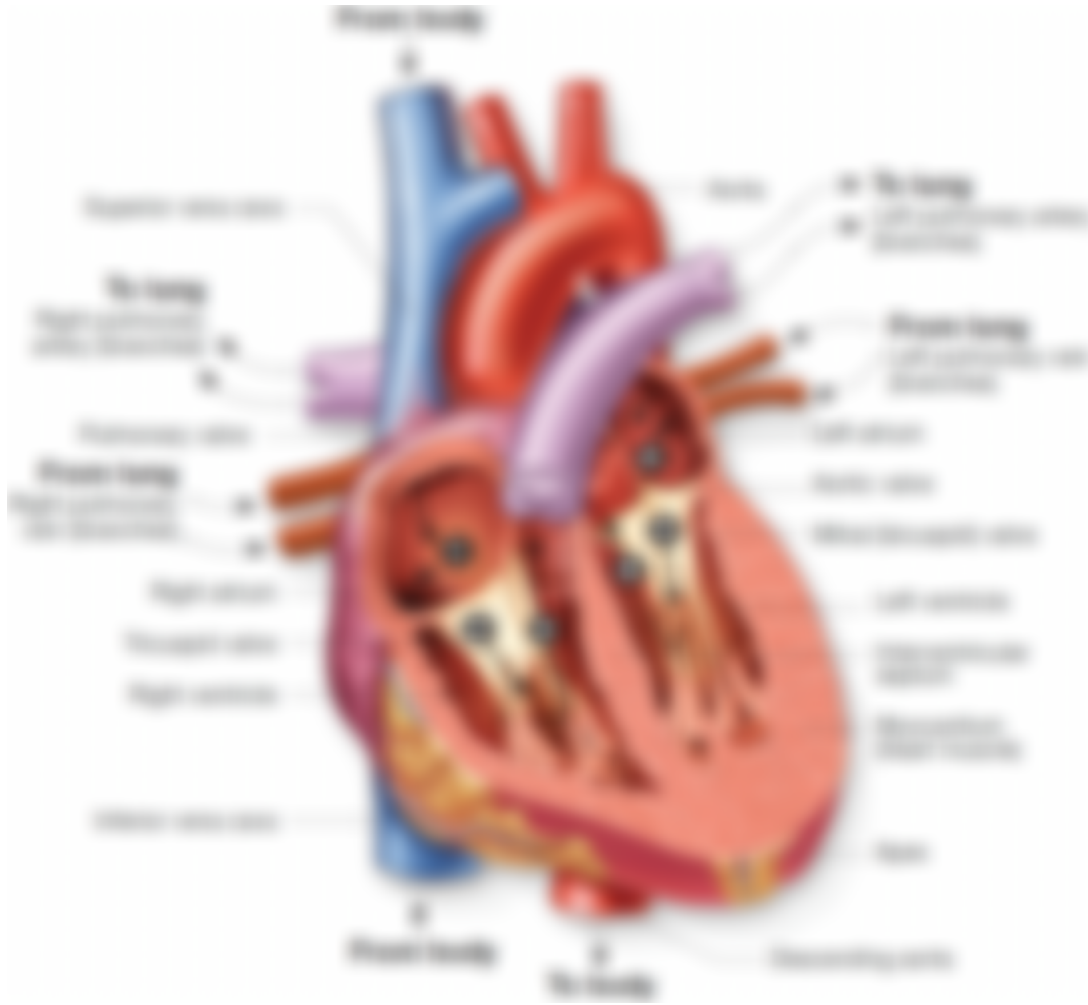


FIGURE 1.2: Circulation of blood through the heart

This process ensures that the heart efficiently supplies oxygen and nutrients to all the tissues in the body.

Source: *The Cardiovascular System, Pearson Education*

The rhythm and timing of the heart's contractions are controlled by the conduction system, which is composed of specialized cells located throughout the heart: the sinoatrial (SA) node, the atrioventricular (AV) node, the bundle of His, and the Purkinje fibers (see Figure 1.3). The SA node, located in the right atrium, serves as the natural pacemaker of the heart. It generates electrical impulses that propagate through the atria and trigger their contractions. This contraction, in turn, facilitates the movement of blood into the ventricles. The impulse then reaches the AV node, located in the lower right atrium, and briefly delays the signal, allowing the ventricles to fill with blood. After the delay, the electrical impulse

travels down the bundle of His, which divides into two branches travelling down the left and right sides of the heart. The branches split into smaller Purkinje fibers, which distribute the impulse throughout the ventricles, causing them to contract and pump blood out of the heart.

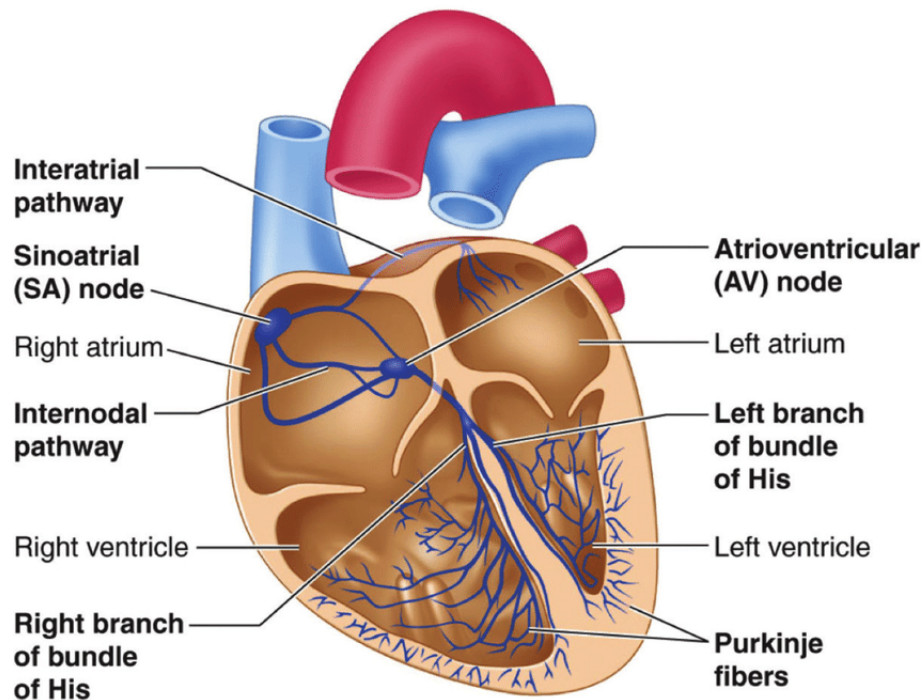


FIGURE 1.3: Electrical conduction system of the heart

The pulmonary circulation picks up oxygen from the lungs, and the systemic circulation delivers oxygen to the body.

Source: Ganesan, P., et al. (2016). *Computer-Aided Clinical Decision Support Systems for Atrial Fibrillation*

The conduction system ensures that the heart's contractions are coordinated and synchronous, preventing any disruption in blood flow to vital organs. Any abnormalities or dysfunction in the conduction system can lead to significant health issues, including arrhythmias, heart block, and heart failure. In the worst cases, these disorders can result in sudden cardiac death.

Sudden Cardiac Death

Definition

Cardiac arrest is a life-threatening medical emergency that occurs when the heart suddenly stops beating efficiently, leading to the cessation of blood flow to vital organs such as the brain and lungs. It is the ultimate mode of any death, regardless of its initial cause. In most cases, it occurs as a terminal complication of a pre-existing condition such as cancer, respiratory failure, or severe infection. However, some of cardiac arrests occur unexpectedly, without any known pre-morbid conditions. When there is no obvious circumstantial cause (such as trauma, suicide, drowning or choking), it is referred to as sudden death, and the underlying cause is presumed to be cardiac. Sudden cardiac death (SCD) is therefore defined as a sudden natural death presumed to be of cardiac cause [Zeppenfeld et al., 2022]. Out-of-hospital cardiac arrest (OHCA) is also often used in the medical literature to refer to SCD. More precisely, two possible definitions of SCD are admitted:

- SCD is certain when it occurs with a sudden collapse witnessed by others, or in the absence of witnesses occurring less than an hour after the onset of the first symptoms.
- SCD is probable when it occurs less than 24 hours after the last contact with the patient.

Mechanism

The underlying mechanism of SCD is usually an abnormal heart rhythm (arrhythmia), that arises from an electrical disturbance in the heart's conduction system. This can result in rapid, irregular, or disorganized electrical activity in the heart that impairs its ability to contract and pump blood effectively. The most common arrhythmias that lead to SCD are ventricular fibrillation (VF) and ventricular tachycardia (VT). VF is characterized by rapid and disorganized electrical activity that results in an erratic and ineffective contraction of the heart muscle (see Figure 1.4), while VT is characterized by a fast and regular heartbeat originating from the lower chambers of the heart. Both of these arrhythmias are considered "shockable" rhythms, as they can be treated with defibrillation, a process in which an electric shock is delivered to the heart to restore its normal rhythm. Prompt defibrillation is crucial in restoring a shockable rhythm and improving survival rates. In this context, public access defibrillation programs have been established in many communities to ensure that bystanders are trained in their use.

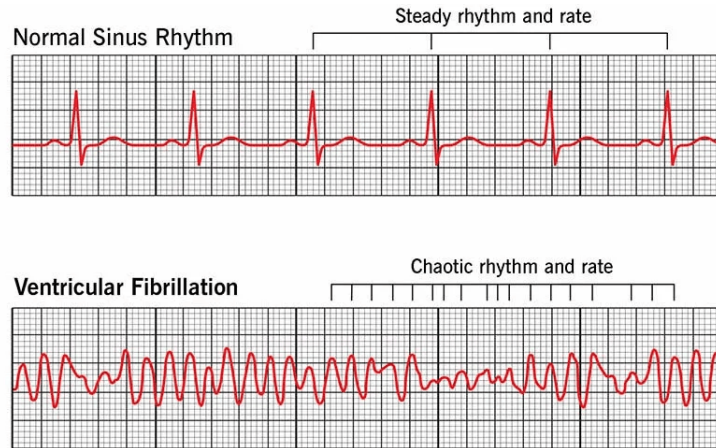


FIGURE 1.4: Ventricular fibrillation
Electrocardiogram of ventricular fibrillation vs normal sinus rhythm.

Source: *Cleveland Clinic*

In contrast, other arrhythmias such as asystole (absence of heartbeat), pulseless electrical activity, and bradyarrhythmias are considered as "non-shockable" rhythms (see Figure 1.5). They do not respond to defibrillation and require other interventions including cardiopulmonary resuscitation (CPR) and advanced life support. The goal of CPR is to provide oxygenated blood to vital organs, such as the brain and heart, until more definitive treatment can be provided. The progression from a shockable to a non-shockable rhythm during SCD can occur due to various reasons, such as delayed or inadequate CPR or an underlying disease process that is not responsive to defibrillation. For instance, VT can progress to pulseless VF and then to asystole if CPR is not initiated or is not successful.

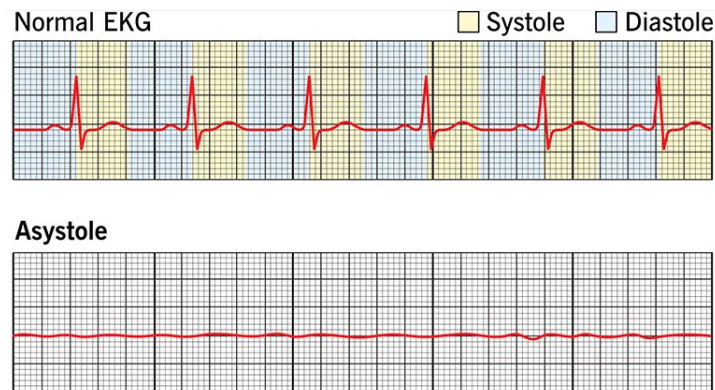


FIGURE 1.5: Asystole
Electrocardiogram of asystole vs normal sinus rhythm.

Source: *Cleveland Clinic*

Underlying cardiac causes


The most common underlying cardiac causes of SCD are coronary artery disease (CAD) and structural heart disease, although other less common causes such as inherited arrhythmia syndromes and ion channelopathies may also be implicated.

CAD, also called ischemic heart diseases, account for approximately 75% of SCD cases [Zeppenfeld et al., 2022]. They occur when the coronary arteries, which supply blood to the heart, become narrowed or blocked due to the buildup of cholesterol (atherosclerosis) and other substances in the artery walls. Over time, this buildup of plaque can restrict blood flow to the heart muscle, leading to a variety of symptoms such as chest pain or discomfort (angina), shortness of breath, fatigue, and weakness. If left untreated, CAD can progress to myocardial infarction (also called heart attack), which occurs when a blood clot forms in a coronary artery and completely cuts off blood flow to a portion of the heart muscle (see Figure 1.6). As a result, that portion of the heart muscle begins to die due to a lack of oxygen and nutrients.

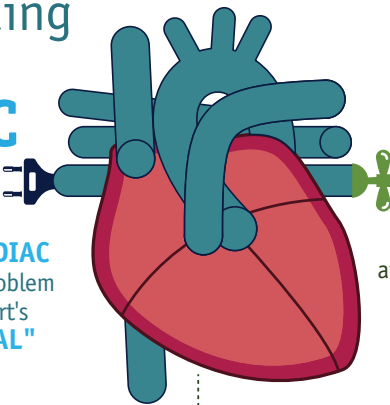
Structural heart diseases such as hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM), and arrhythmogenic right ventricular cardiomyopathy (ARVC) are less common but also important causes of SCD, particularly in younger individuals [Zeppenfeld et al., 2022]. HCM is characterized by thickening of the heart muscle which can impair the heart's ability to pump blood effectively. DCM is characterized by dilation and thinning of the heart chambers, leading to impaired cardiac function and increased susceptibility to arrhythmias. ARVC is a rare inherited condition characterized by replacement of normal heart muscle with fibrous and fatty tissue, leading to arrhythmias and an increased risk of SCD.

Inherited arrhythmia syndromes such as long QT syndrome (LQTS), Brugada syndrome (BrS), and catecholaminergic polymorphic ventricular tachycardia (CPVT) are caused by genetic mutations that disrupt the heart's electrical activity, and also increase the risk of SCD. They are increasingly recognized as important underlying causes of SCD, particularly in young individuals with structurally normal hearts. These conditions can be often diagnosed through specific tests, such as genetic testing and electrocardiogram evaluation.

SUDDEN CARDIAC ARREST



Understanding SUDDEN CARDIAC ARREST



VS. HEART ATTACK

SUDDEN CARDIAC ARREST is a problem with the heart's **"ELECTRICAL"** system

Usually strikes **WITHOUT WARNING**

The heart **SUDDENLY STOPS BEATING**, and no blood is pumped to the rest of the body

People with sudden cardiac arrest **WON'T HAVE A PULSE**

A **HEART ATTACK** affects the **"PLUMBING"** of the heart


People may have **EARLY SIGNS**


BLOOD SUPPLY to the heart muscle is **REDUCED OR BLOCKED**, but the heart **KEEPS BEATING**


People **HAVE A PULSE**, unless the heart attack causes sudden cardiac arrest

Quick Action **SAVES LIVES**

- 1.** Call 911


- 2.** Immediately start CPR, hands only


- 3.** If available, use an automated external defibrillator (AED) to provide an electric shock to the heart, within minutes



SURVIVAL RATES COULD DOUBLE OR TRIPLE if more people **TAKE ACTION AND KNOW** what to do when someone is in sudden cardiac arrest

Sudden cardiac arrest claims **ONE LIFE EVERY 90 SECONDS**

Information provided for educational purposes only. Please consult your health care provider regarding your specific health needs.

For more information, visit [CardioSmart.org/SuddenCardiacArrest](https://www.cardiosmart.org/SuddenCardiacArrest)

If you would like to download or order additional posters on various topics, visit [CardioSmart.org/Posters](https://www.cardiosmart.org/Posters)

FIGURE 1.6: Sudden cardiac death vs heart attack

Source: American College of Cardiology

In the early 1990s, a French cardiologist (Pr. Coumel) proposed a new concept, referred to as “Coumel’s triangle of arrhythmogenesis” [Coumel, 1999], to explain the relationship between the trigger, the substrate, and vulnerability for the development of arrhythmias that can lead to SCD:

- The trigger refers to the acute event that can initiate the arrhythmia. It can be a variety of factors, such as emotional stress, physical exertion, or exposure to certain drugs or toxins. The trigger alone is not sufficient to cause SCD, but it can set the stage for the development of an arrhythmia in a susceptible individual.
- The substrate refers to the underlying cardiac condition or abnormality that can promote the development of arrhythmias. It includes factors such as ischemic heart disease, structural heart disease, or inherited arrhythmia syndromes. The substrate can

make the heart more vulnerable to the effects of the trigger.

- The vulnerability refers to the individual's susceptibility to developing an arrhythmia in response to a trigger and substrate. It can be influenced by various factors, such as age, sex, genetics, and comorbidities.

The Coumel's triangle highlights the complex interplay between factors that lead to SCD, and the importance of considering them to improve preventive strategies. Standardized international tools are also essential to provide a complete understanding of the epidemiology, etiology, and outcomes of SCD. In this context, a set of guidelines for reporting and evaluating OHCA has been proposed. These guidelines, called the Utstein criteria, were developed by an international group of experts in 1991 and have since been updated and refined to improve the consistency and quality of OHCA data [Jacobs et al., 2004]. They outline specific data elements that should be reported for OHCA, including:

- Patient demographics
- Location and time of the event
- Response times of emergency medical services
- Cardiac rhythm at presentation
- Survival outcomes

These criteria also provide guidelines for defining and measuring important variables such as time to defibrillation and quality of CPR [Cummins et al., 1991]. They have facilitated the collection of consistent and reliable OHCA data, allowed for the identification of disparities in survival rates across different regions [Kitamura et al., 2018], and has led to the development of new targeted interventions to improve OHCA management [Gräsner et al., 2016].

Incidence of Sudden Cardiac Death

3 million people are estimated to die from SCD annually, accounting for approximately 15% of all deaths worldwide [Chugh et al., 2008]. The incidence of SCD varies depending on the population studied, but is generally higher in older individuals, men, and those with underlying cardiac disease [Priori et al., 2015]. SCD accounts for approximately 50% of all cardiovascular deaths, with up to 50% being the first manifestation of cardiac disease [Zepfenfeld et al., 2022]. In the Western countries, the epidemiology of SCD is closely related to CAD, which is responsible for up to 75–80% of SCD cases. The incidence also increases markedly with age. It is very low during infancy and childhood (1 per 100,000 person-years (PY)), approximately 50 per 100,000 PY in middle-aged individuals (50-60 years), and at least 200 per 100,000 PY in the eighth decade of life. At any age, males have higher SCD rates compared with females, even after adjustment for risk factors of CAD. A meta-analysis of 27 studies found that the odds ratio for SCD in men compared to women was 2.4 (95% CI, 2.1-2.8) [Chugh et al., 2009]. Although regular physical activity benefits cardiovascular health, sport, particularly when practiced vigorously, has also been shown to be associated with SCD during or shortly after, with an incidence estimated to be around 1 in 50,000 to 1 in 80,000 athletes per year [Chugh et al., 2008].

In the United States, the incidence of SCD is estimated to be approximately 300,000 to 350,000 cases per year [Mozaffarian et al., 2015]. In Europe, 300,000 people have OHCA treated by emergency medical systems every year. Empana et al. [2022] aimed to estimate

the incidence of SCD in the European Union (EU) and to assess the variation in incidence rates between EU countries. The study found that the average annual incidence of SCD in the 4 European registries existing on SCD ranged from 36.8 to 39.7 per 100,000. When extrapolating to each European country and accounting for age and sex, this yields to 249,538 SCD cases per year (see Figure 1.7).

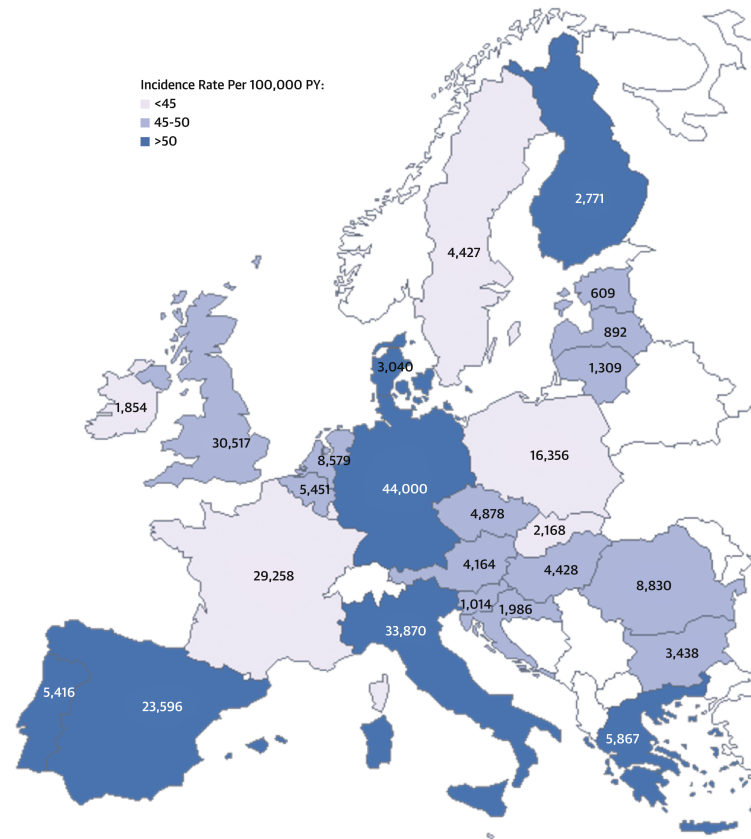


FIGURE 1.7: Average incidence rates of SCD estimated in each European country

Source: *Empana et al. [2022]*

The prognosis of SCD is terrible, with less than 10% surviving to hospital discharge. Survival rates following SCD depend on several factors, including the underlying cause of cardiac arrest, the duration of CPR, and the time to defibrillation. A recent systematic review and meta-analysis of 79 studies found that the overall survival rate to hospital discharge following OHCA was 8.3% [Gräsner et al., 2016], with higher rates observed in patients with shockable rhythms (29.8%) compared to non-shockable rhythms (4.6%). Another study reported that survival rates following OHCA varied widely depending on the location of cardiac arrest, with higher survival rates observed in urban areas and in patients who received bystander CPR [Nichol et al., 2008]. The use of automated external defibrillators and advanced cardiac life support protocols have also been shown to improve survival rates [Meaney et al., 2013].

Risk Factors for Sudden Cardiac Death

A broad range of risk factors for SCD have been described in the medical literature. Most of them are common to other cardiovascular diseases, such as advancing age, male sex,

smoking, diabetes and obesity. Indeed, these risk factors have been shown to increase the risk of atherosclerosis, which in turn can increase the risk of SCD. So far, only a few specific risk factors for SCD in the general population have been identified, including family history and heart rate at rest and during exercise.

Family history

Jouven et al. [1999] aimed to investigate whether a family history of SCD was associated with an increased risk of SCD in the general population. The study included a total of 5,243 individuals aged 45-64 years who participated in a prospective French cohort. They found that individuals with a family history of SCD had a significantly increased risk of SCD compared to those without a family history. Specifically, the risk of SCD was 2.4 times higher in individuals with a first-degree relative who had died of SCD compared to those without a family history. They also found that the association between family history and SCD was stronger in individuals with no prior history of cardiovascular disease. A meta-analysis of 10 case-control studies [Winkel et al., 2011] found that a family history of SCD was associated with a significantly increased risk of SCD, with an odds ratio of 1.9 (95% CI: 1.5-2.5). Chugh et al. [2004] found that the risk of SCD was increased in individuals with a family history of premature CAD, which may be a marker of genetic predisposition to heart disease. Finally, a family history of inherited cardiac disorders such as HCM, ARVC and LQTS has been associated with an increased risk of SCD [Ackerman et al., 2011].

Heart rate

Elevated heart rate at rest and during exercise have also been identified as specific risk factors for SCD. Jouven et al. [2005] investigated the association between heart rate and the risk of SCD in a population-based cohort of 5,713 men and women who were 42 to 53 years old at baseline. The study found that higher resting heart rate was associated with an increased risk of SCD, even after adjusting for other risk factors such as smoking, blood pressure, and cholesterol levels. Specifically, each increase in heart rate by 10 beats per minute (bpm) was associated with a 1.2-fold increase in the risk of SCD in men and a 1.6-fold increase in women. The study also found that the association between heart rate and SCD was stronger among individuals with no history of cardiovascular disease at baseline. Similarly, an elevated heart rate during exercise has been associated with an increased risk of SCD. A study of over 3,000 men found that individuals with an exercise-induced heart rate of greater than 150 bpm had a significantly increased risk of SCD compared to those with a heart rate of less than 120 bpm [Albert et al., 2000].

Genetic factors

Several results have highlighted the role of genetic factors in the pathogenesis of SCD. Inherited cardiac disorders such as HCM, ARVC and LQTS are well-established risk factors for SCD. These disorders are caused by mutations in genes encoding proteins involved in the structure and function of the heart, leading to abnormalities in the heart's electrical system and an increased risk of arrhythmias [Maron, 2009, Basso et al., 2009]. In addition to rare mutations causing inherited cardiac disorders, several common genetic variants have been identified as risk factors for SCD. These variants may influence the structure and function of the heart or its electrical properties, leading to an increased risk of arrhythmias. Genome-wide association studies have also identified loci (positions on a chromosome where a gene or Deoxyribonucleic Acid (DNA) sequence is located) associated with an increased risk of SCD, including variants near genes encoding ion channels and structural proteins of the heart [Bezzina et al., 2010]. Finally, epigenetic modifications, such as DNA methylation and

histone modification, can influence gene expression and contribute to the pathogenesis of SCD.

A summary of age distribution when SCD occurs, and its association with gender, dominant arrhythmia subtypes, triggers and genetic factors are presented in Figure 1.8.

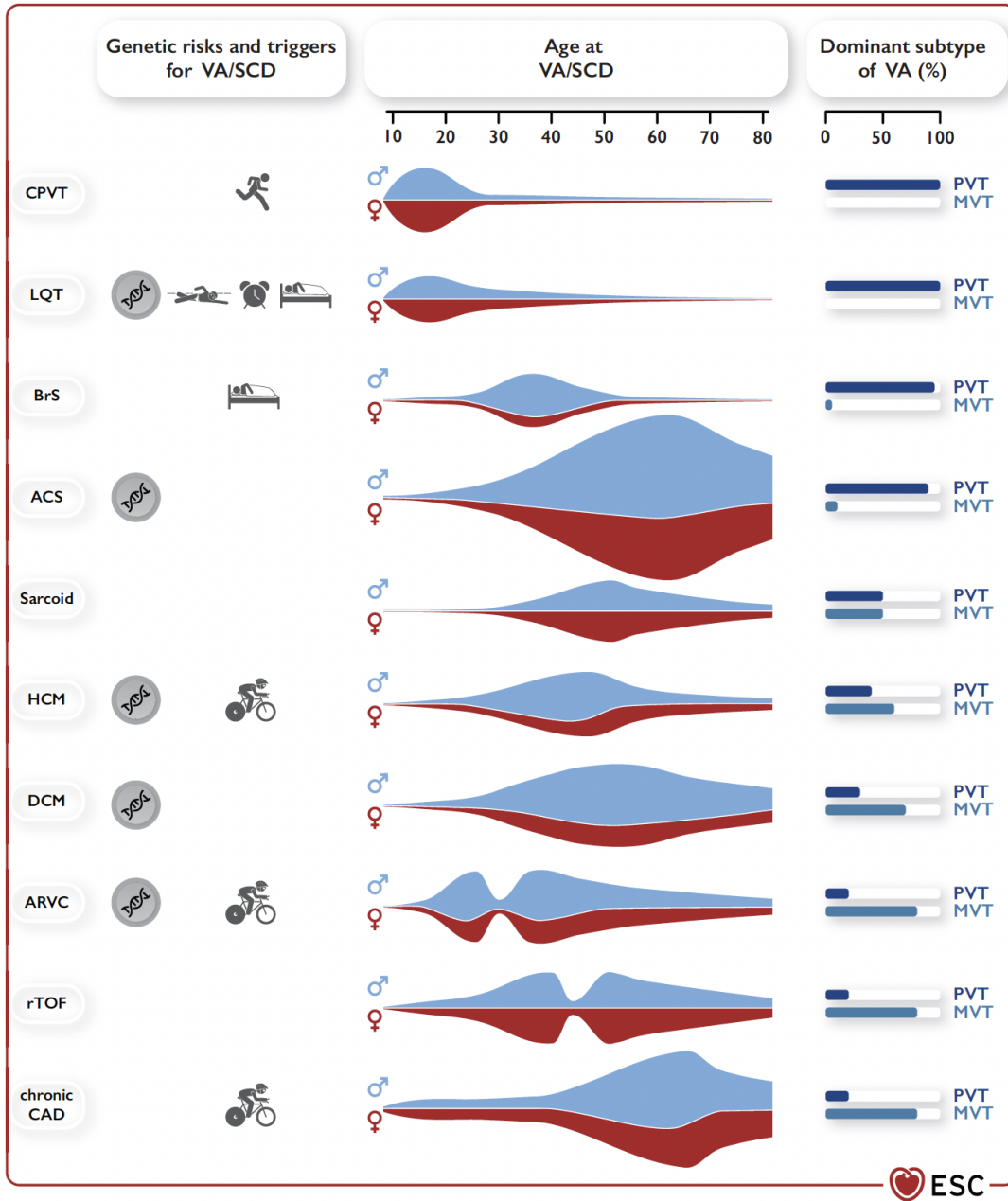


FIGURE 1.8: Risk factors for SCD

Source: Zeppenfeld et al. [2022]

Reduced left ventricular ejection fraction

Reduced left ventricular ejection fraction (LVEF) is one of the most established risk factor for SCD. The LVEF is a quantitative measure of the heart's ability to pump blood out of the left ventricle with each contraction (see Figure 1.9). A reduced LVEF indicates impaired cardiac function, which can be a sign of heart disease or other cardiac conditions. Several studies have investigated the role of depressed LVEF in predicting SCD, and this association is particularly strong in patients with heart failure, where a reduced LVEF is a key feature of the disease.

Moss et al. [2002] evaluated the use of LVEF to predict SCD in patients with coronary artery disease. The study found that patients with an LVEF less than 30% had a significantly higher risk of SCD compared to those with an LVEF of 30% or greater. Similarly, Pocock et al. [2006] evaluated the use of LVEF in patients with heart failure. They found that patients with an LVEF less than 35% had a significantly higher risk of SCD compared to those with an LVEF of 35% or greater.

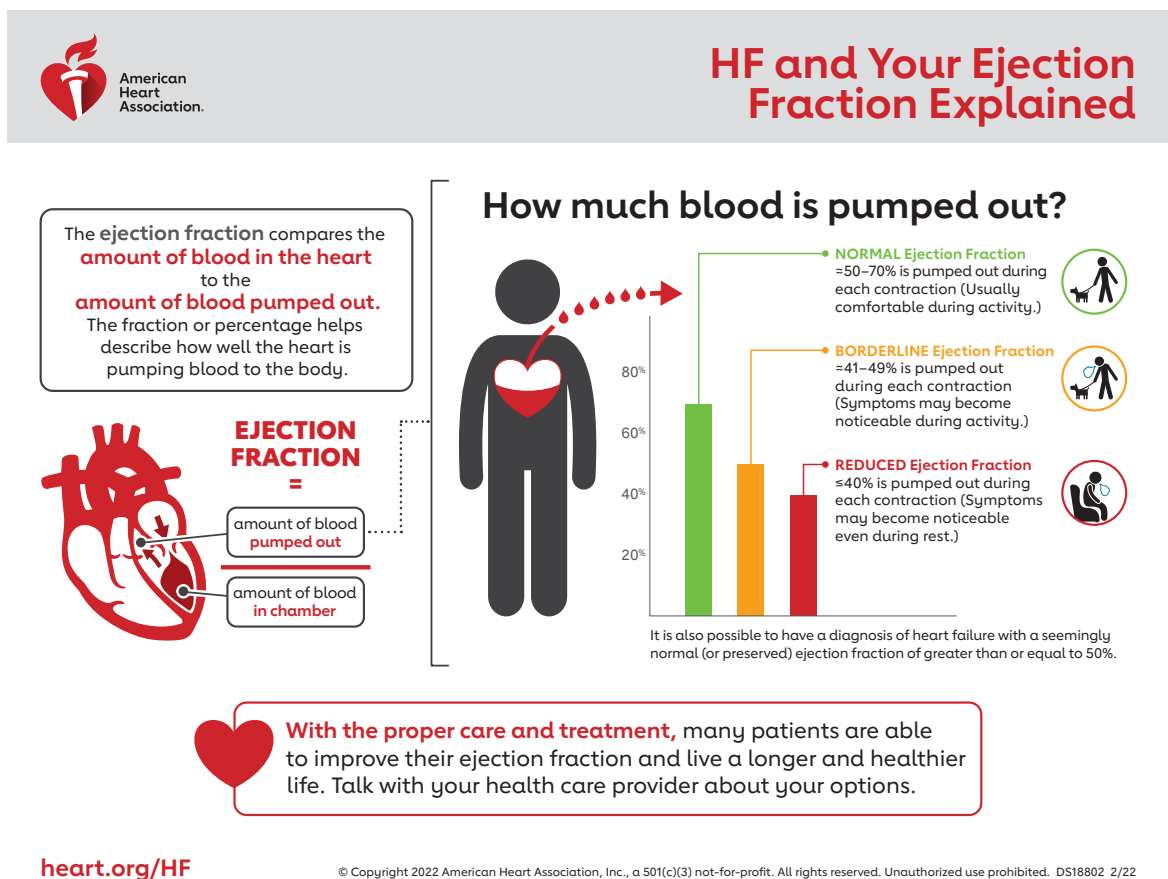


FIGURE 1.9: The role of ejection fraction

Source: *The American Heart Association*

Current medical guidelines recommend therefore the use of LVEF as a key parameter in the decision-making process as a primary criterion for implantable cardio-verter-defibrillator (ICD) therapy in patients with chronic CAD and DCM. These devices are implanted subcutaneously and are designed to detect and terminate potentially fatal arrhythmias through the delivery of an electrical shock (see Figure 1.10).

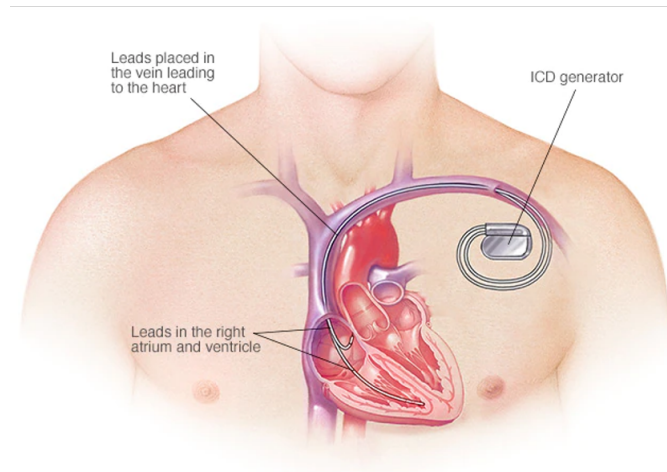


FIGURE 1.10: Implantable cardioverter-defibrillator

Source: *Mayo foundation for medical foundation and research*

The American College of Cardiology/American Heart Association guidelines recommend ICD therapy for primary prevention in patients with $LVEF \leq 35\%$ [Kusumoto et al., 2019]. However, the use of LVEF as a unique criterion for ICD therapy has several limitations. Indeed, LVEF is a static measurement that may not reflect changes in myocardial function over time, and may not accurately predict the risk of SCD in all patients with reduced LVEF. In addition, recent studies have suggested that the standard threshold of 35% in determining eligibility for ICD therapy may be not optimal. Narayanan et al. [2013] found that only 20% had a prior indication for implantation of a defibrillator, meaning that the majority of patients would not have been identified as candidates based on current guidelines. LVEF should be therefore used in conjunction with other relevant clinical parameters.

Risk stratification

Myerburg et al. [1992] proposed a conceptual model for risk stratification to highlight the complex relation between the numerous SCD risk factors. The Myerburg pyramid is divided into four levels, each representing an increasing level of risk for SCD:

1. The first level of the pyramid includes population-based risk factors that are prevalent in the general population and contribute to an increased risk of SCD. These risk factors include age, gender, family history of SCD, race/ethnicity, and socioeconomic status. They cannot be modified but they help identify individuals who may be at increased risk for SCD.
2. The second level includes clinical risk factors that are identifiable through medical history, physical examination, and diagnostic tests. These risk factors include underlying cardiac diseases such as CAD, heart failure, valvular heart disease, and genetic cardiac disorders. Other factors that increase the risk of SCD include hypertension, diabetes, smoking, and high cholesterol. Identification of these risk factors allows for targeted interventions to prevent SCD.
3. The third level includes inducible arrhythmia risk factors that are identified through electrophysiological testing. These tests include exercise stress testing, electrocardiogram monitoring, and electrophysiology studies. The presence of inducible arrhythmias during these tests increases the risk of SCD and may indicate the need for ICD placement or other interventions.

4. The fourth and final level of the pyramid includes symptoms and events that may indicate an increased risk of SCD. These include syncope (fainting), palpitations, and cardiac arrest. Identification of these symptoms and events is critical as they may indicate the need for immediate intervention to prevent SCD.

The Myerburg pyramid emphasizes the need for a more comprehensive approach to SCD risk assessment and prevention, that takes into account multiple risk factors at various stages of the pyramid.

Management and Prevention of Sudden Cardiac Death

Several tools have been developed to improve the prognosis of SCD, with particular focus on enhancing prehospital management through the chain of survival (see Figure 1.11). This concept emphasizes the importance of rapid intervention and coordinated care to improve the chances of survival in SCD patients. It comprises four key links:

- Early recognition and activation of the emergency medical services
- Early CPR
- Prompt defibrillation
- Effective post-resuscitation care

Each link is essential and can significantly impact the survival and neurological outcomes of patients.

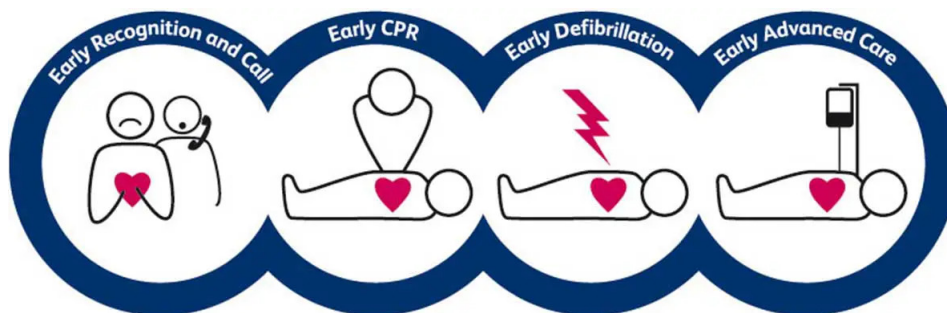


FIGURE 1.11: The chain of survival

In addition to prehospital interventions, effective hospital management is also critical for improving the survival and neurological outcomes of SCD. Early coronary care, including percutaneous coronary intervention and revascularization, can significantly improve survival rates in patients with acute coronary syndromes.

Predicting sudden cardiac death in the population

Despite these interventions, survival rates after SCD remain low, with reported rates ranging from 5% to 20% depending on the patient population. Preventive strategies are therefore needed to identify individuals who are the most at risk of SCD. This population could benefit from antiarrhythmic treatments and ICDs. However, ICDs also carry risks and complications, and their use should be guided by careful patient selection and appropriate follow-up. In this context, prediction of SCD is still a major research challenge, with disappointing results so far. Preventive strategies have mainly focused on using ICD in the highest risk subgroups of the population, such as those with an advanced cardiomyopathy and depressed

LVEF [Moss et al., 2002]. While preventing the onset of SCD in these populations is clinically relevant, the impact on the number of preventable cardiac arrests is small since SCD in high-risk populations represent a small proportion of the total SCD. Most cases, indeed, occur in the general population, with no clinically recognized heart disease prior to the event.

Myerburg [2001] described the inverse relationship between the incidence of SCD and the absolute numbers of events in the various epidemiological and clinical categories (see Figure 1.12). The nominal incidence of SCD increases from the general population aged older than 35 years to specific high-risk post-myocardial infarction patients, while the associated risk accounts for a decreasing absolute number of events annually. Improving prediction in the general population may therefore have a real impact on the total burden of SCD.

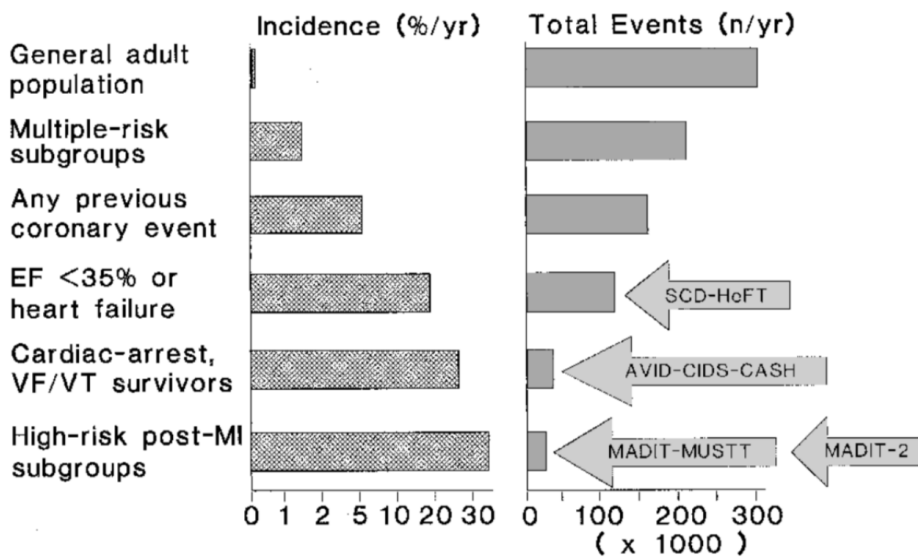


FIGURE 1.12: Risk stratification of SCD

Source: Myerburg [2001]

Litterature Review of Prediction Models for Sudden Cardiac Death

Several prediction models for SCD in the general population have been proposed. Most of them use community-based cohorts and consist in estimating the 10 year-risk of SCD according to risk factors measured at study entry. We conducted a litterature review on Pubmed on prediction models of SCD. We searched for a combination of keywords related to SCD and prediction models between 01/01/1976 and 01/03/2023. 1,948 related citations were identified. After screening titles, abstracts, and results of articles, 20 studies were considered eligible as prediction models. Of those, 16 models predict SCD only few minutes or hours before the onset, and were excluded, resulting in 5 eligible studies [Deo et al., 2016, Waks et al., 2016, Aro et al., 2017, Bogle et al., 2018, Holkeri et al., 2020].

1. Deo et al. [2016] used data from the Atherosclerosis Risk in Communities (ARIC) study (13,677 participants including 171 SCD cases), a prospective cohort study that followed a large sample of individuals from four communities in the United States since 1985. The authors identified a set of risk factors for SCD, including age, sex, race, smoking status, body mass index, diabetes, hypertension, prior myocardial infarction, heart rate, and QT interval. These risk factors were used to develop a 10-year prediction model for SCD based on a competing risk regression analysis. The final

prediction model included five risk factors: age, sex, race, smoking status, and QT interval, with a C-statistic (which measures the discrimination performance) of 0.82 in the ARIC study cohort and 0.74 in an external validation cohort from the Cardiovascular Health Study (CHS) (4,207 participants including 174 SCD cases).

2. Waks et al. [2016] aimed at developing and validating a 10-year risk score, called the Global Electric Heterogeneity Risk Score (GEHRS), for predicting SCD in the general population. The study used data from the ARIC study (14,609 participants including 291 SCD cases) and the CHS (5,568 participants including 195 SCD cases). The GEHRS is based on a non-invasive measurement of ventricular repolarization heterogeneity, which is a marker of electrical instability in the heart that has been shown to be associated with SCD risk. It includes five independent predictors of SCD: QRS duration, spatial QRS-T angle, heart rate, sex, and serum potassium levels. The GEHRS score was able to accurately identify individuals at high risk of SCD in both the ARIC and CHS cohorts (C-statistic = 0.79). It outperformed traditional risk factors, such as age, sex, smoking status, and history of cardiovascular disease, and remained a strong predictor even after adjusting for these variables.
3. Aro et al. [2017] aimed at developing and validating a 10-year risk score that could predict SCD beyond LVEF. The study analyzed data from the Oregon Sudden Unexpected Death Study (1,258 participants including 522 SCD cases) and the ARIC study (3,567 participants including 260 SCD cases). The electrical risk score included four electrocardiographic markers: QRS duration, QTc interval, Tpeak-Tend interval, and spatial QRS-T angle. The electrical risk score was found to be a strong predictor of SCD, even after adjusting for other clinical and demographic factors, with a C-statistic of 0.75 for the Oregon study and 0.77 for the ARIC study.
4. Bogle et al. [2018] used data from the ARIC study (11,335 participants including 145 SCD cases) and the Framingham Heart Study (FHS) (5,626 participants including 64 SCD cases) to identify potential predictors of SCD. They developed a 10-year risk score which is based on age, sex, cholesterol, lipid-lowering medication use, hypertension, systolic and diastolic blood pressures, smoking status, diabetes and body mass index. The C-statistic of the final model was 0.82 in white ARIC participants, 0.75 in black ARIC participants, and 0.82 in white FHS participants.
5. Holkeri et al. [2020] aimed to develop and validate an electrocardiographic 10-year risk score for predicting SCD from two Finnish population-based cohorts: The Mini-Finland Health Survey (6,830 participants including 123 SCD cases) and the Coronary Heart Disease (CHD) Study (10,617 participants including 115 SCD cases). The study identified ECG parameters that were associated with an increased risk of SCD, including prolonged QT interval, fragmented QRS complex, and abnormal T-wave morphology. The study then developed an ECG risk score that incorporated these parameters and assessed its predictive ability for SCD. They found that the ECG risk score was able to accurately predict SCD in both cohorts. In the Mini-Finland Health Survey cohort, the risk score had a C-statistic of 0.86, while in the CHD study, the risk score had a C-statistic of 0.89.

These prediction models demonstrated excellent discrimination capacities, with C-statistic ranging between 0.74 and 0.89. However, several limitations hinder their broad applicability

in the general population. Although these studies have been derived from large community-based cohorts, populations of controls are often not randomly selected in the general population, and the number of SCD cases is often limited, typically ranging between 100 to 500 subjects, which may limit their statistical power and generalizability. Additionally, these models have relied on clinical variables that can be challenging to collect, such as electrocardiogram signals, or not routinely measured, such as potassium, serum albumin, or glomerular filtration rate. Furthermore, these models typically consider risk factors measured at a single time point and do not integrate trajectories of risk factors and treatments over time, which can lead to inaccuracies and misclassification of risk. Finally, these models lack specificity since they predict equally well SCD and acute coronary syndrome. One possible explanation is that most risk factors considered in these models are related only to the development of atherosclerosis and not specific to the susceptibility of arrhythmias. New preventive strategies are therefore needed in the field to improve prediction of SCD in the general population.

The Challenge of Artificial Intelligence in Cardiology

Artificial intelligence (AI) has recently emerged as a powerful tool for transforming health-care and medicine by enabling more accurate diagnosis, personalized treatment, and disease prevention [Rajpurkar et al., 2022]. The use of AI in healthcare is rapidly evolving, with significant potential for improving patient outcomes and reducing healthcare costs. New statistical approaches based on machine learning and deep learning, including natural language processing and computer vision, can now process vast amounts of data and generate insights that were previously difficult or impossible to obtain so far (see Figure 1.13). The use of AI could be therefore a promising avenue towards developing new preventive strategies for SCD.

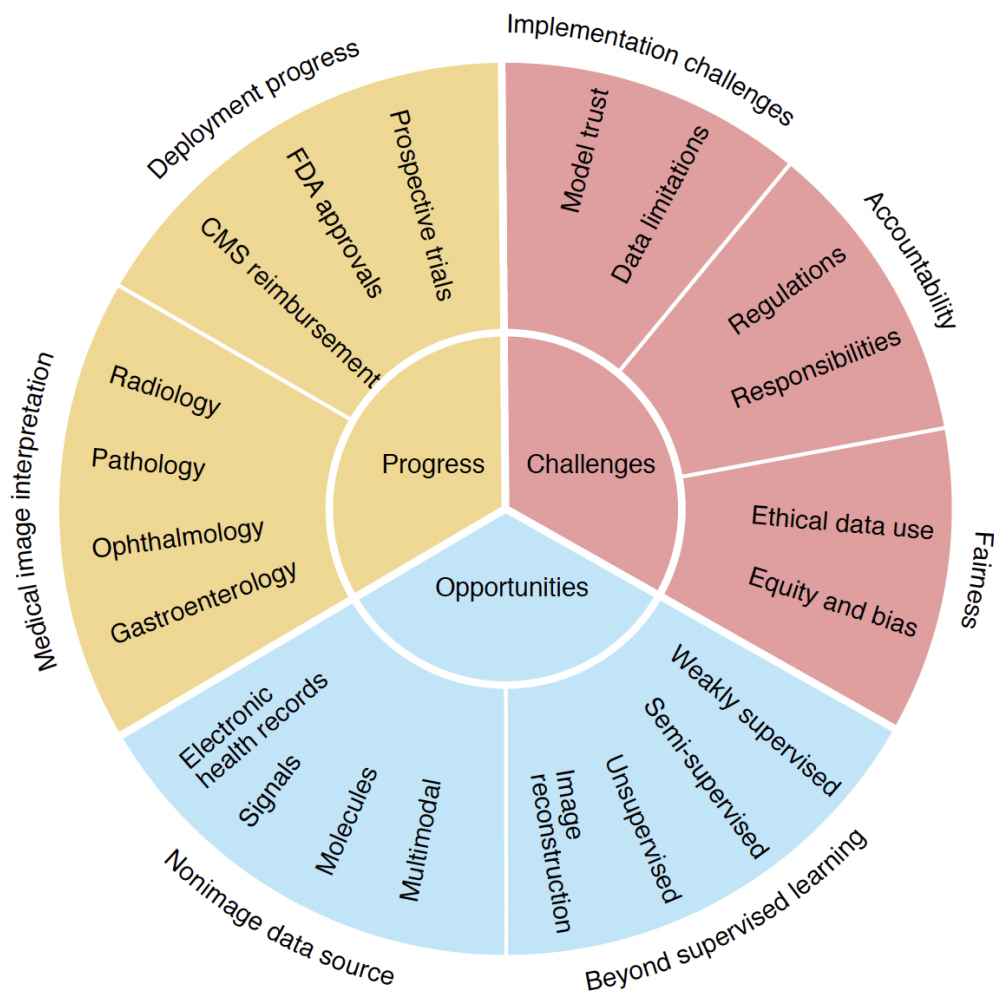


FIGURE 1.13: AI in health
Overview of the progress, challenges and opportunities for AI in health

Source: Rajpurkar et al. [2022]

AI demonstrated particular promise in medical imaging, with algorithms being trained to detect and classify abnormalities in specialties that rely heavily on the interpretation of images, such as radiology, pathology, gastroenterology and ophthalmology. For instance, these models make accurate survival predictions for a wide range of cancer types compared

to conventional histopathological subtyping [Rajpurkar et al., 2022].

AI also offers promising opportunities to expand the limit of our knowledge in cardiovascular diseases. Krittanawong et al. [2020] performed a meta-analysis of 82 studies, including over 200,000 patients, to assess the performance of machine learning and deep learning models in predicting cardiovascular diseases. The results revealed that they demonstrated superior predictive performance compared to traditional models in predicting cardiovascular diseases, including CAD, heart failure, and atrial fibrillation. Notably, these approaches have achieved considerable improvements in the analysis of electrocardiogram signals to predict SCD few minutes before the onset [Houshyarifar and Chehel Amirani, 2016, Ebrahimzadeh et al., 2019, Amezquita-Sanchez et al., 2018]. Multiple studies used AI to predict cardiac arrest in high-risk populations including patients suffering from heart failure [Meng et al., 2019], critically-ill patients admitted to the emergency department or intensive care units [Jang et al., 2020, Kim et al., 2019], or focused on intra hospital cardiac arrest [Kwon et al., 2018]. However, so far, no studies have examined if machine learning can enhance risk prediction over clinical risk models for SCD in the general population. This question remains an important challenge that we aim to address in this thesis.

Among current statistical learning methods, natural language processing (NLP) techniques have advanced significantly in recent years, enabling the exploitation of vast medical text databases and electronic health records (EHR) beyond standard predictive approaches. New models based on contextual word embeddings have indeed improved the ability to consider the surrounding context when analyzing complex medical information. They achieve remarkable success across a wide range of tasks, such as named entity recognition, sentence classification, or question answering (see Figure 1.14). Notable examples are the transformers-based models neural networks, which allow to handle long-range dependencies between words, and have become the state-of-the-art in many NLP benchmarks.

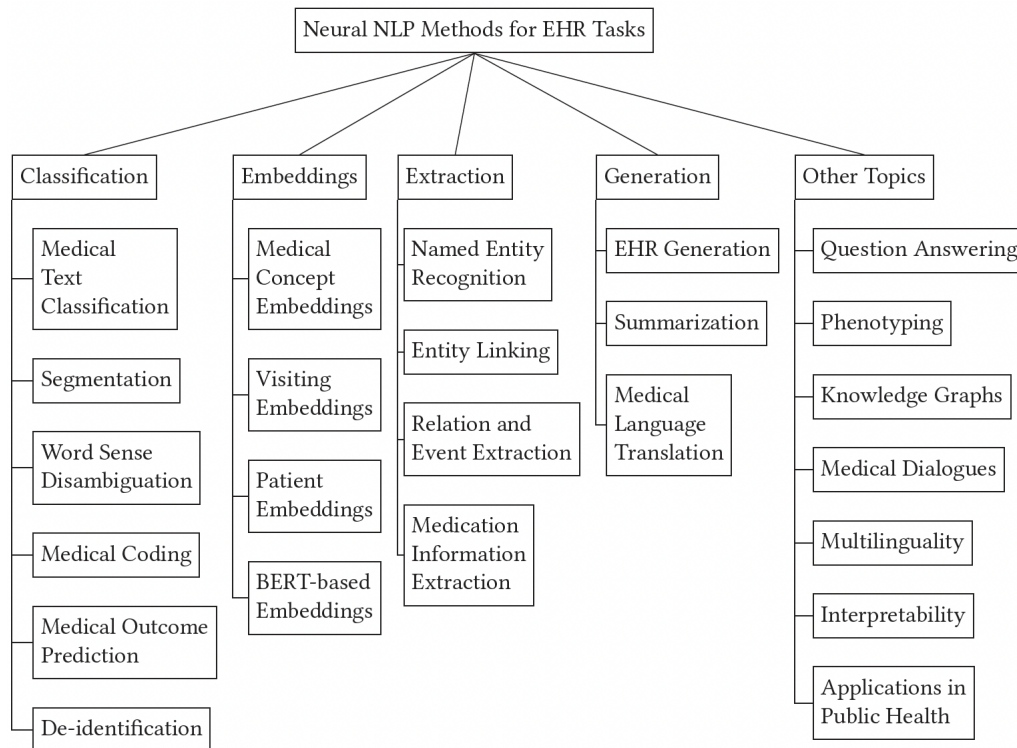


FIGURE 1.14: Various tasks covered by NLP in medicine

Source: *Li et al. [2022]*

Word embeddings methods can be used to improve current clinical prediction models. For instance, Choi et al. [2017] applied NLP algorithms to develop an early prediction model for heart failure. The authors trained a neural network to represent patients by mapping hospital diagnoses and outpatient drugs collected in their medical history into a continuous vector space. This medical embedding allowed them to capture complex relationships between diagnosis codes, that may have been difficult to capture using traditional methods. By processing the embeddings over time, the model was able to identify relevant patterns and trends to predict the onset of heart failure. In this thesis, we explore such approaches, to assess the extent to which it could provide a better understanding of SCD.

These methods also highlight the potential value of electronic health records, and the current paradigm shift towards incorporating new data sources into standard epidemiological approaches. EHR routinely collect data from millions of patients across diverse healthcare institutions, including demographic information, diagnoses, laboratory test results, medication, prescriptions, clinical notes, and medical images. EHR have changed the data analytic modeling paradigm for many biomedical applications, compared with standard data sources such as cohort studies or randomised controlled trials. Indeed, they reduce administrative efforts and costs to collect data. They are also more representative of the total target population and less subject to inclusion bias than randomised controlled trials, because they are obtained from all individuals who interact with health systems. Over the past few years, an increasing body of literature confirmed the success of epidemiological models derived from large EHR databases [Xiao et al., 2018]. Notable examples of such data sources include:

- The Medical Information Mart for Intensive Care III (MIMIC-III): a publicly available database of de-identified EHR data from over 40,000 patients who were admitted to

critical care units at Beth Israel Deaconess Medical Center in Boston (United States).

- The UK Biobank: a large-scale prospective cohort study that includes genetic and EHR data from over 500,000 participants in the United Kingdom.
- The SNDS database (Système National des Données de Santé): the database of the French Universal Health Insurance System, which manages all reimbursements of healthcare for all people affiliated to a health insurance scheme in France, resulting in one of the largest EHR databases in the world. This database is used in this thesis, and is described in detail in Section 1.2.

1.2 Description of the Data

This thesis is based on a large retrospective case-control investigation, using two main data sources: a unique population-based registry on SCD, and a large database of electronic health records (see Figure 1.15). This section describes in detail each data source.

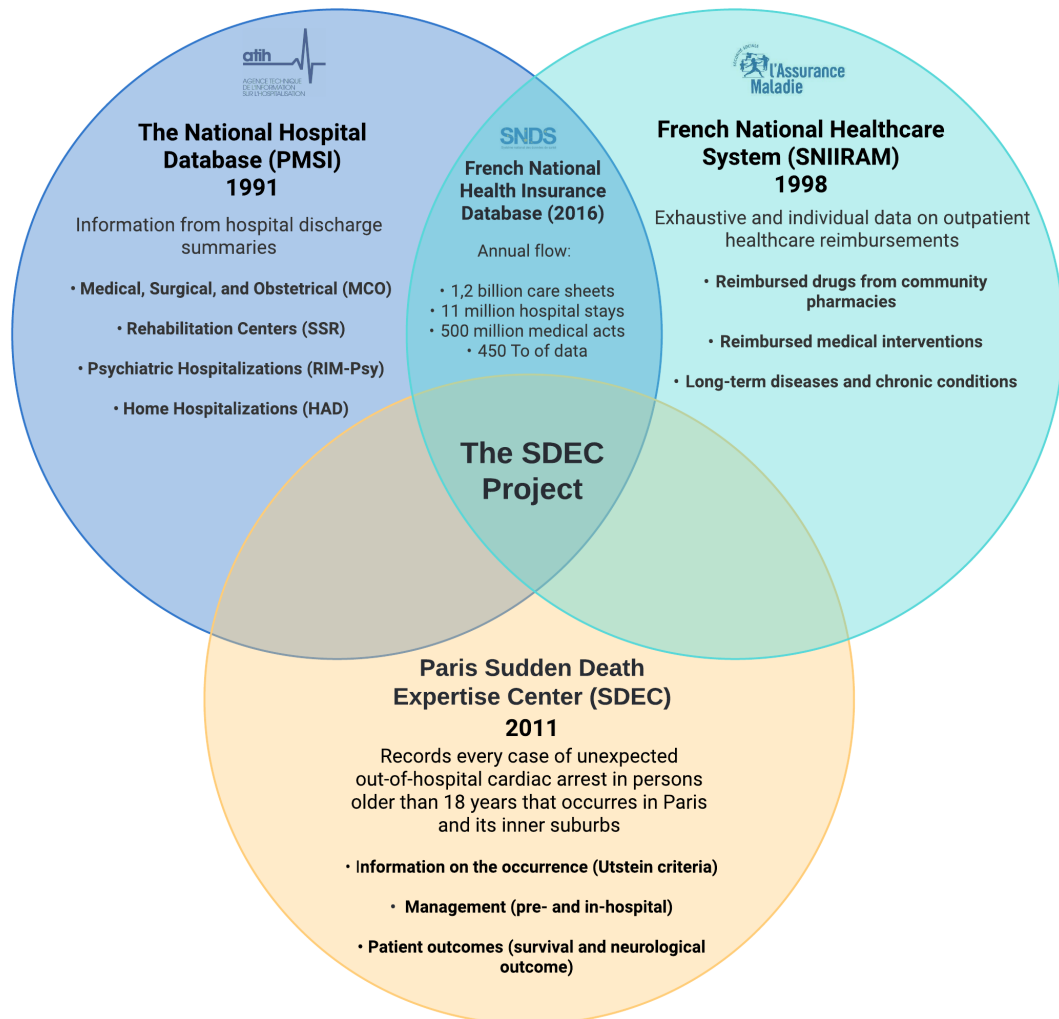


FIGURE 1.15: Data sources

The Paris Sudden Death Expertise Center

The Paris Sudden Death Expertise Center (SDEC) registry is a multicenter population-based registry system which collects every case of unexpected OHCA occurring among adults (aged 18 years and older) in Paris (France) and its inner suburbs (Hauts-de-Seine, Seine-Saint-Denis, Val-de Marne) since 16 May 2011, covering a population of 6.7 million inhabitants (10% of the French population) [Bougouin et al., 2014]. It records prospectively and continuously information on the occurrence (Utstein criteria), management (pre- and in-hospital) and patient outcomes (regarding survival and neurological outcomes) of all cases. Exclusion criteria are a prior terminal condition, no attempt at advanced cardiac life support by the emergency medical service personnel, or an obvious non cardiac cause according to the Utstein templates. The SDEC registry therefore includes only cases who experienced SCD.

SCD cases included in the SDEC registry are managed by the emergency medical service, which is composed of a two-tiered physician-manned system. The first tier is known as the basic life support tier and is staffed by the Brigade des Sapeurs Pompiers de Paris (BSPP). They provide essential first-aid treatment, including stabilizing patients and transporting them to the appropriate medical facility. The second tier is the advanced cardiac life support (ACLS) tier, which is staffed by physicians who are specially trained in advanced medical techniques. They are equipped with the knowledge and technology necessary to provide rapid and effective treatment in critical situations, such as cardiac arrest, stroke, and severe trauma. The SDEC registry is derived with the following procedure:

1. First, a nominative case report form is sent daily for every cardiac arrest supported by BSPP.
2. Second, an electronic query algorithm is performed in the ACLS computer system to identify every case of SCD.
3. Third, retrospective controls based on diagnostic codes are conducted in selected intensive care units.

This method therefore involves every link of the chain of survival, to ensure completeness of the registry. (Bougouin et al. [2014]) performed a retrospective control among a sample of 3 intensive care units, and combination of both sources (BSPP and ACLS) detected 99% of cases of cardiac arrests admitted alive in this sample. In addition, each case is reviewed separately by two investigators of the SDEC, to ensure accuracy of classification and to avoid the over-estimation often experienced in retrospective collection.

The SDEC registry has been described in multiple studies [Bougouin et al., 2014, Maupain et al., 2016, Jabre et al., 2016, Bougouin et al., 2018, 2020]. For instance, Marijon et al. [2020] used it during the COVID-19 pandemic as a real-time multisource surveillance system set up to assess the incidence and outcomes of OHCA. The study found that the maximum weekly OHCA incidence during the pandemic period increased from 13.4 to 26.6 per million inhabitants, compared to the same weeks in the non-pandemic period, with a survival rate to hospital admission reduced from 22.8% to 12.8%. The results demonstrated therefore a major rise in OHCA-related deaths during the pandemic period (see Figure 1.16).

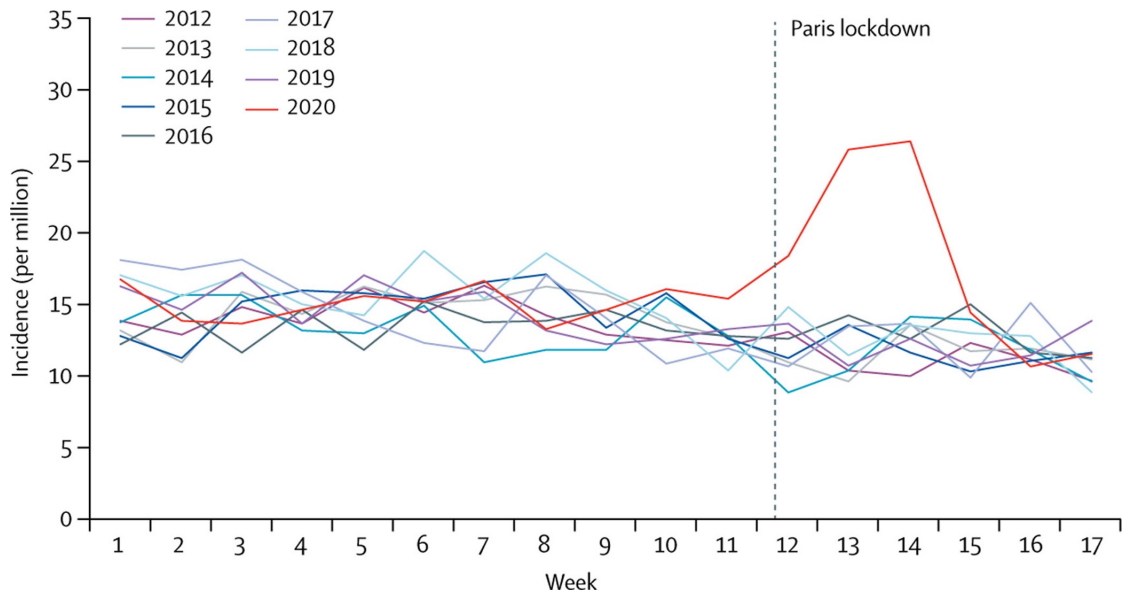


FIGURE 1.16: The SDEC registry
Weekly incidences of OHCA during the first 17 weeks of years 2012 to 2020.

Source: Marijon et al. [2020]

The French National Healthcare Insurance Database

The French healthcare system was established in 1945 and is defined by a mix of public and private healthcare providers. It ensures a comprehensive healthcare coverage to all individuals through a public health insurance scheme, called Social Security. The system is funded through payroll taxes, contributions from employers and employees, and government subsidies. To ensure efficient reimbursement processes, a dedicated national database was established and collects individual data on all healthcare expenses for all people affiliated to a health insurance scheme, covering 98% of the French population (67 million inhabitants). It includes:

- Outpatient visits, procedure, and reimbursed drugs relative to outpatient medical care claims.
- Information from hospital discharge summaries.
- Chronic conditions.
- Paramedical activities such as nursing or physiotherapy, lab tests, and devices.

This database enables the system to monitor and manage healthcare costs effectively. Data acquisition is permanent, from birth to death, irrespective of wealth, age, or work status, resulting in one of the largest electronic health records databases in the world. The data are anonymized but individually linked, which allows individual longitudinal follow-up. As individuals are identified in the database by a unique identifier, double counting of medical information documented from multiple sources is avoided.

The SNDS database links 3 existing databases:

- SNIIRAM (Système National d'Informations Inter-Régimes de l'Assurance Maladie) is the nationwide claims database of the French National Healthcare System. It contains exhaustive, anonymous, and individual data on outpatient healthcare reimbursements. Main data includes data on ambulatory care with reimbursed drugs from community pharmacies and reimbursed medical interventions. It also includes long-term diseases and chronic conditions as well as information about occupational accidents and diseases.
- PMSI (Programme de Médicalisation des Systèmes d'Information) is the national hospital discharge database, concerning both French public- and private-sector hospitals. Main data includes admission and discharge dates, duration of stay, diagnoses (main, related, and associated), as well as procedures (medical acts and biology) and especially costly drugs administered in hospital. Some specific databases exist and depend on the type of hospital admission : medical surgical, and obstetrical wards (PMSI-MCO), home hospitalizations (PMSI-HAD), psychiatric hospitalizations (PMSI-PSY) and rehabilitation centers (PMSI-SSR). Main hospital diagnosis are defined as the condition which occasioned the admission to the hospital, and secondary diagnoses (related and associated) are conditions that coexist at the time of admission, that develop subsequently, or that affect the treatment received and/or length of stay.
- CepiDC (Centre d'épidémiologie sur les causes médicales de décès) is the French national database that collects and analyzes information on the causes of death in France, based on death certificates and hospital records. The registry is managed by the French National Institute of Health and Medical Research and is used for public health research, policy-making, and epidemiological surveillance. The primary purpose of the CepiDC is to provide a comprehensive and accurate record of all deaths in France and their causes.

Hospital diagnoses are coded according to the International Classification of Diseases, 10th revision (ICD-10), which is a classification tool developed by the World Health Organization for epidemiology, health management and clinical purposes. Drugs are coded according to the Anatomical Therapeutic Chemical (ATC) system, that classifies drugs according to the organ or system on which they act and their therapeutic, pharmacological, and chemical properties. Results relating to biological tests and other medical procedures are not recorded, and medical indications are not specified for the reimbursed medical cares. Demographic (age, sex) and socioeconomic (affiliate insurance scheme, universal healthcare coverage and state medical assistance) information are available in the SNDS database. Notably, the universal healthcare coverage is obtained for all individuals whose income is below a specific threshold and was used as a proxy variable for social deprivation in this work.

The SNDS database is managed by the Health Data Hub, a French initiative launched in 2019 to centralize and make available healthcare data for research purposes. The French National Health Insurance Fund (CNAM) is responsible for ensuring that data is collected, processed, and used in compliance with data privacy regulations. The SNDS has been increasingly used for research in recent years. It was described in many studies [Moulis et al., 2015, Bezin et al., 2017, Tuppin et al., 2017, Revet et al., 2022] and has been used to conduct multiple studies in cardiovascular epidemiology [Tuppin et al., 2016, Weill et al., 2016, Giral et al., 2019, Feldman et al., 2021, Piot et al., 2022, Lecoœur et al., 2023].

Importantly, the CNAM developed a standardized tool, called the Healthcare Expenditures and Conditions Mapping (HECM) algorithm, to describe the national annual prevalence of 58 health conditions, grouped into 15 categories and including treated diseases, chronic treatments (without a specific diagnosis identified), and episodes of care (such as maternity) (see Figure 1.17). In 2019, 66.3 million people were identified by the HECM algorithm, including 52% women and 21% people aged 65 years or older, with a median age of 42 years [Rachas et al., 2022]. We used this algorithm to identify several comorbidities in our work.

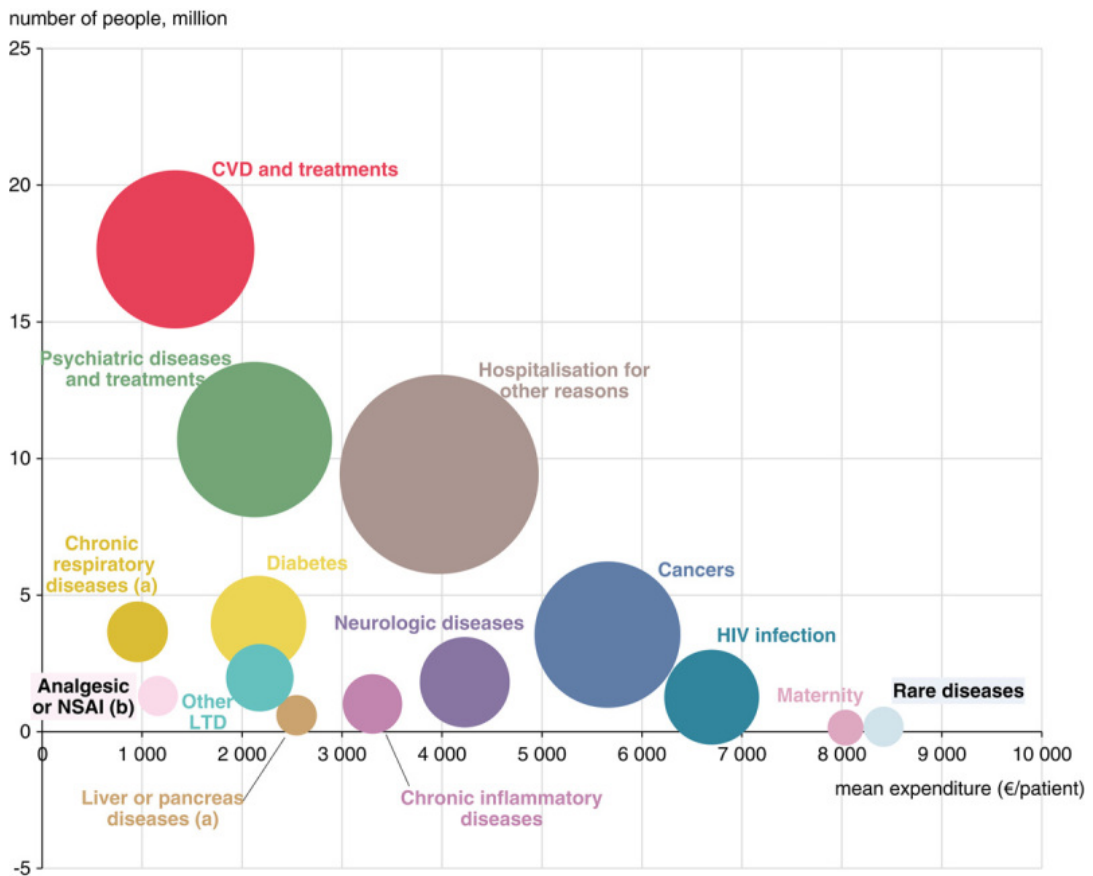


FIGURE 1.17: HECM algorithm

Expenditures by health condition category in 2019 and their components: number of patients and mean expenditure per patient. The size of the bubbles is proportional to the expenditure.

Source: Rachas et al. [2022]

Definition of the populations

In this thesis, we performed a case-control analysis using data extracted from the SNDS database between January 1, 2006 and December 31, 2020. The main group studied was 23,958 SCD (60.4% men, mean age 70.6 ± 17 years) cases collected from the SDEC registry between May 16, 2011 and December 31, 2020. Among them, 4,336 (18.1%) SCD occurred in a public area. A bystander was present in 15,667 (66.6%) of cases and performed CPR in 9,376 (59.5%) of cases. 3,737 (18.8%) had an initial shockable rhythm, 5,167 (21.6%) were transported alive to the hospital and 1,261 (5.3%) patients survived at hospital discharge. SCD cases were matched by age, sex and residence area with control groups through an individual case-control matching:

- A group of 71,919 controls were randomly sampled from the French general population, using the French National Health Insurance database. The endpoint of this group was the day on which SCD occurred among their corresponding cases.
- A group of 71,919 controls experienced myocardial infarction and was identified using the HECM algorithm. The endpoint of this group was defined as the day of the occurrence of the MI event.
- A group of 71,919 has been diagnosed with heart failure and was identified using the HECM algorithm. The primary endpoint for this group was defined as the first day of hospitalization due to heart failure.
- A group of 71,919 has been diagnosed with chronic coronary disease and was identified using the HECM algorithm. The primary endpoint for this group was defined as the first day of hospitalization due to chronic coronary disease.

For each SCD case, 3 corresponding subjects were included in each control group. These groups were selected to identify specific risk factors for SCD, compared to other main cardiovascular diseases, and the general population as well. The work described in this thesis focuses on the two first control groups, while the three last ones were also involved in other related works. Overall, we collected EHR data for 311,634 participants, that represents 282,255,786 medical data points over a period of 15 years (see Figure 1.18).

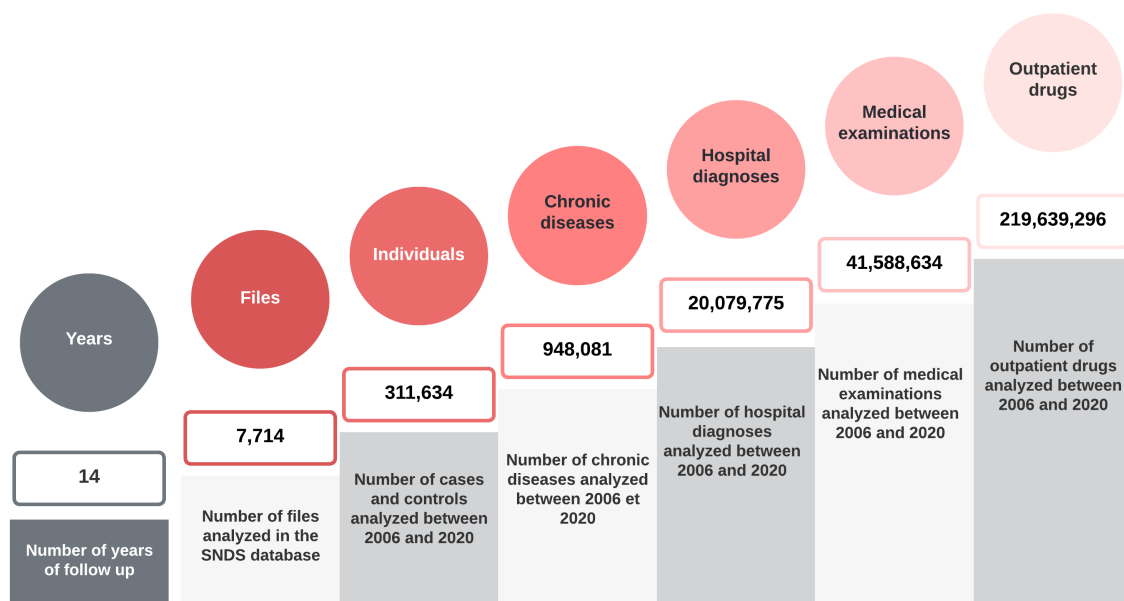


FIGURE 1.18: Data extracted from the SNDS database

1.3 Objective of the Thesis

In this thesis, we aim to provide a better understanding of SCD using statistical modeling. The main goal is to improve current risk stratification guidelines and decrease the global burden of SCD in the population. Three related problems are addressed:

- Problem 1: identifying clinical subgroups of SCD, which may provide new insights about the heterogeneity of profiles and risk factors in the population. To achieve this objective, we use unsupervised learning methods. These methods refer to the analysis of data without the use of a target variable. The underlying goal is to identify hidden patterns in the unlabelled data. One of the most commonly employed unsupervised techniques is called clustering, which involves grouping observations with similar features. Mathematically, we can define occurrence of SCD as a random variable X , and consider a dataset $\mathcal{D} = \{x_1, \dots, x_n\}$ of n observations (i.e patients). Each patient x_i is represented by a vector $x_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$ which describes its medical history prior to the event. Given the dataset \mathcal{D} , we want to find a partition of the observations into K clusters $C = \{C_1, \dots, C_K\}$ (the optimal number of groups is not known *a priori*), which should be both mathematically and clinically relevant.
- Problem 2: predicting SCD in the general population, to identify high-risk subjects who could benefit from specific interventions. To achieve this objective, we use supervised classification methods, which aim to explain the value of a categorical variable Y , based on a set of predictors X . We suppose that we have collected a dataset $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ of n patients, where $x_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$ is the medical history of patients, and $y_i \in \{0, 1\}$ is a binary outcome that represents SCD occurrence. The goal of the classification task is to find the function $F : \mathbb{R}^p \rightarrow \{0, 1\}$ which minimizes the expected loss $\mathcal{L}(F) = \mathbb{E}(L(y, F(x)))$ where $L(\cdot, \cdot)$ is a given loss function. Generally, the data are separated in two sets: the training set is used to build and train the model, while the test set is used to measure its predictive performance on new and unseen observations.
- Problem 3: selecting the most relevant information in large medical history of patients, in order to identify multi-level effects of drugs and diseases associated with SCD. To achieve this objective, we use variable selection methods, which aim to find the optimal subset $S \subseteq \{1, \dots, p\}$ of predictors for our binary regression model, defined in Problem 2. In this work, we suppose that the predictors X belong to some groups (according to medical classification systems) and could affect SCD through individual and/or group effects.

The work proposed for Problem 1 and Problem 2 are based on models that have been already developed, and are commonly used in various statistical learning tasks. The objective is therefore to assess to which extent these models can provide additional insights on SCD, as compared to the current medical literature. For Problem 3, we propose a new model of variable selection, using a Bayesian approach and Monte Carlo methods.

1.4 Summary of Contributions

Clustering Model of Sudden Cardiac Death

Context and objective

Current classifications of SCD are primarily based on cardiovascular phenotypes, and the underlying mechanism is often explained by arrhythmias which occur in patients with cardiomyopathies, mostly from ischemic coronary origin (see section 1.1). However, these

classifications fail to explain numerous cases, and a significant proportion of SCD cases remain unidentified until the event occurs. In many cases, no clear cause is identified despite exhaustive medical and para-medical examinations, including autopsies and genetic analyses. In order to improve the current risk stratification guidelines, we should assess to which extent non-cardiovascular conditions and variables could also play a role in the occurrence of SCD. To this end, we developed a data-driven approach, based on unsupervised statistical models, that can identify new relevant clusters of SCD. We used a wide range of both cardiovascular and non-cardiovascular variables extracted from the SNDS database up to 15 years before the event.

Methodology

The main challenge of this work was to create a meaningful representation of SCD cases based on massive amounts of electronic health records, comprising both structured and unstructured medical information. To this end, we used word embedding methods, which have become a major reference to tackle the issue of medical concepts representation. Each medical code of the patients' medical history was treated as a word, such that a patient can be represented by a sentence whose number of words is equal to the number of medical events that occurred before SCD. We then built a model that transforms the medical codes to numerical vectors of fixed dimensionality, whose relative geometrical positions reflect the medical proximities. Words that co-occurred more frequently should be close together in this embedding space. For this task, we used the Skip-gram architecture of the Word2Vec algorithm, a neural network-based approach, to exploit the co-occurrence information of the medical trajectories.

Word2Vec algorithm

Word2Vec is a word embedding model proposed by Mikolov et al. [2013] that is trained to reconstruct linguistic contexts of words. It takes a corpus of text as input and produces a vector space, called embedding space, as output. Each unique word is assigned a corresponding vector in the embedding space, and word vectors are positioned such that words sharing common contexts in the corpus are located close to one another in the embedding space. There are 2 different types of Word2Vec (see Figure 1.19):

- The continuous bag of words model is trained to predict a word given its surrounding words, called context.
- The Skip-gram model does the opposite and tries to predict the context of a single word.

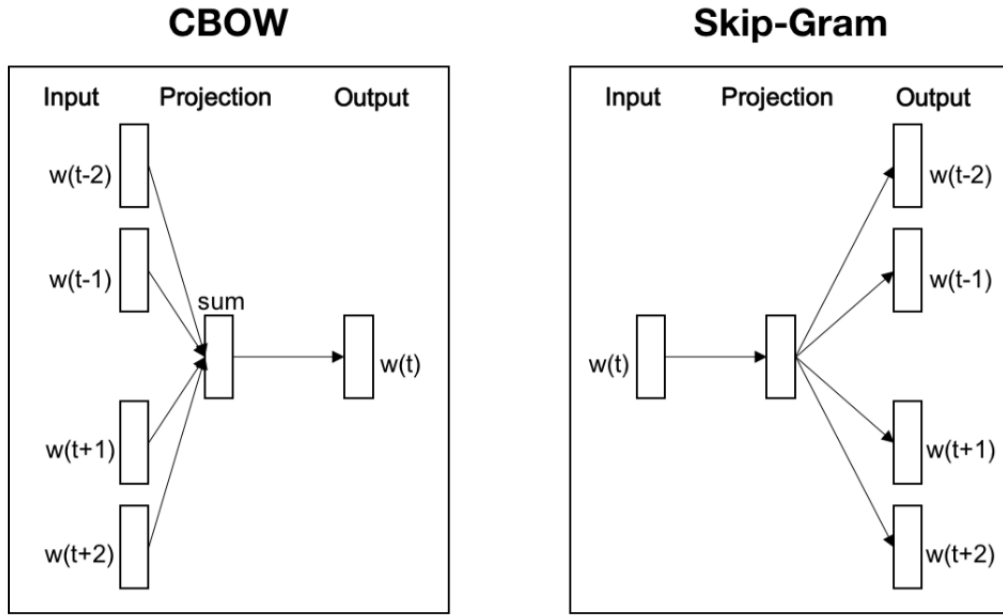


FIGURE 1.19: architectures of Word2Vec

The main objective of these models is to find word representations that are useful for semantic analysis. In our work, we used the Skip-gram version, which works well with a small amount of data, and represents well even rare words or sentences.

We consider a sequence of words (w_1, \dots, w_n) of size n , and composed of r unique elements. These elements are defined in a dictionary D . The objective of Skip-gram is to maximize the conditional probability:

$$\prod_{i=1}^r P(w_{i-m}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+m} | w_i) = \prod_{i=1}^r \prod_{\substack{-m \leq j \leq m \\ j \neq 0}} P(w_{i+j} | w_i) \quad (1.1)$$

where $(w_{i-m}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+m})$ are the m nearest neighbors of the word w_i (i.e. its context). The model supposes that maximizing Equation 1.1 will result in good word embedding in the sense that similar words will have similar vectors.

The Skip-gram model is a neural network built with a single hidden layer, and trained with all pairs of words / contexts. The input of the model is a word w_i of the dictionary D , and the outputs are its m corresponding neighbors $(w_{i-m}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+m})$. The model is defined as follows:

1. In the input layer, the word w_i is converted into a vector $x_i \in \mathbb{R}^r$ using one hot encoding, such that $\forall j \in \llbracket 1, r \rrbracket$:

$$x_{ij} = \begin{cases} 1 & \text{if } w_i = w_j \\ 0 & \text{otherwise} \end{cases}$$

2. The one-hot vector x_i is passed to the hidden layer, which performs the dot product between the embedding matrix W_{xh} and the vector x_i . The output is $H = x_i^T W_{xh}$ where $W_{xh} \in \mathcal{M}_{r,d}$ and d is the number of nodes in the hidden layer. The i th row represents the weights of the i th word of the dictionary D .

3. H is directly passed to the output layer, which performs the dot product between H and the context matrix $W_{hy} \in \mathcal{M}_{d,r}$, whose the i th column represents the embedding weights of the i th word in the dictionary D . The corresponding output is $O = HW_{hy}$.
4. The outputs of the model are the one hot vectors $(y_{i-m}, \dots, y_{i-1}, y_{i+1}, \dots, y_{i+m}) \in \mathbb{R}^r$ that represent the context of the word w_i , such that $\forall -m \leq j \leq m, j \neq 0, z \in \llbracket 1, r \rrbracket$

$$y_{ijz} = \begin{cases} 1 & \text{if the word } w_z \text{ is a neighbor of the word } w_i \\ 0 & \text{otherwise} \end{cases}$$

The probability that w_z is a neighbor of w_i is then computed with the softmax function:

$$P(w_{i+j} = w_z | w_i) = \frac{e^{W_{hyz}}}{\sum_{z'=1}^r e^{W'_{hyz}}}$$

where W_{hyz} represents the z th column of the context matrix W_{hy} . The weights of the model are updated such that the loss function \mathcal{L} of the neural network is minimized using gradient descent and backpropagation methods:

$$\mathcal{L} = \sum_{i=1}^r \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \log P(w_{i+j} | w_i)$$

Once the embedding matrix W_{xh} has been obtained, we can use it to perform semantic analysis and measure the similarity between 2 words (for instance by calculating the cosine similarity between their corresponding word vectors). If 2 words have very similar contexts (*i.e* similar surrounding words), the model should generate similar vectors for these words. Such representation can capture many linguistic regularities. For instance, the representation of the word "Rome" can be obtained with the vector operation "Paris" - "France" + "Italy".

In our work, the medical embedding space obtained from the Word2Vec algorithm corresponds to the hidden representation learned by the model at the end of the training. It represents each medical code by a vector of length 100 (the size of the vector is chosen when we build the model), which was used to represent patients, by computing the mean of vectors corresponding to their medical events occurred before SCD. These vectors therefore summarize their temporal information, such that 2 patients who have similar medical trajectories are expected to be close to each other in the embedding space. We finally used this new representation to perform a clustering analysis, based on the K-Means algorithm, in order to find subgroups of SCD with homogeneous clinical characteristics.

K-Means algorithm

The K-Means clustering is an unsupervised algorithm that aims to partition a set of data points $X = (x_1, x_2, \dots, x_n)$ into K clusters $S = \{S_1, S_2, \dots, S_K\}$, with $K \leq n$. The objective is to generate clusters of patients with a high degree of similarity within each cluster, and a low degree similarity between clusters. It is a very popular method because of its ease of implementation, computational efficiency and low memory consumption. It was proposed for the first time in 1956 by Steinhaus [1956] and developed by Selim and Ismail

[1984]. Formally, the objective is to find S that minimizes the intra-class variance:

$$S^* = \operatorname{argmin}_S \sum_{k=1}^K \sum_{i \in S_k} \|x_i - \mu_k\|^2$$

where μ_k is the mean of data points in S_k , also called centroid, and defined by:

$$\mu_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} x_i$$

Finding an exact solution to the K-Means problem is difficult. Several approximate solutions have been proposed to address this issue. Lloyd's algorithm is the most standard approach. Given an initial set of K centroids (μ_1, \dots, μ_K) , the algorithm proceeds by alternating between 2 steps:

1. Assignment step: assign each datapoint to the cluster with the nearest mean, i.e with the least squared Euclidean distance

$$\forall k \in 1, \dots, K \ S_k = \{x_i : \|x_i - \mu_k\|^2 \leq \|x_i - \mu_{k'}\|^2 \ \forall 1 \leq k' \leq k\}$$

Each data point x_i is assigned exactly 1 cluster.

2. Update step: recalculate the means of the cluster S_k obtained with the assignment step:

$$\forall k \in 1, \dots, K \ \mu_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} x_i$$

Steps 1 and 2 are repeated until the data points assignments (step 1) no longer change. Given enough time, K-Means will always converge. However, this procedure does not guarantee convergence to the global optimum. The result highly depends on the initialization of the clusters. Two main methods have been proposed to initialize the clusters:

- The Forgy method randomly chooses K data points and uses these as the initial centroids.
- The Random Partition method randomly assigns a cluster to each observation and the proceeds to the update step.

In our work, we used the K-Means++ procedure [Arthur and Vassilvitskii, 2007], which is another popular and faster procedure to initialize the cluster centers:

1. We choose one center uniformly at random among the datapoints.
2. For each datapoint x_i not chosen yet, we compute the Euclidean distance d_i between x_i and the nearest center that has already been chosen.
3. We choose one new datapoint x_j as new center, with probability proportional to d_j .

Steps 2 and 3 are then repeated until K centers have been chosen.

The K-Means algorithm is summarized in Algorithm 1:

Algorithm 1: K-Means algorithm

Input : Observations (x_1, \dots, x_n) , number of clusters K

- 1 Choose μ_1, \dots, μ_K centers.
- 2 $t = 0$
- 3 **while** Stopping criteria has not been met **do**
- 4 **for** $i \leftarrow 1$ to n **do**
- 5 **for** $k \leftarrow 1$ to K **do**
- 6 $d_{i,k} = \|x_i - \mu_k\|^2$
- 7 $c_i^{(t)} \rightarrow \underset{k}{\operatorname{argmin}} d_{i,k}$
- 8 **for** $k \leftarrow 1$ to K **do**
- 9 $N_k^{(t)} = \sum_{i=1}^n \mathbb{1}_{[c_i^{(t)}=k]}$
- 10 $\mu_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^n x_i \mathbb{1}_{[c_i^{(t)}=k]}$
- 11 $t = t + 1$

The K-Means algorithm is a simple and efficient clustering method that has been widely used in many fields of machine learning, especially in medical applications. However it has some limitations and drawbacks. One of the main limitations is that it assumes that the clusters are spherical, equally sized, and have similar densities. This assumption may not hold for all datasets, and could therefore lead to suboptimal clustering results. In addition, the algorithm requires the user to specify the number of clusters, which is often challenging, especially for high-dimensional datasets.

To overcome the limitations of K-Means, various clustering algorithms have been proposed in the literature. One popular alternative is hierarchical clustering, which does not require the user to specify the number of clusters and can handle non-spherical clusters. Hierarchical clustering builds a tree-like structure, called a dendrogram, that represents the hierarchy of clusters in the dataset. Spectral clustering have been introduced more recently, and is based on the eigenvectors of the data's similarity matrix. One of its main advantages is its ability to handle non-linearly separable data, which is a common challenge in many clustering problems. Another popular clustering algorithm is DBSCAN (Density-Based Spatial Clustering of Applications with Noise) which groups datapoints based on their density and efficiently identifies outliers and noise.

Main results

This work, which combines a word embedding method with a clustering analysis, provides a new relevant medical representation of SCD cases, that were difficult or impossible to obtain so far with standard statistical approaches. We show that SCD patients can be classified into 8 relevant clinical clusters:

- Cardiovascular diseases (30% of cases)
- Mainly aged women with multiple comorbidites (27% of cases)
- No apparent risk factor (22% of cases)
- Psychiatric and neurologic diseases (7% of cases)
- Respiratory diseases (6% of cases)

- Oncologic diseases (4% of cases)
- Kidney diseases (2% of cases)
- Social deprivation (2% of cases)

The first 2 clusters explain more than 50% of all SCD and are part of the cardiovascular phenotypes already known. The third cluster represents 22% of all SCD and is composed of relatively young subjects without known cardiovascular risk factor, that were very difficult to identify until now. The other 5 clusters are smaller and much less expected. By extending far beyond cardiovascular pathology, our approach provides a global picture of SCD, revealing the involvement of other medical fields that might eventually lead to discover new pathways and help identifying high risk subjects in the general population.

Prediction Model of Sudden Cardiac Death

Context and objective

Despite decades of research, the prognosis of SCD remains poor, with a survival rate below 10%. Recent therapeutic trials have been disappointing, leading to a paradigm shift towards new preventive strategies. In this context, Efficient treatments for SCD management are available. One of the most effective options is the use of implantable cardioverter defibrillators. However, identifying the best candidates for ICD implantation remains challenging. Some very-high risk patients (survivors of SCD, high-risk cardiomyopathies) are clearly identified in current risk stratification guidelines. However, these patients only account for a small proportion of the overall burden of SCD, whereas most patients who experience SCD emerge from the general population without previously known heart disease (see Section 1.1).

Current prediction models in the general population remain disappointing. Improving their performance would require to analyze larger SCD populations, to include more exhaustive patient characteristics (both cardiovascular and non- cardiovascular) and to propose individualized prevention strategies. To this end, we developed and validated a population-based model of SCD prediction, using all cases collected from the SDEC registry, and large-scale data analysis of electronic health records extracted from the SNDS database.

Methodology

In this work, we developed a 3-month prediction model of SCD in the general population, using a supervised learning classification model. The prediction model was trained on SCD cases and matched controls collected between 2011 and 2015 (derivation cohort), with a cross-validation approach, and was then validated on SCD cases and matched controls collected between 2016 and 2020 (validation cohort). We assessed to which extent machine learning approaches could outperform standard statistical methods and compared the Logistic Regression model with 3 ensemble methods (Random Forest, Extreme Gradient Boosting and CatBoost).

CatBoost algorithm

We suppose that we have collected a dataset $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ where $x_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$ is a vector of p predictors and $y_i \in \{0, 1\}$ is a binary outcome. In our case, y_i corresponds to the occurrence of SCD, and x_i is related to the drugs and hospital diagnoses that are observed before the event. The goal of the classification task is to find

the function $F : \mathbb{R}^p \rightarrow \{0, 1\}$ which minimizes the expected loss $\mathcal{L}(F) = \mathbb{E}(L(y, F(x)))$ where $L(\cdot, \cdot)$ is a given loss function.

Gradient boosting models are a family of procedures which build iteratively a sequence of functions $F^t : \mathbb{R}^p \rightarrow \{0, 1\}, \forall t = 1, \dots, T$. Each function F^t is obtained from the previous approximation F^{t-1} in an additive manner:

$$F^t = F^{t-1} + \alpha h^t \quad (1.2)$$

where α is a well-chosen step size and $h^t : \mathbb{R}^p \rightarrow \{0, 1\}$ is a base predictor chosen from a family of functions \mathcal{H} . h^t is chosen from \mathcal{H} in order to minimize the expected loss:

$$\begin{aligned} h^t &= \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{L}(F^{t-1} + h) \\ &= \operatorname{argmin}_{h \in \mathcal{H}} \mathbb{E}(L(y, F^{t-1} + h)) \end{aligned} \quad (1.3)$$

In practice, the expectation in Equation 1.2 is unknown and is approximated with gradient descent by taking a gradient step. The gradient step is chosen in such a way that $h^t(x)$ approximates $-g^t(x, y)$ defined by:

$$g^t(x, y) = \left. \frac{\partial L(y, s)}{\partial s} \right|_{s=F^{t-1}(x)}$$

One common family often chosen for \mathcal{H} is binary decision trees. A decision tree is a model defined by a recursive partition of the predictors space \mathbb{R}^p into several disjoint regions (tree nodes) according to the values of some splitting attributes. These attributes are usually binary variables that identify when predictors x_i^j exceeds some threshold t , where x_i^j is either numerical or binary. Each final region (leaf of the tree) is assigned to a value, which corresponds to the majority class label of the region in the case of a classification problem. A decision tree h can therefore be summarized as:

$$h(x_i) = \sum_{j=1}^p b_j \mathbb{1}_{\{x \in \mathcal{R}_j\}}$$

where $\{(b_j, \mathcal{R}_j)\}_{i=1, \dots, n}$ are the labels attributed to the leaves of the tree and their corresponding disjoint regions respectively.

CatBoost is a popular implementation of gradient boosting proposed by Dorogush et al. [2018] which uses binary decision trees as base predictor and provides a method called ordered boosting to prevent overfitting. Standard gradient boosting models provide an unbiased estimate of the true outcome $y = f^*(x)$ when they use independent datasets at each gradient step. However, in most cases, the same dataset is used at each step, that affects the generalization ability of the model. Dorogush et al. [2018] proposed a boosting algorithm which does not suffer from such a bias. At the start, CatBoost generates $s + 1$ independent random permutations of the training dataset. The permutation $(\sigma_1, \dots, \sigma_s)$ are used for building the new tree at each step, while σ_0 serves for choosing the leaf values of the obtained trees. The final estimator M is obtained following Algorithm 2.

Algorithm 2: CatBoost classification model

Input : Data $\{(x_i, y_i)\}_{i=1, \dots, n}$, number of trees I , step size α , number of permutations s , number of leaves F

- 1 **for** $t \leftarrow 0$ to I **do**
- 2 **if** $t = 0$ **then**
- 3 **for** $r \leftarrow 0$ to s **do**
- 4 $\sigma_r \leftarrow$ random permutation of $[1, n]$
- 5 **for** $i \leftarrow 0$ to n **do**
- 6 $M_{r,0}(x_{\sigma_r}(i)) \leftarrow 0$ [initialize the s models M_r]
- 7 **else**
- 8 $r' \leftarrow$ random($1, s$) [select randomly a permutation to build the tree T_t]
- 9 **for** $i \leftarrow 0$ to n **do**
- 10 $G(i) \leftarrow \frac{\partial L(y_{\sigma_r(i),s})}{\partial s} \Big|_{s=M_{t-1,r}(x_{\sigma_r(i)})}$ [compute the gradient of the loss]
- 11 $T_t \leftarrow$ empty tree
- 12 **for** each step of top-down procedure **do**
- 13 **for** each candidate split c **do**
- 14 **for** $i \leftarrow 0$ to n **do**
- 15 $E_i = \{m \in \sigma_{r'}/m \in \text{leaf}(\sigma_{r'}(i)) \text{ and } m < \sigma_{r'}(i)\}$
- 16 $\Delta(i) \leftarrow \frac{1}{|E_i|} \sum_{m \in E_i} \frac{\partial L(y_m, z)}{\partial z} \Big|_{z=T_{t,c}(x_m)}$
- 17 $c^* = \underset{c}{\operatorname{argmin}} \cos(\Delta, G)$
- 18 $T_t \leftarrow$ add split c^* to T_t [build T_t]
- 19 **for** $f \leftarrow 1$ to F **do**
- 20 $E_f \leftarrow \{m \in \sigma_0/m \in f\}$
- 21 $l_{t,f} \leftarrow \frac{1}{|E_f|} \sum_{m \in E_f} \frac{\partial L(y_m, z)}{\partial z} \Big|_{z=T_t(x_m)}$ [compute the leaves of T_t]
- 22 **for** $i \leftarrow 1$ to n **do**
- 23 $M_{\sigma_0,t}(x_{\sigma_0(i)}) \leftarrow M_{\sigma_0,t-1}(x_{\sigma_0(i)}) - \alpha \sum_{f=1}^F l_{t,f} \mathbb{1}_{\{\text{leaf}(x_{\sigma_0(i)})=f\}}$ [update model M_0]
- 24 **for** $r \leftarrow 1$ to s **do**
- 25 **for** $i \leftarrow 0$ to n **do**
- 26 $E_i = \{m \in \sigma_r/m \in \text{leaf}(\sigma_r(i)) \text{ and } m < \sigma_r(i)\}$
- 27 $M_{r,t}(x_{\sigma_r(i)}) \leftarrow M_{r,t-1}(x_{\sigma_r(i)}) - \alpha \frac{1}{|E_i|} \sum_{p \in E_i} \frac{\partial L(y_m, z)}{\partial z} \Big|_{z=T_t(x_m)}$ [update models $M_r, r \neq 0$]

Our prediction model is based on outpatient drugs and hospital diagnoses that occurred up to 5 years before SCD. To investigate whether non-cardiovascular variables could enhance predictive performance beyond standard risk factors of SCD, we compared 2 different strategies for variable inclusion. The first approach includes only medical codes that attempt to represent traditional risk factors for cardiovascular diseases, based on an exhaustive literature review of SCD prediction models. The second approach, more agnostic, includes all medical codes that occurred up to 5 years before SCD, without any prior selection.

Once the prediction model was trained and optimized, we used the Shapley additive explanations (SHAP) algorithm to explain how the variables relate to the predicted risk at the individual level. SHAP is a model-agnostic representation of variable importance where the impact of each variable on a particular prediction is represented using Shapley values, inspired by cooperative game theory.

Shapley values

Shapley values belong to the class of Additive Feature Attribution methods. Let f be the original prediction model and g a local method designed to explain a prediction $f(x_i)$ based on the predictors x_i of the instance i , $\forall i \in \llbracket 1, n \rrbracket$. Local explanation models often use simplified inputs z_i that map to the original inputs x_i through a function $x_i = h(z_i)$, such that:

$$\forall w \approx z_i, g(w) \approx f(h(z_i))$$

The explanation model g is assumed to be a linear function of binary variables:

$$\forall i \in \llbracket 1, n \rrbracket, g(z_i) = \phi_0 + \sum_{j=1}^p \phi_{ij} z_{ij} \quad (1.4)$$

where $\phi_0 \in \mathbb{R}$, $\phi_{ij} \in \mathbb{R}$, p is the number of predictors and $z_i \in \{0, 1\}^p$ with $x_i = h(z_i)$.

Classic methods of the Additive Feature Attribution class come from the cooperative game theory to compute explanations of model prediction. They use a method called Shapley values, defined as follows. We suppose a coalition of players that cooperate and obtain an overall gain from that cooperation. Since some players may contribute more to the coalition than others, we want to know the fairest way to divide the gain among the players. Shapley values answer this question by computing the average marginal contribution for each player over all possibilities of coalitions. In the context of machine learning, the players are the variables of the model that cooperate to make a prediction, and the gain is the variable importance. It can be theoretically proved that Shapley values are the only additive feature attribution method that satisfies the following properties:

1. Efficiency: the sum of the Shapley values of all variables for a given instance equals the value of the prediction for this instance, so that all the prediction's value is distributed among the variables:

$$\begin{aligned} f(x_i) &= g(h(z_i)) \\ &= \phi_0 + \sum_{j=1}^p \phi_{ij} z_{ij} \end{aligned}$$

where ϕ_{ij} is the Shapley values of the variable j for the instance i .

2. Symmetry: the contribution of two variables should be the same if they contribute equally to all possible coalitions. Let F be the set of all variables included in the model. $\forall i \in \llbracket 1, n \rrbracket, \forall j, k \in \llbracket 1, p \rrbracket^2$

$$f_{S \cup \{j\}}(x_i) = f_{S \cup \{k\}}(x_i) \forall S \subseteq F \setminus \{j, k\} \Rightarrow \phi_{ij} = \phi_{ik}$$

3. Dummy: a variable j that does not change the predicted value of an instance, regardless of which coalition of variables it is added to, should have a Shapley values of 0

$$f_{S \cup \{j\}}(x_i) = f_S(x_i) \forall S \subseteq F \setminus \{j\} \Rightarrow \phi_j = 0$$

4. Linearity: $\forall i \in \llbracket 1, n \rrbracket, \forall j, k \in \llbracket 1, p \rrbracket^2$ and $\alpha \in \mathbb{R}$, let $x_{ij+k} = x_{ij} + x_{ik}$ and $x_{i(\alpha j)} = \alpha x_j$. Then:

$$\begin{cases} \phi_{i(j+k)} = \phi_{ij} + \phi_{ik} \\ \phi_{i(\alpha j)} = \alpha \phi_{ij} \end{cases}$$

To compute the contribution of the variable j on the model prediction of the instance i , we train the models $f_{S \cup \{j\}}$ (with the variable j) and f_S (without the variable j), and compute the difference between $f_{S \cup \{j\}}(x_i)$ and $f_S(x_i)$ for any subset $S \subseteq F \setminus \{j\}$. The Shapley value ϕ_{ij} is then defined by the weighted average of all possible differences:

$$\phi_i = \sum_{S \subseteq F \setminus \{j\}} \binom{|S|}{|F| - 1} (f_{S \cup \{j\}}(x_i) - f_S(x_i)) \quad (1.5)$$

However, Equation 1.5 has an exponential time complexity which makes the method infeasible for practical use. Lundberg and Lee [2017] proposed a solution called SHapley Additive Explanation, using the following assumption:

$$f_S(x_i) = \mathbb{E}(f(x_i | x_{iS}))$$

where x_{iS} corresponds to the input values of the subset S for the instance i . This assumption means that the prediction $f_S(x_i)$ is the expected value of the prediction $f(x_i)$ (with all variables) given x_{iS} . SHAP values attributes to each variable the change in the expected model prediction when conditioning on that variable, in an additive way:

$$\begin{aligned} \phi_0 &= \mathbb{E}(f(x_i)) \\ \phi_{i1} &= \mathbb{E}(f(x_i | x_{i1})) - \mathbb{E}(f(x_i)) \\ &\vdots \\ \phi_{ip} &= \mathbb{E}(f(x_i | x_{i1:p})) - \mathbb{E}(f(x_i)) \end{aligned}$$

When the prediction model is non-linear or the variables are not independent, the order in which the variables are added to the expectation matters. In these cases, the SHAP values are computed by averaging the ϕ_{ij} values across all possible orderings. The exact computation of SHAP is still therefore computationally challenging. We can use sampling procedures to approximate them. Shapley sampling values apply sampling approximation to Equation 1.5. We start by writing a different but equivalent formulation of Equation

1.5:

$$\begin{aligned}\phi_i &= \sum_{S \subseteq F \setminus \{j\}} \binom{|S|}{|F| - 1} (f_{S \cup \{j\}}(x_i) - f_S(x_i)) \\ &= \frac{1}{|F|!} \sum_{S \in E_F} (f_{P_S \cup \{j\}}(x_i) - f_{P_S}(x_i))\end{aligned}\quad (1.6)$$

where E_F is the set of all ordered permutations of the variable indices $\{1, \dots, |F|\}$ and P_S is the set of all indices that precede j in the permutation $S \in E_F$. E_F could be approximated using a simple sampling procedure, where $(f_{P_S \cup \{j\}}(x_i) - f_{P_S}(x_i))$ would be one sample. However, the computational complexity of computing these terms is still exponential. We can simplify Equation 1.6 using:

$$f_S(x_i) = \frac{1}{n} \sum_{l \in n} f(x_l[x_{lj} = x_{ij}, j \in S])$$

where the notation $x_l[x_{lk} = x_{ij}, j \in S]$ denotes the values of the instance l with x_{lj} replaced with x_{ij} . In this case, Equation 1.6 becomes:

$$\phi_i = \frac{1}{|F|!} \sum_{S \in E_f} \sum_{l \in n} \underbrace{(f(x_l[x_{lj} = x_{ij}, j \in S \cup \{j\}]) - f(x_l[x_{lj} = x_{ij}, j \in S]))}_{(d)_{S,l}} \quad (1.7)$$

We can then use a sampling procedure to approximate Equation 1.7. Let $(d)_{S,l}$ be the sampling population. We draw M samples d_1, \dots, d_M with probability $\frac{1}{n}$ with replacement and define $\hat{\phi}_{ij} = \frac{1}{M} \sum_{m=1}^M d_m$. It follows that $\hat{\phi}_{ij}$ is approximately normally distributed with mean ϕ_{ij} and variance $\frac{\sigma_j^2}{M}$, where σ_j^2 is the variance of the population. $\hat{\phi}_{ij}$ is therefore an unbiased and consistent estimator of ϕ_{ij} .

Main results

Our prediction model was trained on 23,958 SCD cases against 23,958 controls, and selected 188 medical codes to predict SCD, both cardiovascular and non-cardio-vascular, among 9,460 potential predictors. The CatBoost algorithm offered the best performance in the cross-validation results. We achieved an AUC of 0.80 (95% CI 0.78 - 0.82) in the derivation cohort, with a positive predictive value of 77% and a sensitivity of 68%. Notably, our model demonstrated excellent discrimination performance in the highest deciles of predicted risk. We detected 2,908 (24%) SCD cases with a predicted risk exceeding 90%, achieving a positive predictive value of 94% in this range. In the validation cohort, we obtained an AUC of 0.80 (0.77 - 0.81), a positive predictive value of 73% and a sensitivity of 71%, which was consistent with the results of the derivation cohort. For each patient, we then identified the most important variables that drive its predicted risk, based on SHAP values. These scores could be used to improve preventive strategies at the individual level. Our model provides therefore a new efficient and personalized approach to accurately identify subjects who are the most at risk of SCD in the general population.

Bi-level Variable Selection for Generalized Linear Models

Context and objective

The main challenge that arises when we analyze heterogeneous data from the SNDS database is the large number of potential drugs and diseases observed, and their infrequent occurrence in the medical history of patients. Indeed, their incidence in the studied populations vary a lot, which makes it difficult to assess the impact of medical codes that are rarely prescribed or observed. On the other hand, recognized nomenclatures for drugs (ATC system) and diseases (ICD-10 classification) exist (see Section 1.2). They facilitate the classification of medical entities into groups with shared properties. For instance, "ST elevation myocardial infarction involving left main coronary artery" is a medical diagnose reported at hospitals by the ICD-10 nomenclature, and which can be classified into 5 hierarchical groups, as follows:

1. ST elevation myocardial infarction involving left main coronary artery (ICD-10 I21.01)
2. ST elevation myocardial infarction of anterior wall (ICD-10 I21.0)
3. Acute myocardial infarction (ICD-10 I21)
4. Ischemic heart diseases (ICD-10 I20-I25)
5. Diseases of the circulatory system (ICD-10 I00-I99)

However, the complexity of these classifications makes it challenging for physicians to identify the optimal level of information to select for each medical code. Therefore, there is clear medical interest in determining automatically whether a particular drug or disease affects SCD, or, if not, whether the groups it belongs to does. To this end, we developed a bi-level variable selection procedure, based on a binary regression model, which should work reliably for a fairly large number of individuals, variables and groups. We wanted this procedure to be Bayesian, in order to be able to obtain posterior probabilities of inclusion (rather than simply 0/1 answers). While this work was initially motivated by SCD prediction models, it can be useful more generally, and applied to other datasets where similar challenges are encountered.

Model

In this work, we suppose that we have collected a dataset $\mathcal{D} = \{X, U, Z, y\}$ with sample size n , where $y \in \{0, 1\}^n$ is a vector of binary responses, $X = (x_{ij}) \in \mathbb{R}^{n \times p}$, $U = (u_{ij}) \in \mathbb{R}^{n \times q}$, and $Z = (z_{ij}) \in \mathbb{R}^{n \times r}$, are design matrices that contain, respectively, individual variables, group variables (both subject to variable selection), and extra variables that one may want to include systematically (e.g. the intercept, socio-demographic effects such as sex, age, etc.).

We suppose that each of the p variables belongs to one (and only one) of the q groups, which may represent different types of group effects. For instance, in our work, the variables in group k may be the indicator that the patient was delivered a certain outpatient drug in the last 5 years before SCD, and the group variable may be the indicator that the patient took any drug in that group in the same period. Let $g(j)$ be the group of variable j . We propose a general approach to capture sparsity at both the group and variable levels. To this end, we introduce a set of two types of binary variables $\theta = (\gamma, \eta)$:

- γ_k indicates whether group k is active ($\gamma_k = 1$) or not ($\gamma_k = 0$).
- η_j indicates whether individual variable j , which is in group $g(j)$, is active ($\eta_j = 1$) or not ($\eta_j = 0$).

We consider a hierarchical structure such that the variable j may be selected only if then group it belongs to, $g(j)$, is selected, that is:

$$P(\eta_j = 1 | \gamma_{g(j)} = 0) = 0$$

As compared to existing models, we propose to keep the flexibility of selecting variables within a group. For example, when a group of drugs is related to SCD, it does not necessarily mean that all drugs of this group are related to SCD. Therefore, we may want to not only remove unimportant groups effectively, but also identify important variables within important groups as well.

The distribution of each data point is defined by:

$$\forall i = 1, \dots, n, P(Y_i = 1 | \beta, \theta) = F \left(\sum_{j=1}^p \eta_j \beta_j^x x_{ij} + \sum_{k=1}^q \gamma_k \beta_k^u u_{ik} + \sum_{l=1}^r \beta_l^z z_{il} \right) \quad (1.8)$$

where $\beta = (\beta^x, \beta^u, \beta^z)$ is the vector of regression parameters and F is the link function (e.g. $F = \Phi$, the unit Gaussian cumulative distribution function for a probit model). We assign independent Gaussian priors to β , and suppose that the prior density of γ is a product of Bernoulli distributions with probabilities p_j^γ . For the predictors, we introduce a spike-and-slab prior defined by:

$$P(\eta_j = 1 | \gamma) = \begin{cases} p_j^\gamma & \text{if } \gamma_{g(j)} = 1 \\ 0 & \text{otherwise.} \end{cases} \quad (1.9)$$

The objective is to perform Bayesian bi-level variable selection in order identify both groups and individuals effects, and therefore to approximate the posterior distribution of $\theta = (\gamma, \eta)$ defined by $\pi(\theta) = p(\theta | \mathcal{D}) \propto p(\theta) L(\theta)$, where $L(\theta)$ is the integrated likelihood obtained by integrating out β :

$$L(\theta) = \int L(\beta, \theta) p(\beta) d\beta, \quad L(\beta, \theta) = \left\{ \prod_{i=1}^N P(Y_i = y_i | \beta, \theta) \right\}$$

To this end, we use a tempering waste-free Sequential Monte Carlo (SMC) sampler proposed by Dau and Chopin [2022] to approximate the joint posterior distribution $\pi(\theta)$. SMC methods are iterative stochastic algorithms that approximate a sequence of probability distributions through successive importance sampling, resampling and Markov steps. In Bayesian modeling, this sequence can be used to interpolate between a distribution $p(\theta)$ which is easy to sample from (e.g. the prior distribution) and a distribution of interest $\pi(\theta)$ which may be difficult to simulate directly (i.e. the posterior distribution). A standard SMC sampler is described in Algorithm 3:

Algorithm 3: SMC sampler

Input : Prior distribution $p(\theta)$, likelihood function $\theta \rightarrow L(\theta)$, integers N, M, P such that $N = MP$, sequence $0 = \lambda_0 < \dots < \lambda_T = 1$, Markov kernels K_t that leave invariant $\pi_{t-1} \forall t \geq 1$

```

1 for  $t \leftarrow 0$  to  $T$  do
2   if  $t = 0$  then
3     for  $n \leftarrow 1$  to  $N$  do
4        $\theta_0^n \sim p(\theta)$ 
5   else
6      $A_t^{1:N} \sim \text{resample}(N, W_{t-1}^{1:N})$  (Draw IID variables such that
7        $P(A_t^k = n) = W_{t-1}^n$  for  $n = 1, \dots, N$ )
8     for  $n \leftarrow 1$  to  $N$  do
9        $\theta_t^n \sim K_t(\theta_{t-1}^{A_t^n}, d\theta_t)$ 
10    for  $n \leftarrow 1$  to  $N$  do
11       $w_t^n \leftarrow L(\theta_t^n)^{\lambda_t - \lambda_{t-1}}$ 
12    for  $n \leftarrow 1$  to  $N$  do
13       $W_t^n \leftarrow w_t^n / \sum_{m=1}^N w_t^m$ 

```

We also replace the marginal likelihood $L(\theta)$ with the approximate Laplace approximation (ALA) approach proposed by Rossell et al. [2021], which offers reliable performance on large datasets, and is less expensive than the standard Laplace approximation.

Main results

Our bi-level variable selection approach based on a waste-free SMC sampler and the ALA approximation demonstrated good performance on both large simulated data and real datasets, within a reasonable computation time. Importantly, this approach offers greater flexibility than most of existing schemes, which impose only “all-in” or “all-out” selection for variables in the same group. To evaluate our model, we applied it on 23,958 SCD cases against 23,958 controls, using all outpatient drugs and hospital diagnose that occurred up to 5 years before SCD. This resulted in a dataset with $q = 36$ groups and $p = 337$ binary variables (using the first 2 levels of ATC and ICD-10 classifications), from which the algorithm selected 16 groups (44%) and 55 variables (16%). The majority of these variables were previously established as well-known risk factors for SCD and described in the medical literature, illustrating the quality of our approach.

Chapter 2

Introduction (en français)

Contents

2.1	La Mort Subite de l'Adulte	61
2.2	Description des Données	62
2.3	Résumé des Contributions	64

Ce chapitre introduit et présente les concepts principaux abordés dans cette thèse. La Section 2.1 définit le concept de mort subite de l'adulte, et les enjeux de recherche qui y sont associés. La section 2.2 décrit les données utilisées pour ce travail. Enfin, la section 2.3 présente les différents travaux réalisés dans le cadre de cette thèse.

2.1 La Mort Subite de l'Adulte

L'arrêt cardiaque est le mécanisme final de tout décès, quelle qu'en soit la cause. Dans la plupart des cas, il survient comme la complication terminale d'une pathologie pré-existante, mais certains arrêts cardiaques sont subis, sans condition prémorbide connue. Lorsqu'il n'existe pas de cause circonstancielle évidente, il s'agit alors d'une mort subite, définie comme une mort inattendue sans cause extracardiaque évidente, survenant avec un effondrement rapide en présence d'un témoin, ou en l'absence de témoin survenant dans l'heure après le début des symptômes.

Cette pathologie touche 30,000 à 40,000 personnes par an en France, et environ 300,000 par an en Europe. Malgré les progrès réalisés dans la prise en charge, la pronostic demeure extrêmement sombre. Moins de 10% des patients sortent en effet vivants de l'hôpital après la survenue d'une mort subite. Son incidence varie fortement en fonction de l'âge et du sexe, et bien qu'elle soit en baisse ces dernières années, notamment du fait d'un meilleur contrôle des facteurs de risque cardiovasculaire et de l'amélioration de la prise en charge des patients atteints de cardiopathie, la mort subite reste responsable de 10% de la mortalité globale et de 50% des décès d'origine cardiovasculaire dans le monde.

Plusieurs outils ont été proposés pour améliorer le pronostic, concernant la prise en charge pré-hospitalière, notamment via la chaîne de survie, le massage cardiaque précoce par les témoins, la défibrillation précoce ou la prise en charge hospitalière (par la prise en charge coronaire précoce ou l'application d'une hypothermie thérapeutique). Les résultats en terme de survie restent néanmoins décevants. Considérant ces résultats modestes sur le versant thérapeutique, plusieurs alternatives préventives ont été proposées pour prévenir la survenue de tels événements. Ainsi, le développement des traitements antiarythmiques et des défibrillateurs automatiques implantables ont permis une prévention significative chez

des patients identifiés à haut risque de mort subite. L'optimisation de l'usage de ces traitements préventifs repose néanmoins sur l'identification préalable des patients à risque de survenue de mort subite. L'amélioration des outils de prédiction, en particulier en population générale, demeure donc un enjeu de recherche majeur. Les travaux réalisés dans le cadre de cette thèse tentent d'y apporter des réponses, à travers l'utilisation d'outils statistiques et d'analyses épidémiologiques.

2.2 Description des Données

Le travail réalisé dans cette thèse repose sur l'analyse de deux sources de données principales : le registre du Centre d'Expertise de la Mort Subite, situé au Centre de Recherche Cardiovasculaire de Paris, et le Système National des Données de Santé, qui regroupe les données médico-administratives de l'Assurance Maladie.

Le Centre d'Expertise de la Mort Subite

Le Centre d'Expertise de la Mort Subite (CEMS) collecte depuis mai 2011 l'ensemble des cas de morts subites survenus dans une zone géographique donnée (Paris et les 3 départements adjacents, Hauts de Seine, Seine-Saint-Denis et Val de Marne), qui représente au total un bassin de population de 6.7 millions d'habitants soit 10% de la population française. Cette collection est rendue possible par une collaboration étagée entre les services de secours préhospitaliers (Brigade des Sapeurs-pompiers de Paris, SAMU), hospitaliers (services de réanimation et de cardiologie) et l'Institut Médico-Légal de Paris. Pour l'ensemble des cas inclus, les informations relatives à la survenue de l'événement (critères Utstein), à la prise en charge (pré et intra hospitalière) et au devenir des patients (en termes de survie et de pronostic neurologique) sont recueillies prospectivement, avec des sources multiples et des contrôles qualité fréquents, permettant d'évaluer l'exhaustivité à 99% des cas dans la zone d'intérêt. Cette collection a fait l'objet de plusieurs publications internationales [Bougouin et al., 2014, Maupain et al., 2016, Jabre et al., 2016, Bougouin et al., 2018, 2020].

Le Système National des Données de Santé

Le Système National des Données de Santé (SNDS) est un entrepôt de données médico-administratives pseudonymisées créé en 2016, couvrant l'ensemble de la population française et contenant l'ensemble des soins présentés au remboursement par l'Assurance Maladie. Il vise à améliorer la santé des patients et l'analyse des dépenses publiques en santé. Le SNDS est géré par la Caisse Nationale de l'Assurance Maladie (CNAM) et par la Plateforme des données de santé (Health Data Hub), qui en propose une documentation détaillée et que nous résumons ici. Il contient ainsi les grandes catégories de données suivantes :

- Les consommations de soins de ville : consultations médicales, prescriptions de médicaments, actes techniques, ...
- Les soins et séjours hospitaliers.
- Les affections de longue durée.
- Les indemnités journalières : maladie, accidents du travail et maladies professionnelles, maternité et invalidité.
- Des informations socio-démographiques sur les bénéficiaires : âge, sexe, commune et département de résidence, Couverture Maladie Universelle Complémentaire, Aide à la Complémentaire Santé et Aide Médicale d'Etat.

- Des informations sur le décès : date, commune et causes médicales de décès.
- Des informations sur les professionnels de santé : médecin traitant, spécialité, mode d'exercice, sexe, âge et département d'implantation

Ces données représentent un flux annuel de 1,2 milliards de feuilles de soins, 1 millions de séjours hospitaliers et 500 millions d'actes, dont la collecte vise à faire de la France l'un des pays pionniers dans le domaine de la promotion et de la valorisation des données de santé. Trois sources principales alimentent le SNDS :

- Le SNIIRAM (Système National d'Information Inter-Régimes de l'Assurance Maladie) contient les données relatives à toutes les dépenses de l'assurance maladie.
- Le PMSI (Programme de Médicalisation des Systèmes d'Information) contient les données relatives à l'activité des établissements hospitaliers.
- Les données du CépiDc (Centre d'épidémiologie sur les causes médicales de décès) contiennent les données relatives aux causes de décès.

Le Système National d'Information Inter-Régimes de l'Assurance Maladie

Le SNIIRAM est un entrepôt de données anonymes regroupant les informations issues des remboursements effectués par l'ensemble des régimes d'assurance maladie pour les soins du secteur libéral. Il comporte le codage détaillé des médicaments délivrés, des actes techniques réalisés, des dispositifs médicaux et des prélèvements biologiques. Il renseigne également les dates de soin ainsi que les montants remboursés par l'assurance maladie et payés par le patient. 3 sources de données alimentent le SNIIRAM :

- Une base de données individuelles des bénéficiaires, appelée DCIR (Datamart de Consommation Inter Régime) pour réaliser des études sur la consommation de soins des bénéficiaires et les pratiques des professionnels de santé. Le DCIR contient l'ensemble des soins de ville remboursés pour les bénéficiaires de l'Assurance Maladie.
- 15 bases de données thématiques de données agrégées, appelées datamarts et orientées vers une finalité particulière : suivi des dépenses, analyse de l'offre de soins libérale, biologie, pharmacie, dispositifs médicaux, établissements privés.
- 1 échantillon général des bénéficiaires (ENSD), représentant 2/100e de la population protégée : l'ESND permet de réaliser des études longitudinales et d'analyser le parcours individuel de près de 1,280,000 bénéficiaires en ville et à l'hôpital.

Le Système National d'Information Inter-Régimes de l'Assurance Maladie

L'objectif principal du PMSI est d'analyser l'activité médicale des établissements hospitaliers à des fins d'allocation budgétaire. Les données relatives à l'ensemble des séjours réalisés dans un établissement de santé, public ou privé, font l'objet d'un recueil systématique, et sont utilisées pour le financement des établissements de santé (tarification à l'activité) ainsi que l'organisation de l'offre de soins. Ce recueil systématique couvre l'ensemble des hospitalisations regroupées en 4 secteurs distincts :

- Les hospitalisations de courte durée (médecine, chirurgie, obstétrique).
- Les soins de suite et de réadaptation.
- Les hospitalisations à domicile.

- Les données relatives à la psychiatrie.

Les informations recueillies par les établissements sont centralisées au niveau national sur la forme d'un résumé de sortie anonyme (RSA). Chaque RSA contient des informations médicales (diagnostics et actes médicaux réalisés) et administratives (identification de l'établissement, durée de séjour, mode d'entrée et de sortie dont, éventuellement, le décès).

Description des Populations

Le travail réalisé dans cette thèse repose sur l'analyse comparative de 23,958 cas de mort subite collectés au CEMS, et de témoins appariés individuellement (à partir du SNDS) selon l'âge, le sexe, et le département de résidence. 4 groupes contrôles ont ainsi été définis:

- 71,919 individus ayant présenté un syndrome coronaire aigu (infarctus du myocarde).
- 71,919 individus atteints d'une insuffisance cardiaque chronique.
- 71,919 individus présentant un antécédent de cardiopathie ischémique.
- 71,919 individus sélectionnés aléatoirement au sein de la population générale.

Au sein des groupes contrôles, 3 témoins ont été appariés pour chaque cas de mort subite. Ces groupes contrôles ont été choisis afin d'identifier des facteurs de risque spécifiques de la mort subite, en comparaison des principales autres maladies cardiovasculaires, ainsi que de la population générale. Les travaux décrits dans cette thèse se concentrent sur les 2 premiers groupes, bien que les 3 derniers aient également été étudiés dans des travaux secondaires. Pour les 5 groupes ainsi définis, nous analysons l'ensemble des données du SNDS collectées entre 2006 et 2020, ce qui représente pour l'ensemble de l'étude :

- 311,634 individus
- 219,639,296 délivrances de médicaments
- 41,588,634 examens médicaux
- 20,079,775 diagnostics hospitaliers
- 948,081 affections de longue durée

2.3 Résumé des Contributions

Clustering de la Mort Subite

Objectif

Les modèles actuels de classification de la mort subite de l'adulte reposent essentiellement sur des causes cardiovasculaires, dont le mécanisme sous-jacent est le plus souvent expliqué par des troubles du rythme survenant chez des individus atteints de cardiopathies ischémiques. Toutefois, ces modèles présentent des limites certaines, et restent insuffisants pour expliquer de nombreux cas observés en population générale. La majorité des morts subites demeure en effet non identifiée jusqu'à leur survenue, et pour lesquelles aucune cause cardiaque n'est souvent clairement établie, malgré un nombre important d'examen médicaux réalisés, incluant notamment des autopsies et des analyses génétiques. Dans ce contexte, il apparaît nécessaire d'évaluer dans quelle mesure l'ajout de facteurs non-cardiovasculaires pourrait permettre de mieux comprendre la pathogénèse de la mort subite.

L'objectif de ce travail consiste donc à développer une nouvelle approche non-supervisée et agnostique, s'affranchissant des classifications actuelles, pour identifier de nouveaux sous-groupes homogènes de morts subites. Ce modèle repose sur l'analyse de l'ensemble de consommations de soin observées jusqu'à 15 ans avant la survenue de l'événements, à partir du SNDS.

Méthode

Le principal enjeu de ce travail consiste à représenter de façon pertinente les trajectoires de soin observées avant la mort subite. Ces trajectoires contiennent des variables hétérogènes et interdépendantes (délivrances de médicaments et hospitalisations), collectées par des sources de données multiples. L'objectif est donc d'en extraire l'information essentielle, tout en occultant le bruit "clinique" n'apportant aucune valeur ajoutée pour l'identification de groupes homogènes de patients. Pour répondre à ce problème, nous nous sommes appuyés sur l'analyse du langage naturel, dont le principe et les méthodes s'appliquent aujourd'hui à de nombreuses problématiques médicales. Nous avons utilisé un modèle de représentation de mots (*word embedding*), en entraînant un réseau de neurones artificiels appelé *Word2Vec*.

Word2Vec est un modèle de traitement du langage qui représente les mots d'un corpus de texte sous une forme vectorielle. Ces vecteurs renseignent des informations sur les relations sémantiques et syntaxiques du corpus, dans un nouvel espace mathématique appelé *embedding*. 2 architectures possibles du modèle existent : *skip-gram* et *bag-of-words*. L'architecture *skip-gram*, que nous avons choisie dans ce travail, consiste à prédire le contexte d'un mot donné, c'est-à-dire les mots qui apparaissent dans son voisinage, en analysant l'ensemble des paires contexte-mot contenues dans le corpus de texte. Cette tâche prédictive est réalisée par un réseau de neurones contenant une couche cachée dont les poids estimés lors de l'entraînement permettent de générer les vecteurs associés aux mots.

Dans notre contexte médical, les délivrances de médicaments et les diagnostics hospitaliers observés dans les trajectoires de soins correspondent à une suite d'événements ordonnés dans le temps. Chacun de ces événement peut être considéré comme un mot, permettant de représenter un patient par une phrase dont le nombre de mots est égal au nombre d'événements survenus avant sa mort subite. Cette approche méthodologique permet ainsi de mesurer la proximité temporelle et médicale de ces événements, et de construire une nouvelle représentation pertinente de son histoire médicale.

L'étape suivante du travail consiste à générer un vecteur pour chaque patient, en calculant la moyenne des vecteurs associés à sa trajectoire de soin. Nous avons utilisé cette dernière représentation vectorielle comme variable d'entrée dans un algorithme de *clustering*, appelé *K-Means*, afin d'identifier des sous-groupes d'individus partageant des caractéristiques cliniques homogènes. L'algorithme *K-Means* est l'une des techniques les plus couramment utilisées en apprentissage non-supervisé. Son objectif est de partitionner un ensemble de données en K groupes homogènes, où K est un nombre pré-défini de groupes. L'initialisation du *K-Means* consiste le plus souvent à sélectionner aléatoirement des individus, servant de barycentres initiaux pour les K groupes. Chaque individu est ensuite affecté au groupe dont le barycentre est le plus proche en termes de distance euclidienne. Les barycentres sont alors recalculés, en effectuant la moyenne des individus qui lui sont attribués. Ce processus d'affectation et de mise à jour se répète jusqu'à ce que les barycentres ne se déplacent plus. La Figure 2.1 résume les différentes étapes de notre modèle.

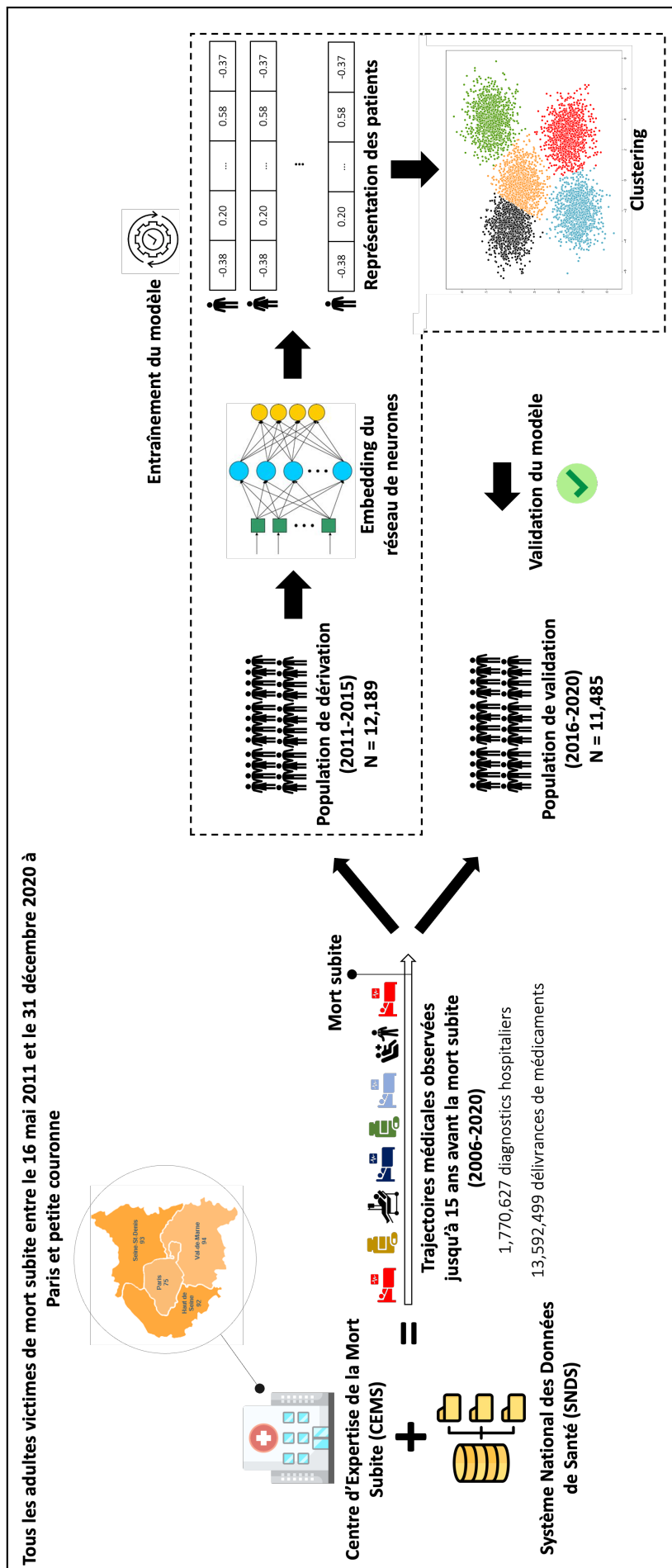


FIGURE 2.1: Schéma méthodologique du modèle de clustering

Résultats

Ce travail combine un modèle d'analyse du langage naturel avec un algorithme de *clustering*. Il nous a permis de construire une nouvelle représentation médicale pertinente des trajectoires de soin observées avant la mort subite. Cette représentation était en effet jusqu'à présent difficile à obtenir en utilisant des modèles statistiques plus conventionnels. Nous montrons ainsi que la population de morts subites peut être répartie en 8 groupes cliniques homogènes :

- Un groupe principalement atteint de maladies cardiovasculaires (30% de la population).
- Un groupe principalement constitué de femmes âgées (27% de la population).
- Un groupe d'individus jeunes ne présentant aucun facteur de risque apparent (22% de la population).
- Un groupe atteints de troubles psychiatriques et neurologiques (7% de la population).
- Un groupe atteints de maladies respiratoires (6% de la population).
- Un groupe d'individus atteints de cancer (4% de la population).
- Un groupe atteints de maladies rénales (2% de la population).
- Un groupe principalement décrit par des désavantages socio-économiques (2% de la population).

Les deux premiers sous-groupes constituent plus de la moitié de la population de mort subite (57%), et se caractérisent par des facteurs cardiovasculaires bien décrits dans la littérature médicale. Le troisième groupe, qui représente 22% des cas, est composé d'individus relativement jeunes, sans facteur de risque cardiovasculaire connu, et qu'il était jusqu'à présent difficile d'identifier en population générale. Enfin, les 5 groupes suivant contiennent des effectifs plus modestes, et dont les caractéristiques sont beaucoup moins décrits dans la littérature médicale. Notre modèle pourrait donc permettre de mieux identifier les facteurs de risque qui leurs sont associés.

En conclusion, ce travail propose une nouvelle classification plus fine des individus victimes de mort subite. Les résultats obtenus pourraient permettre, à terme, d'améliorer leur prise en charge clinique et d'individualiser les stratégies de prévention. Elle offre également de nouvelles perspectives de recherche pour mieux comprendre les causes et mécanismes sous-jacents à la mort subite, et ainsi identifier les sujets à haut risque en population générale.

Prédiction de la Mort Subite

Objectif

Si de nombreuses recherches sur la mort subite ont été menées sur le versant post-événement, notamment sur le soin apporté aux victimes, la prédiction de la survenue d'un tel événement reste difficile. L'identification d'individus à risque de mort subite est donc un enjeu de santé publique majeur, avec des résultats jusqu'à présent décevants. Certains groupes de patients à très haut risque ont été identifiés, mais ils ne constituent qu'une fraction très restreinte de l'ensemble de la population concernée. Ces patients bénéficient d'ores et déjà d'une prise en charge rythmologique spécialisée, visant à la prévention de tels événements. D'un point de

vue épidémiologique, la grande majorité des victimes d'une mort subite ne font cependant pas partie de ces populations à très haut risque. Le principal pourvoyeur de morts subites demeure en effet la cardiopathie ischémique, que ce soit à l'occasion d'un événement aigu (infarctus du myocarde) ou lors du suivi de ces malades. La cohorte des patients atteints d'une cardiopathie ischémique est très importante, et seule une faible proportion d'entre eux présentera une mort subite au cours de son évolution. Il existe par conséquent une discordance entre une population à très haut risque individuel mais d'effectif limité (les cardiopathies pro-arythmogènes spécifiques, structurelles ou électriques) et une population à faible risque individuel mais d'effectif très important (les cardiopathies ischémiques), qui constitue l'essentiel des patients victimes de mort subite en population générale.

Le défi de la prédiction de la mort subite demeure donc entier. Dans ce contexte, ce travail vise à développer un modèle de prédiction en population générale, à partir de l'ensemble des cas inclus dans le CEMS, en les comparant à des témoins appariés. Contrairement aux modèles existants, nous souhaitons proposer un score de risque personnalisé pour chaque individu, et qui intègre un grand nombre de facteurs cardiovasculaires et non-cardiovasculaires collectés à partir du SNDS.

Méthode

Nous avons développé un modèle permettant de prédire la survenue de la mort subite à un horizon temporel de 3 mois, en utilisant un algorithme de classification supervisée. Ce modèle a été entraîné par validation croisée sur les cas collectés au CEMS entre 2011 et 2015 (population de dérivation), et validé sur les cas collectés entre 2015 et 2020 (population de validation). Pour chaque cas, nous avons inclus un témoin apparié individuellement, et sélectionné en population générale.

Afin de mesurer l'apport prédictif des modèles d'apprentissage automatique par rapport aux approches statistiques plus conventionnelles, nous avons comparé le modèle de régression logistique à 3 algorithmes de classification : *Random Forest*, *Extreme Gradient Boosting* et *CatBoost* :

- L'algorithme *Random Forest* (forêts aléatoires) appartient à un ensemble de méthodes appelée *Bagging* (*Bootstrap Aggregating*). Elles consistent à entraîner plusieurs modèles de classification sur des échantillons aléatoires de données, et à les combiner pour obtenir une prédiction finale plus performante que celle obtenue par chacun des modèles séparément, tout en minimisant le risque de surapprentissage. Le *Random Forest* agrège un ensemble d'arbres de décision. Chaque arbre est défini par une suite de critères visant à maximiser la séparation entre les différentes classes à prédire, en choisissant à chaque étape la variable qui offre la meilleure séparation possible. La prédiction finale du *Random Forest* est obtenue en moyennant les prédictions de l'ensemble des arbres de décision.
- Les algorithmes *Extreme Gradient Boosting* et *CatBoost* appartiennent la famille du *boosting*. De la même manière que le *bagging*, ces méthodes s'appuient sur une combinaison de modèles faibles (*weak learner*) pour construire un modèle plus robuste. La principale différence réside dans la façon dont les modèles sont construits et combinés entre eux. Dans le cas du *boosting*, ces derniers sont construits de façon itérative, de telle sorte que chaque modèle corrige les erreurs commises par le précédent, en accordant plus de poids aux observations mal classées. L'algorithme *CatBoost* est une version particulièrement performante du *boosting* qui utilise une technique appelée "boosting ordonné" (*ordered boosting*). Cette technique prend en compte l'ordre des

variables dans la construction de l'algorithme: les variables sont classées et ajoutées séquentiellement selon leur ordre d'importance dans le modèle, générant des arbres de décision moins complexes et plus rapides à construire, tout en réduisant le risque de surapprentissage.

Notre modèle de prédiction a été entraîné à partir de l'ensemble des délivrances de médicaments et diagnostics hospitaliers observés sur une période de 5 ans avant la survenue de la mort subite. Pour déterminer dans quelle mesure l'ajout de facteurs non-cardiovasculaires améliore l'identification des sujets à risque, nous comparons deux stratégies différentes d'inclusion des variables :

- La première approche se limite à l'inclusion de variables uniquement cardiovasculaires, en s'appuyant sur les facteurs de risques de mort subite décrits dans la littérature médicale.
- La seconde approche, plus agnostique, s'affranchit de ces hypothèses et inclut l'ensemble des codes médicaux observés dans les trajectoires de soin.

Après avoir entraîné et validé le modèle en comparant les 2 stratégies décrites ci-dessus ainsi que les différents algorithmes de classification supervisée, nous proposons d'expliquer à l'échelle individuelle les scores de risque ainsi générés. Pour chaque patient, nous identifions les variables les plus importantes qui sont associées à son risque, en utilisant l'algorithme *SHAP*. Cet algorithme produit pour chaque variable et pour chaque individu donné, un score d'importance, appelé valeur de Shapley, à partir du modèle de prédiction préalablement entraîné. Les valeurs de Shapley reposent sur la théorie des jeux collaboratifs, et permettent d'estimer la contribution marginale d'un joueur qui réalise une action avec d'autres joueurs. Dans le cadre d'un modèle de prédiction :

- Le joueur correspond à une variable explicative.
- L'action correspond à la prédiction réalisée par le modèle auquel il appartient.
- La contribution correspond au score d'importance qui lui est associé.

Pour un patient donné, les valeurs de Shapley se calculent en considérant l'ensemble des interactions possibles que les variables génèrent entre elles pour réaliser la prédiction. Pour chaque permutation, la contribution marginale d'une variable correspond à la différence obtenue en comparant le score de risque généré avec, et sans cette variable. Ainsi, la valeur de Shapley associée s'obtient en calculant la moyenne de toutes les contributions marginales, en tenant compte de l'ensemble des permutations possibles. En pratique, des solutions existent pour approximer cette méthode combinatoire, qui est souvent impossible à calculer lorsque le nombre de variables et d'individus est important.

La méthode *SHAP*, et de façon générale les méthodes d'explicabilités, sont particulièrement utiles pour expliquer des prédictions réalisées par des modèles complexes et peu interprétables, tels que les réseaux de neurones ou les algorithmes de *boosting*. Elles présentent un intérêt tout particulier pour la prédiction de la mort subite, car elles permettraient d'identifier les facteurs de risques modifiables à l'échelle individuelle, et ainsi d'individualiser les stratégies de prévention. La Figure 2.2 résume les différentes étapes de notre modèle.

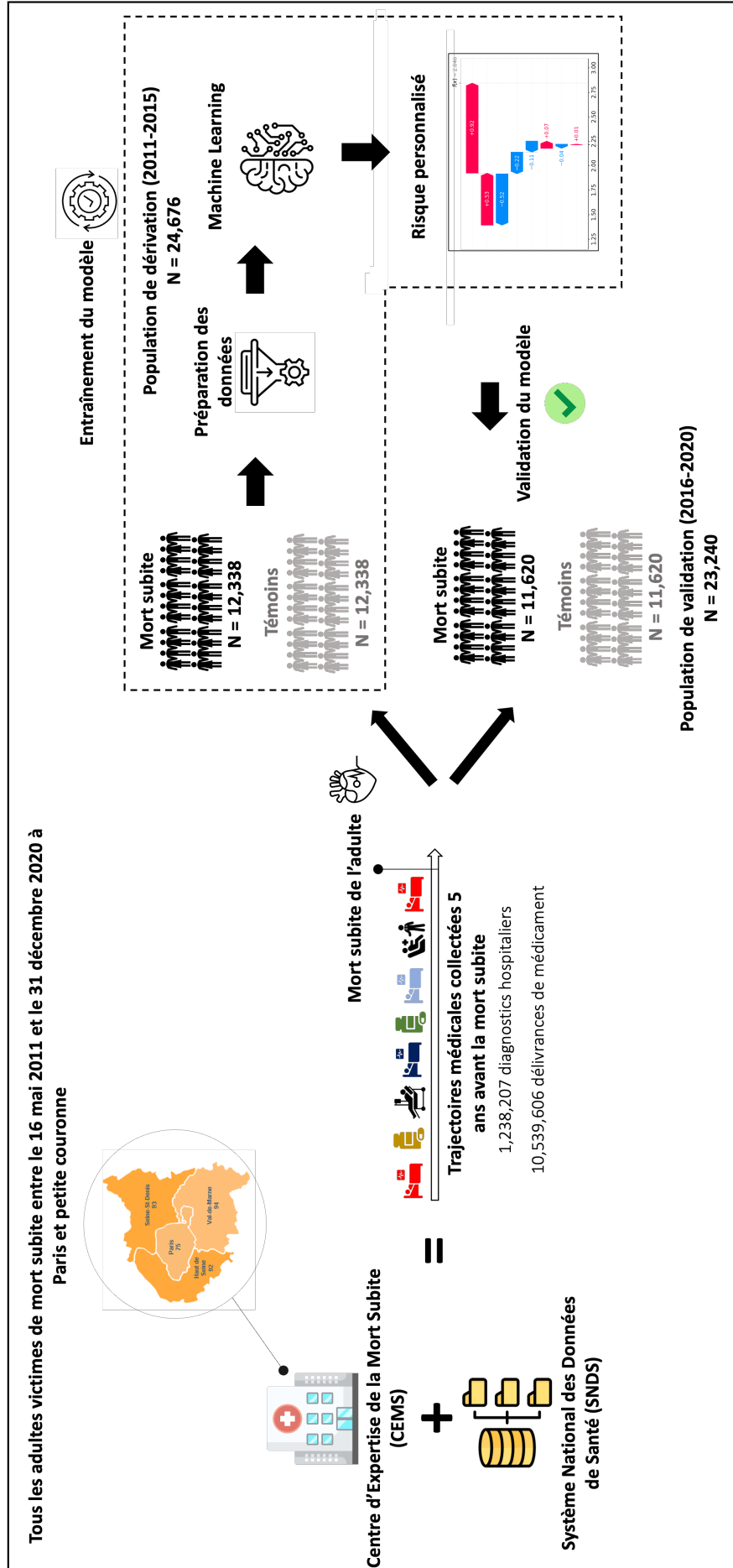


FIGURE 2.2: Schéma méthodologique du modèle de prédiction

Résultats

Nous avons entraîné notre modèle de prédiction sur 23,958 cas de morts subites et 23,958 témoins individuellement appariés et sélectionnés en population générale. L'algorithme *CatBoost*, combiné à une stratégie agnostique d'inclusion des variables (non restreinte aux facteurs cardiovasculaires) obtient les meilleures performances prédictives en validation croisée. Ce modèle a sélectionné 188 codes médicaux d'intérêt, cardiovasculaires et non cardiovasculaires, parmi 9,460 prédicteurs potentiels. Nous avons obtenu une AUC de 0,80 (CI 95% 0,78 - 0,82) dans la population de dérivation, avec une valeur prédictive positive de 77% et une sensibilité de 68%. Par ailleurs, notre modèle présente une excellente performance prédictive dans les déciles les plus élevés du risque prédit. En effet, 2,908 (24%) des cas de mort subites ont été identifiés avec un risque supérieur à 90%, atteignant une valeur prédictive positive de 94% dans cet intervalle. Dans la population de validation, nous avons obtenu une AUC de 0,80 (CI 95% 0,77 - 0,81), une valeur prédictive positive de 73% et une sensibilité de 71%, validant les résultats décrits précédemment. Pour chaque individu, nous avons ensuite identifié les variables les plus importantes qui expliquent son risque prédit, à partir des valeurs de Shapley. En conclusion, notre modèle fournit donc une nouvelle approche efficace et individualisée permettant d'identifier avec précision les individus qui sont les plus à risque de mort subite en population générale.

Modèle de Sélection de Variables Bi-Niveaux

Objectif

La principale difficulté méthodologique que pose l'analyse des données du SNDS est liée au nombre très important de codes médicaux qu'il est possible d'observer dans les trajectoires de soins. En effet, la fréquence d'apparition de ces variables est très hétérogène, et leur incidence dans les populations étudiées varie beaucoup. Cela rend ainsi difficile l'analyse des effets associés à certains médicaments ou diagnostics hospitaliers rarement prescrits ou observés avant la mort subite.

Par ailleurs, il existe aujourd'hui des nomenclatures reconnues dans la littérature médicale qui permettent de regrouper les médicaments (classification ATC) ou les diagnostics (classification CIM-10) en groupes d'entités partageant des propriétés et caractéristiques communs. A titre d'exemple, *l'infarctus du myocarde avec sus-décalage du segment ST impliquant l'artère coronaire principale gauche* est un diagnostic codé dans les hôpitaux selon la nomenclature CIM-10, et qui peut-être regroupé en 5 classes hiérarchiques :

1. Infarctus du myocarde avec sus-décalage du segment ST impliquant l'artère coronaire principale gauche (CIM-10 I21.01)
2. Infarctus du myocarde avec sus-décalage du segment ST de la paroi antérieure (CIM-10 I21.0)
3. Infarctus aigu du myocarde (CIM-10 I21)
4. Cardiopathies ischémiques (CIM-10 I20-I25)
5. Maladies de l'appareil circulatoire (CIM-10 I00-I99)

Dans les approches de clustering et de prédiction décrites précédemment, nous avons choisi d'appliquer des stratégies systématiques (i.e non individualisées) de regroupement des codes, reposant sur des critères uniquement médicaux. Tous les codes ont ainsi été regroupés au niveau 3 ou 4 des nomenclatures médicales, selon les modèles considérés. Cette

stratégie est facile à mettre en oeuvre et présente l'avantage de pouvoir facilement comparer les variables entre elles. Cependant, elle n'est pas toujours pertinente du point de vue clinique, car certains actes de soin doivent être analysés à un niveau détaillé, tandis que d'autres peuvent être considérés dans des groupes plus larges.

L'objectif de ce travail consiste donc à proposer une méthode automatique et individualisée pour discriminer les effets de groupes des effets individuels, et ainsi identifier le niveau optimal d'information. Nous avons développé un algorithme de sélection de variables bi-niveaux pour un problème de régression binaire et pouvant s'appliquer à un grand nombre d'individus et de variables en entrée. Nous construisons cette approche dans un cadre bayésien afin de proposer, dans un souci d'interprétabilité, des probabilités *a posteriori* d'inclusion plutôt que de simples réponses binaires. Enfin, bien que ce travail soit motivé par le développement d'outils pour la prédiction de la mort subite, il peut-être utilisé et adapté pour un grand nombre d'autres situations, où les mêmes enjeux méthodologiques se posent.

Modèle

Nous supposons disposer d'un ensemble de données $\mathcal{D} = \{X, U, Z, y\}$ de taille n , où $y \in \{0, 1\}^n$ est un vecteur de réponses binaires, et où $X = (x_{ij}) \in \mathbb{R}^{n \times p}$, $U = (u_{ij}) \in \mathbb{R}^{n \times q}$, et $Z = (z_{ij}) \in \mathbb{R}^{n \times r}$ correspondent respectivement aux p variables individuelles, aux q variables de groupe, et aux r variables supplémentaires que l'on souhaite inclure systématiquement dans l'analyse (par exemple des variables socio-démographiques tels que l'âge ou le sexe).

Nous supposons que chacune des p variables appartient à un (et seulement un) des q groupes précédemment définis. Ces variables peuvent représenter différents types d'effet. Par exemple, dans notre contexte médical, les variables individuelles du groupe k peuvent indiquer que le patient a reçu un certain médicament au cours des 5 dernières années avant la survenue de la mort subite, et la variable de groupe peut indiquer que le patient a pris n'importe quel médicament de ce groupe au cours de la même période. Soit $g(j)$ le groupe contenant la variable j . Nous proposons un modèle permettant de sélectionner simultanément les effets de groupes et les effets individuels. Pour cela, nous introduisons deux variables binaires $\theta = (\gamma, \eta)$, définies de la façon suivante :

- γ_k indique si le groupe k est actif ($\gamma_k = 1$) ou non ($\gamma_k = 0$).
- η_j indique si la variable individuelle j , appartenant au groupe $g(j)$, est active ($\eta_j = 1$) ou non ($\eta_j = 0$).

Nous considérons une structure hiérarchique définie telle que la variable j ne peut être sélectionnée que si le groupe $g(j)$, auquel elle appartient, est également sélectionné, c'est-à-dire $P(\eta_j = 1 | \gamma_{g(j)} = 0) = 0$. Nous proposons également de conserver une flexibilité dans la sélection des variables au sein d'un même groupe, par rapport à l'approche dite *all-in all-out* des modèles existants. Cela signifie qu'au sein d'un groupe sélectionné, toutes les variables ne sont pas nécessairement conservées par le modèle. Cette flexibilité est pertinente du point de vue clinique, car il est souvent peu probable que l'effet de tous les médicaments au sein d'une classe donnée soit identique.

Notre modèle est ainsi défini de la façon suivante :

$$\forall i = 1, \dots, n, P(Y_i = 1|\beta, \theta) = F \left(\sum_{j=1}^p \eta_j \beta_j^x x_{ij} + \sum_{k=1}^q \gamma_k \beta_k^u u_{ik} + \sum_{l=1}^r \beta_l^z z_{il} \right) \quad (2.1)$$

où $\beta = (\beta^x, \beta^u, \beta^z)$ est le vecteur des paramètres de régression et F est la fonction de lien (par exemple, la fonction de répartition gaussienne $F = \Phi$ d'un modèle Probit). Dans le cadre bayésien, nous assignons des lois *a priori* gaussiennes indépendantes aux paramètres β , et nous supposons que la densité *a priori* de γ est un produit de distributions de Bernoulli avec des probabilités individuelles p_j^γ . Pour les paramètres de sélection de variables, nous introduisons une loi *a priori* de type *slope-and-slab* définie par

$$P(\eta_j = 1|\gamma) = \begin{cases} p_j^\gamma & \text{si } \gamma_{g(j)} = 1 \\ 0 & \text{sinon} \end{cases} \quad (2.2)$$

L'objectif est de réaliser une sélection de variables bi-niveaux, en estimant la loi *a posteriori* jointe de $\theta = (\gamma, \eta)$ définie par $\pi(\theta) = p(\theta|\mathcal{D}) \propto p(\theta)L(\theta)$, où $L(\theta)$ correspond à la vraisemblance marginale obtenue en intégrant selon β :

$$L(\theta) = \int L(\beta, \theta)p(\beta)d\beta, \quad L(\beta, \theta) = \left\{ \prod_{i=1}^N P(Y_i = y_i|\beta, \theta) \right\}$$

Nous utilisons un échantillonneur de Monte-Carlo Séquentiel (SMC), et plus particulièrement la version *waste-free* développée par Dau and Chopin [2022], pour estimer la loi $\pi(\theta)$, combinée avec la méthode *approximate Laplace approximation* (ALA) proposée par Rossell et al. [2021] pour approximer la vraisemblance marginale à chaque étape de l'algorithme. Les méthodes SMC sont des algorithmes itératifs permettant d'estimer une suite de distributions de probabilités en appliquant des étapes successives d'échantillonnage préférentiel (*importance sampling*) et de Monte-Carlo par chaîne de Markov (MCMC). Elles sont particulièrement utiles en analyse bayésienne pour estimer une loi *a posteriori* multimodale et complexe à obtenir, en générant des distributions intermédiaires à partir de la loi *a priori*. A chaque étape t de l'algorithme, l'échantillonneur SMC génère un ensemble de particules $\theta_t = \{\theta_t^1, \dots, \theta_t^N\}$ permettant d'estimer π_t . Chaque particule est pondérée selon des poids W_1^1, \dots, W_1^N , calculés en fonction de la vraisemblance marginale $L(\theta_t)$. Les particules les plus probables sont conservées par un processus de ré-échantillonnage, tandis que les autres sont éliminées. Toutefois, ce ré-échantillonnage présente l'inconvénient de créer des groupes de particules identiques, ce qui appauvrit l'échantillon généré. Pour remédier à ce problème, les particules conservées après ré-échantillonnage sont injectées dans un noyau MCMC laissant invariante la loi π_t , sur plusieurs itérations, afin de produire un échantillon de meilleure qualité. Les particules restantes à la fin de l'algorithme permettent d'estimer les quantités d'intérêt de la loi π , en utilisant l'estimateur $\sum_{i=1}^N W_T^i \phi(\theta_i)$ pour $\pi(\phi) := \mathbb{E}_\pi(\phi(\theta))$.

Résultats

Nous avons développé un algorithme de sélection de variables bi-niveaux dans un cadre bayésien, à partir d'un échantillonneur SMC et pour un problème de régression binaire. Cet algorithme offre notamment une plus grande flexibilité de sélection de variables au sein

des groupes, par rapport aux modèles existants. Notre approche obtient de bonnes performances sur de grands jeux de données, à la fois réels et simulés, dans un temps de calcul raisonnable. En particulier, nous l'avons évalué sur les données du SNDS pour la prédiction de la mort subite, en incluant 23,958 cas du CEMS, 23,958 témoins, ainsi que l'ensemble des variables représentant les médicaments et les diagnostics hospitaliers observés sur une période de 5 ans avant l'événement. En agréant ces variables selon les niveaux 1 et 2 des classifications médicales, cela représente $q = 36$ groupes et $p = 337$ variables individuelles. Notre algorithme a sélectionné 16 groupes (44% du total) et 55 variables individuelles (16% du total). La majorité de ces codes médicaux sélectionnés étaient pertinents du point de vue clinique, et bien décrits comme des facteurs de risque associés à la mort subite dans la littérature médicale, démontrant ainsi l'intérêt de notre approche.

Chapter 3

Identifying Subgroups of SCD with Clustering Analysis

Contents

3.1 Introduction	75
3.2 Methods	76
3.3 Results	80
3.4 Discussion	92

This Chapter comes from a joint work with Patricia Jabre, Wulfran Bougouin, Frankie Beganton, Yseult Masson, Matthieu Bricaire, Jean-Philippe Empana, Nicolas Chopin and Xavier Jouven, and was submitted in Nature Journal in April 2023.

3.1 Introduction

Sudden death is an unexpected and natural death that occurs within one hour after the occurrence of the first symptoms [Fishman et al., 2010]. After exclusion of obvious extra cardiac causes, about 90% of sudden death are considered from cardiovascular origin, and the term of sudden cardiac death (SCD) or sudden cardiac arrest is generally used [Fishman et al., 2010, Hinkle and Thaler, 1982]. SCD is a major public health burden estimated to account for 10-20% of all deaths [Myerburg and Castellanos, 2009]. The presumption based on epidemiological studies, is that such rapid deaths are often because of lethal ventricular arrhythmias [Hayashi et al., 2015]. Ventricular fibrillation (VF), followed by asystole, is the most frequent cause of SCD and is definitely a cardiovascular event, sometimes the first and usually the last [Weisfeldt et al., 2010]. According to the current classification, VF and asystole occur in patients with underlying cardiomyopathies, mostly from ischemic coronary origin [Myerburg and Junttila, 2012]. However, in many cases of SCD, no cause is clearly found despite numerous medical and para medical examinations including autopsy and genetics [Hayashi et al., 2015, Fishman et al., 2010]. The prevention is even worse since the identification of high risk subjects is very difficult and limited, even among high level athletes with many cardiac investigations [Jouven et al., 1999, 2017, Malhotra and Sharma, 2018].

Obviously, something is missing precluding to explain properly the possible diagnosis and mechanisms. One possible reason is that the current approach tries to explain these last cardiovascular events (VF and asystole) by cardiovascular causes essentially neglecting the overall past medical history. We might have to relax this strong hypothesis and to consider to a larger extent that non cardiovascular conditions and variables could play a role in the occurrence of SCD. To address this issue, we assessed whether a data-driven

non-supervised approach based on artificial intelligence could identify new clusters of SCD, using an unselected wide range of cardiovascular and non-cardiovascular variables collected from electronic health records up to 10 years before the occurrence of SCD.

3.2 Methods

Data acquisition

This study was conducted in accordance with all relevant French regulatory requirements. Access to the French National Health Insurance Database is regulated by the Committee of Expertise for Research, Studies and Evaluations in the field of Health (CEREES) and the French National Data Protection Agency (CNIL). The study protocol was submitted and approved in 2016 for SCD cases collected between 2011 and 2015 (Institut des Données de Santé, approval N°183, 2016; CCTIRS approval N°12-336, 2016; CNIL authorization DR-2016-401, 2016) and in 2022 for cases collected between 2016 and 2020 (CEREES, approval 2785673, 2020). The Paris Sudden Death registry was also approved by the CNIL (CNIL authorization DR-2012-445, 2012). In accordance with the regulations in force, informed patient consent was not required due to the retrospective and observational nature of the study.

The Paris Sudden Death Expertise Center

Every case of unexpected out-of-hospital cardiac arrest in persons older than 18 years that occurred between 16 May 2011 and 31 December 2020 in Paris (France) and its inner suburbs (Hauts-de-Seine, Seine-Saint-Denis, Val-de Marne) was collected throughout the Paris Sudden Death Expertise Center (SDEC). The SDEC is a multidisciplinary consortium dedicated to research, education, and care of SCD [Bougouin et al., 2014, Maupain et al., 2016, Jabre et al., 2016, Bougouin et al., 2018, 2020]. Exclusion criteria are a prior terminal condition, no attempt at advanced cardiac life support by Emergency Medical System personnel, or an obvious noncardiac cause according to Utstein templates for resuscitation registries reporting data on cardiac arrest [Jacobs et al., 2004, Perkins et al., 2015]. Hence, the included subjects in this study were SCD cases.

The SDEC registry is a multicenter population-based registry system covering a population of 6.7 million inhabitants (10% of the French population). It records prospectively and continuously information on the occurrence (Utstein criteria), management (pre- and in-hospital) and patient outcomes (regarding survival and neurological outcomes) of all SCD cases. This includes information about age, sex, location of SCD, presence of a bystander, initial cardiac rhythm, cardiopulmonary resuscitation, alive transportation to the hospital, coronary angiogram and survival.

To ensure completeness of collection in the area, an intensive and prospective epidemiological case-ascertainment programme was applied. In France, the emergency medical service is a two-tiered physician-manned system, with a basic life support tier served by firefighters of the Brigade de Sapeurs Pompiers de Paris (BSPP), and an advanced cardiac life support tier (ACLS) [Adnet and Lapostolle, 2004]. The SDEC Registry is derived with the following procedure. First, a nominative case report form is sent daily for every cardiac arrest supported by BSPP. Second, an electronic query algorithm is performed in the advanced cardiac life support computer system to identify every case of SCD. Third, retrospective controls based on diagnostic codes are conducted in selected intensive care units. This method therefore involves every link of the chain of survival, to ensure completeness of the registry. We performed a retrospective control among a sample of 3 intensive care

units, and combination of both sources (BSPP and ACLS) detected 99% of cases of cardiac arrests admitted alive in this sample [Bougouin et al., 2014]. In addition, each case is reviewed separately by two investigators of the SDEC, to ensure accuracy of classification and to avoid the over-estimation often experienced in retrospective collection. Our clustering analysis was derived from cases of the Paris SDEC registry collected between 2011 and 2015, and validated on cases collected between 2016 and 2020 to assess the temporal transportability of our approach.

The French National Healthcare System Database

The SDEC registry was linked with the French National Health Insurance Database (SNDS) [Tuppin et al., 2017, Moulis et al., 2015, Bezin et al., 2017]. The French Universal Health Insurance System manages all reimbursements of healthcare for all people affiliated to a health insurance scheme in France, covering 98% of the population (67 million inhabitants). It provides information on all healthcare expenses, on an individual level, including outpatient visits, procedure, and reimbursed drugs relative to outpatient medical care claims; information from hospital discharge summaries; chronic conditions. Data acquisition is permanent, from birth to death, irrespective of wealth, age, or work status, resulting in one of the largest electronic health records databases in the world. The data are anonymized but individually linked, which allows individual longitudinal follow-up. As individuals are identified in the database by a unique identifier, double counting of medical information documented from multiple sources is avoided.

The SNDS database links 2 existing databases:

- The nationwide claims database of the French National Healthcare System (SNIIRAM) contains exhaustive, anonymous, and individual data on outpatient healthcare reimbursements. It includes data on ambulatory care with all reimbursed drugs from community pharmacies and all reimbursed medical interventions. It also includes long-term diseases and chronic conditions as well as information about occupational accidents and diseases.
- The national hospital database (PMSI) is the national hospital discharge database, concerning both French public- and private-sector hospitals. Main data includes admission and discharge dates, duration of stay, diagnoses (main, related, and associated), as well as procedures (medical acts and biology) and especially costly drugs administered in hospital. Specific databases exist for hospital admissions in medical, surgical, and obstetrical wards (PMSI-MCO), home hospitalizations (PMSI-HAD), psychiatric hospitalizations (PMSI-PSY) and rehabilitation centers (PMSI-SSR). In this study, we only collected principal and secondary hospital diagnosis codes from the PMSI-MCO database. Principal diagnosis is defined as the condition, after study, which occasioned the admission to the hospital, and secondary diagnoses are conditions that coexist at the time of admission, that develop subsequently, or that affect the treatment received and/or length of stay.

Hospital diagnoses and long-term diseases are coded according to the International Classification of Diseases, 10th revision (ICD-10), which is a classification tool developed by the World Health Organization for epidemiology, health management and clinical purposes. Drugs are coded according to the Anatomical Therapeutic Chemical (ATC) system, that classifies drugs according to the organ or system on which they act and their therapeutic, pharmacological, and chemical properties. Results relating to biological tests and other medical procedures are not recorded, and medical indications are not specified for the

reimbursed medical cares.

Demographic (age, sex, place of residence) and socioeconomic (affiliate insurance scheme, universal healthcare coverage and state medical assistance) information are available in the SNDS database. The universal healthcare coverage is obtained for all individuals whose income is below a specific threshold and was used as a proxy variable for low income. The state medical assistance covers the healthcare costs of foreigners who do not meet the requirement of legal residence allowing them access to the universal healthcare coverage.

The SNDS database has been described in detail previously [Tuppin et al., 2017, Moulis et al., 2015, Bezin et al., 2017, Revet et al., 2022] and has been used to conduct multiple studies in cardiovascular epidemiology [Tuppin et al., 2016, Weill et al., 2016, Giral et al., 2019, Feldman et al., 2021, Piot et al., 2022]. More details are available at <https://www.health-data-hub.fr/>.

Data processing

To perform our clustering analysis, we analyzed all reimbursed drugs relative to outpatient medical care claims and all hospital diagnoses relative to medical, surgical, and obstetrical wards that occurred up to 15 years before SCD.

Drugs were grouped according to their pharmacological subgroups, corresponding to the fourth level of the ATC system. For instance, the combination of nadolol and thiazides (ATC C07BA12) was labeled as selective beta blocking agents, non-selective, and thiazides (ATC C07BA) as follows :

- C: Cardiovascular system
- C07: Beta blocking agents
- C07B: Beta blocking agents and thiazides
- C07BA: Beta blocking agents, non-selective, and thiazides
- C07BA12: Nadolol and thiazides

Hospital diagnoses were grouped according to the fourth level of the ICD-10 classification. For instance, atherosclerotic heart disease of native coronary artery with unstable angina pectoris (ICD-10 I25.110) was labeled as atherosclerotic heart disease of native coronary artery (ICD-10 I25.1) as follows:

- I00-I99: Diseases of the circulatory system
- I20-I25: Ischemic heart diseases
- I25: Chronic ischemic heart disease
- I25.1: Atherosclerotic heart disease of native coronary artery
- I25.11: Atherosclerotic heart disease of native coronary artery with angina pectoris
- I25.110 Atherosclerotic heart disease of native coronary artery with unstable angina pectoris

Construction of the medical embedding

We performed a non-supervised approach to identify relevant clinical clusters of SCD, using NLP and clustering methods. A key challenge was to create a meaningful representation of patients based on massive amounts of electronic health records, comprising both structured and unstructured information. For this task, we used word embedding methods, which have become a major reference to tackle the issue of medical concepts representation [Li et al., 2022]. We treated each medical code (outpatient drug or hospital diagnosis) as a word, such that a patient can be represented by a sentence whose number of words is equal to the number of medical events that occurred before SCD.

We built a model that transforms all medical codes to numerical vectors of fixed dimensionality, whose relative geometrical positions reflect medical proximities. Words that co-occurred more frequently should be close together in the embedding space. For instance, antidepressant drugs (ATC N06) are expected to be close to mood affective disorders (ICD-10 F30-F39), and beta-blocking agents (ATC C07) should be close to hypertensive diseases (ICD-10 I10-I16). But this unsupervised approach also allows to learn new medical relationships, that could be useful to find unknown groups of patients and to highlight the heterogeneity of SCD, especially for non-cardiovascular risk factors.

We used the Skip-gram architecture of the Word2Vec algorithm [Mikolov et al., 2013], a neural network-based approach, to exploit the co-occurrence information of the medical trajectories. It consists of training a shallow neural network to predict the nearest neighboring words within the context window of a single input word. We fed it with all combinations of medical codes and associated surrounding contexts observed in the medical trajectories before SCD. The medical embedding space then corresponds to the hidden representation (weights of the input layer) learned by the model at the end of the training. Each medical code is finally represented by a vector of length 100.

We used the open-source Python library Gensim (see <https://radimrehurek.com/gensim> version 4.2.0) to train our model. We followed the default options proposed by the authors for the hyperparameters, except for the min count parameter (min count = 1) and for the window size, which corresponds to the mean number of medical codes observed within a 3 month time window among all medical trajectories (window = 186).

Patients with no hospital diagnoses or outpatient drugs before SCD were not used to feed the model and therefore were excluded from the whole clustering analysis. Information on the occurrence, management, and patient outcomes of SCD, as well as socio-demographic information were not used to build the medical embedding space.

Representation of the patients in the medical embedding space

Once the medical embedding space had been obtained, each patient was mapped to this space and represented by a vector of length 100, by computing the mean of vectors corresponding to its medical events occurred before SCD. Patient's vectors therefore summarize their temporal information, such that two patients who have similar medical trajectories are expected to be close to each other in the embedding space.

Clustering analysis

The patients' vectors computed from the medical embedding space were used to perform a clustering analysis, in order to find subgroups of SCD with homogeneous clinical characteristics. We used the K-Means method [Arthur and Vassilvitskii, 2007] available in the open-source Python library Scikit-Learn (see <https://scikit-learn.org/>, version 1.1.3). This method identifies homogeneous subgroups, such that patients in each cluster are as similar as possible according to the Euclidean distance. We applied the Elbow method to find the optimal clustering setting. We varied the number of clusters between 2 and 20 and selected the one above which the total intra-class variance did no longer improve [Satopaa et al., 2011].

Visualization of the clusters

We visualized the clusters of patients in a 2-dimensional space, using the t-distributed stochastic neighbor embedding algorithm [Van der Maaten and Hinton, 2008] available in the Python library Scikit-Learn. This method first computes a probability distribution over the vectors of patients in their original 100-dimensional embedding space, and assigns random coordinates for each patient in the target 2-dimensional map. These coordinates are then iteratively updated based on the objective to minimize the Kullback-Leibler divergence between the probability distribution computed in the high- and low-dimensional spaces. Patients who are close to each other in the t-SNE map can be considered more similar than others.

Evaluation and validation of the clusters

The evaluation of the clusters was based on hospital diagnoses and outpatient drugs that occurred up to 5 years before SCD. For each cluster, we identified the medical codes that are under or overrepresented, compared to their average occurrence in the whole population, using the Cramer's V test score. Cluster labels were then assigned by examining the most under or overrepresented medical codes. For each cluster, we also described information that was not used to build the clustering model, including age, sex, universal healthcare coverage, and characteristics of SCD. In order to assess the temporal validity of our approach, we finally applied our model (Word2Vec and K-Means algorithms) trained with the 2011-2015 SDEC Registry, to SCD cases collected between 2016 and 2020 in the same geographic area.

3.3 Results

The Paris Sudden Death Expertise Center Registry and the French National Health Insurance Database

In this study, we combined data from a large population-based registry on SCD with the French National Health Insurance Database (SNDS). Every case of SCD in persons older than 18 years that occurred between 16 May 2011 and 31 December 2020 in Paris (France) and its inner suburbs (Hauts-de-Seine, Seine-Saint-Denis, Val-de Marne) was included. The cases were collected throughout the Paris Sudden Death Expertise Center (SDEC), a multidisciplinary consortium dedicated to research, education and care of SCD2. For all SCD cases, we collected data from the SNDS database, which provides information on healthcare expenses, on an individual level, for all people affiliated to an insurance scheme in France. The SDEC Registry and the SNDS database are described more fully in the Methods section.

First, we analyzed hospital diagnoses and outpatient drugs that occurred up to 15 years before SCD in the derivation cohort (2011-2015 SDEC population). This population includes 12,189 SCD patients (60.1% men, mean age 69.5 ± 18 years) who had at least 1 hospital diagnosis or 1 outpatient drug up to 15 years before SCD. Among them, 2,403 (19.8%) SCD occurred in a public area. A bystander was present in 8,528 (70.8%) of cases and performed cardiopulmonary resuscitation (CPR) in 4,432 (50.4%) of cases. An initial shockable rhythm was performed in 1,802 (16.5%) patients and 2,714 (22.3%) patients were transported alive to the hospital and 621 (5.1%) patients survived at hospital discharge. We performed a non-supervised statistical approach to identify relevant clinical clusters of SCD in the derivation cohort based on their medical trajectories. A methodological overview of the study is provided in Figure 3.1. Our clustering analysis was then validated on the 2016-2020 SDEC registry to assess the temporal transportability of the results. The validation cohort (2016-2020 SDEC population) included 11,485 SCD patients (60.3% men, mean age 71.9 ± 17 years) who had at least 1 hospital diagnosis or 1 outpatient drug up to 15 years before SCD. Among these 11,485 SCD, 1,835 (16.0%) SCD occurred in a public area. A bystander was present in 6,941 (62.0%) of cases and performed CPR in 4,810 (71.0%) of cases. 1,846 (21.3%) had an initial shockable rhythm, 2,336 (20.3%) patients were transported alive to the hospital and 589 (5.1%) patients survived at hospital discharge.

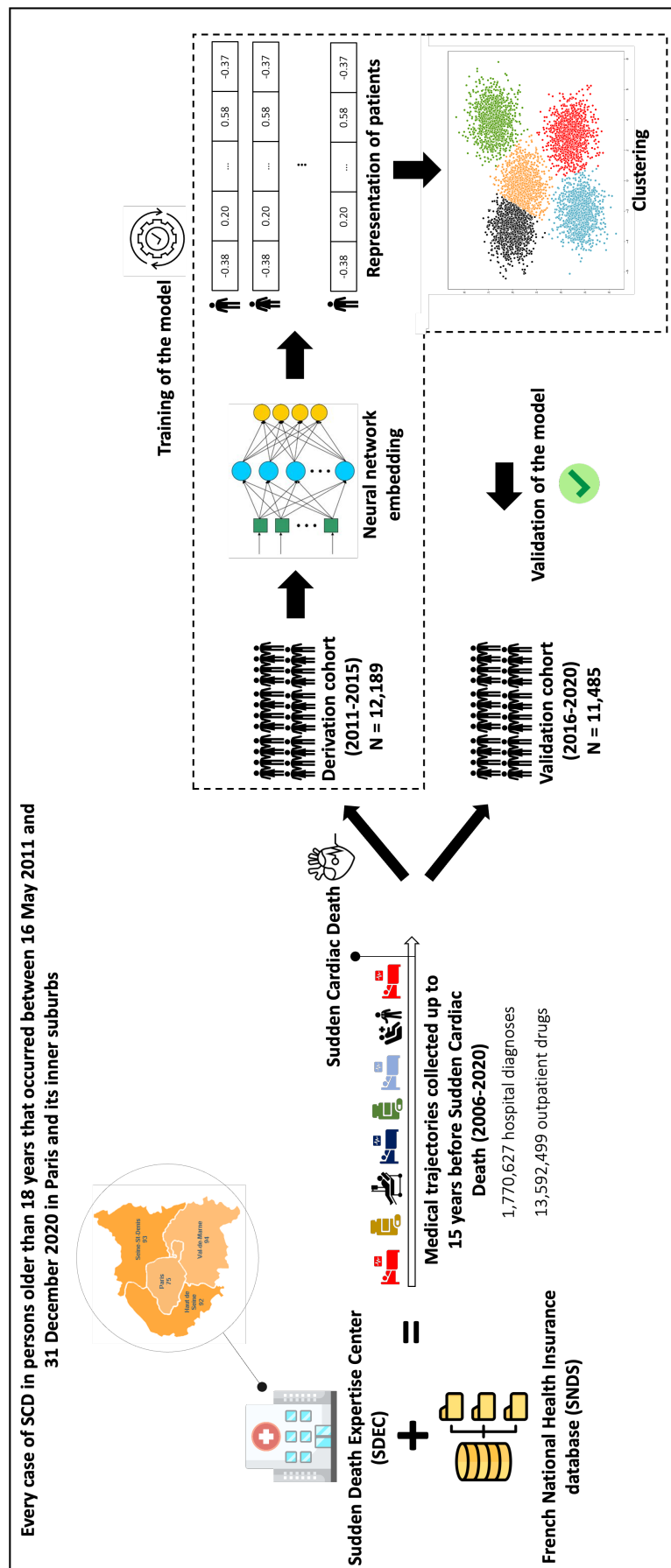


FIGURE 3.1: Methodological overview of the clustering study

Clustering analysis

We used natural language processing (NLP) and clustering methods to identify clusters of SCD based on 676,930 hospital diagnoses and 5,461,318 outpatient drugs that occurred up to 15 years before SCD for the 12,189 patients in the derivation cohort. We first applied the Word2Vec algorithm [Mikolov et al., 2013], a neural network-based embedding model, to exploit the co-occurrence information of medical trajectories before SCD. Each medical code was treated as a word, such that a patient can be represented by a sentence whose number of words is equal to the number of medical events that occurred before SCD. Our model transformed all medical codes into numerical vectors of length 100 using the Skip-gram architecture of Word2Vec. This approach builds a new representation of medical knowledge, in which relative geometrical positions reflect medical proximities between diseases and treatments. Information on the occurrence, management, and patient outcomes of SCD, as well as socio-demographic information were therefore not used to build the medical embedding space.

Once the medical embedding space had been obtained, each patient was mapped to this space, by computing the mean of vectors corresponding to its medical events. The patients' vectors were then used to perform a clustering analysis in order to find clusters of SCD with homogeneous clinical characteristics. We used the K-Means method and applied the Elbow method to find the optimal clustering setting. Eight distinct homogeneous clusters of SCD were identified in the 2011-2015 SDEC population. We visualized the clusters of patients in a 2-dimensional space, using the t-distributed stochastic neighbor embedding algorithm (t-SNE) [Van der Maaten and Hinton, 2008] (see Figure 3.2). Patients who are close to each other in the t-SNE map can be considered more similar than others. We identified 3 large central groups and 5 small peripheral groups (Figure 3.2). Baseline characteristics of each cluster are presented in Table 3.5). Figures 3.3 and 3.4 describe the hospital diagnoses and outpatient drugs that were under or overrepresented in each cluster up to 5 years before SCD. We assigned a label to each cluster based on these clinical determinants, age, sex, universal healthcare coverage and characteristics of SCD.

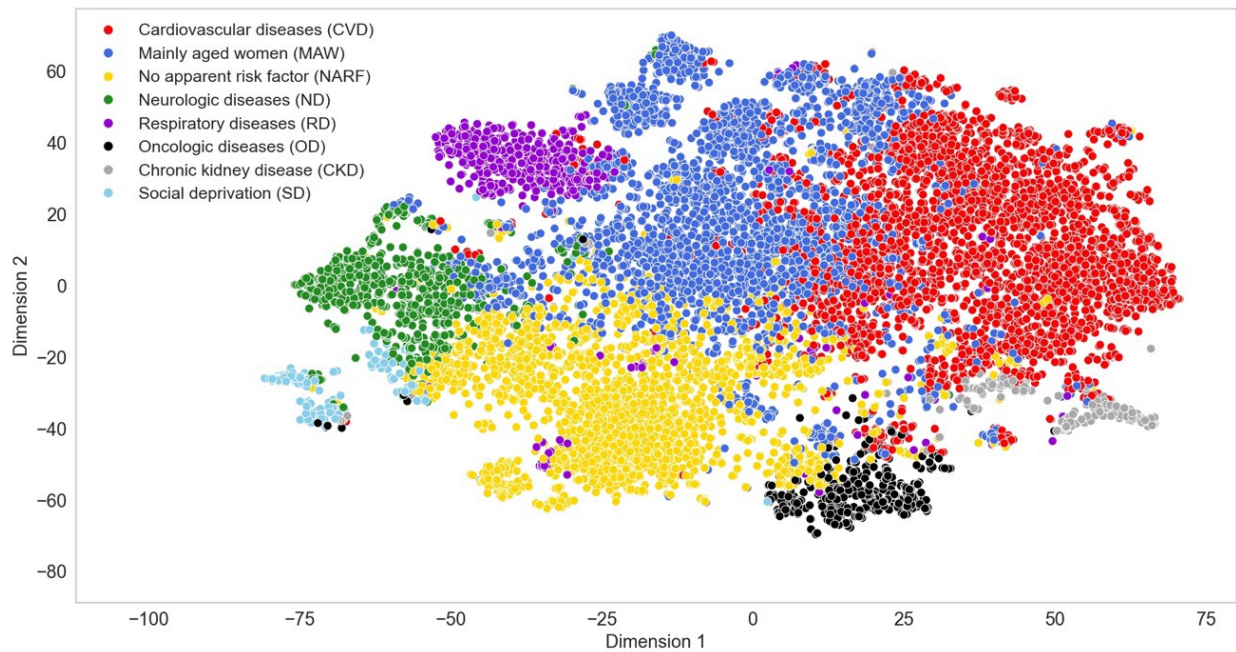


FIGURE 3.2: Visualization of the clusters (derivation cohort)

Cluster 1, including 3,638 (30%) of the 12,189 patients, with mainly circulatory system diseases and cardiovascular drugs, located at the right side of the map was the largest group identified in the SCD population (Figures 3.2, 3.3 and 3.4). Patients of this cluster were older (74.1 years vs. 69.5 years, $p < 0.001$) and over-represented by men (74.3% vs 60.1%, $p < 0.001$) as compared to the total population. It was labelled as SCD cluster with cardiovascular diseases (CVD) because of the old age, the over-representation of men and the cardiovascular comorbidities and drugs associated with this cluster. Among cardiovascular diseases, coronary heart disease was already known to be the most common pathology underlying SCD, followed by cardiomyopathies, inherited arrhythmia syndromes, and valvular heart disease [Hayashi et al., 2015].

Cluster 2, including 3,307 (27%) patients, was located in the center of the map. Patients of this cluster were older (81.3 y. vs 69.5 y, $p < 0.001$) and over-represented by women (62.0% vs. 39.9%, $p < 0.001$) as compared to the total population. It was labelled as SCD cluster with mainly aged women (MAW). Previous studies have shown that women are considerably older than men at the time of SCD. They have more often nonischemic causes such as primary myocardial fibrosis. Women are also more likely to have a normal electrocardiogram (ECG) prior to SCD than men, but more women have ECG markers of left ventricular hypertrophy than do men who have SCD [Haukilahti et al., 2019].

Cluster 3, including 2,645 (22%) patients, composed of young subjects (54.8 y. vs. 69.5 y, $p < 0.001$) over-represented by men (64.6% vs. 60.1%, $p < 0.001$) who had few comorbidities and drugs before SCD, was located under the MAW cluster and at the left side of the CVD cluster. It was labelled as SCD cluster with no apparent risk factor (NARF) since the subjects do not have the usual cardiovascular risk factors, conditions and related drugs. It is usually very difficult to identify these high risk subjects before the occurrence of their SCD. Here, our approach, collecting micro signals over the 15 years before the event permitted their identification. Since they represent more than one fifth of the total number of SCD, the present result is of peculiar importance and will require additional works. The NARF cluster, like the CVD cluster, had the most frequent proportion of patients with an initial

shockable rhythm (n = 606, 25%) and was associated with a better survival rate (n = 234, 8.8%, $p < 0.001$). SCD in young athletes < 35 years old can match well with SCD with NARF subjects. Hypertrophic cardiomyopathy is the most common cardiac abnormality implicated in SCD in young athletes < 35 years old in the United States followed by congenital anomalies of the coronary artery and arrhythmogenic right ventricular cardiomyopathy. Hence, athletic pre participation screening has been made essential for minimizing the risk for SCD in them [Kumar et al., 2021].

On the periphery of these 3 big clusters, the model identified 5 small and lesser known clusters.

Cluster 4, including 862 (7%) patients, was associated with mental, behavioral, and neurodevelopmental disorders and with drugs of the nervous system. Patients of this cluster were younger (54.4 y. vs. 69.5 y., $p < 0.001$) as compared to the total population. It was labelled as SCD with neurologic diseases (ND). Some studies have suggested that clinical depression, phobic anxiety and antipsychotic drugs may be associated with a higher risk of SCD independently of established coronary heart disease risk factors [Empana et al., 2006, Albert et al., 2005, Whang et al., 2009, Weeke et al., 2014]. Other studies have suggested that acute stroke can disturb central autonomic control, resulting in myocardial injury, electrocardiographic abnormalities, cardiac arrhythmias, and ultimately sudden death^{21,22}. Similarly, myocardial damage and an increase of troponins and Takotsubo syndrome have been reported in humans suffering from epileptic seizures^{23–25}. Finally, McMillan and Teasdale reported a high incidence of sudden death in humans who had mild traumatic brain injury years ago [McMillan and Teasdale, 2007].

Cluster 5, including 688 (6%) patients, was associated with diseases and drugs of the respiratory system, while Cluster 6, including 518 (4%) patients, was characterized by tumors and their associated medical codes, without one cancer being identified more than another. These two clusters were well separated from the other ones and located at the top of the map for Cluster 5 and at the bottom of the map for Cluster 6. Cluster 5 was labelled as SCD with respiratory diseases (RD) and Cluster 6 as SCD with oncologic diseases (OD). Indeed, cumulating evidence associates chronic obstructive pulmonary disease with an increased risk of SCD both in cardiovascular patient groups and in community-based studies, independent from cardiovascular risk profile. Underlying mechanisms explaining this association require further investigation [Empana et al., 2006, Van den Berg et al., 2016]. Similarly, patients with bronchial asthma may die unexpectedly and with no obvious cause for the severity of this process [Robin and Lewiston, 1989]. Finally, cardiac involvement and fibrosis in sarcoidosis occur in 5-10% of cases and lead to congestive heart failure, arrhythmias and sudden cardiac death [Markatis et al., 2020]. In cancer patients, SCD differs significantly when compared to non-cancer patients. Coronary events are less prominent whereas respiratory causes (pulmonary embolism and hypoxia) are common etiologies in cancer³⁰. There is variability in the incidence of corrected QT prolongation of various cancer drugs (0%-22%); however, the clinical consequence, as defined by arrhythmias or sudden cardiac death, remains rare [Weeke et al., 2014, Porta-Sanchez et al., 2017]. As for primary defibrillator therapy in patients with cancer, its relative benefit is limited because of competing risk of nonarrhythmic mortality and a personalized cardiologic and oncologic coevaluation is needed [Itzhaki Ben Zadok et al., 2023].

Cluster 7, including 279 (2%) patients, was associated with kidney diseases and over-represented by men (72.4% vs. 60.1%, $p < 0.001$). Indeed, chronic kidney disease (CKD) patients demonstrate an increased incidence of SCD with declining kidney failure, mainly from

cardiovascular causes³³. Current evidence suggests that coronary artery disease (CAD) associated risk factors may play a lesser role in CKD patients. Complex relationships between CKD-specific risk factors, structural heart disease, and VA contribute to the high risk of SCD. In dialysis patients, the occurrence of VA and SCD could be exacerbated by electrolyte shifts, divalent ion abnormalities, sympathetic overactivity, inflammation and iron toxicity [Di Lullo et al., 2016].

Finally, Cluster 8, including 252 (2%) patients, composed of young subjects (51.0 y. vs. 69.5 y, $p < 0.001$) over-represented by men (82.5% vs. 60.1%, $p < 0.001$), was associated with mental, behavioral, and neurodevelopmental disorders (like in the ND cluster), but also with infectious diseases including human immunodeficiency virus infection and socioeconomic disadvantages. We found that 21.1% of patients in this cluster had access to the universal healthcare coverage (vs. 6.1% in the total population, $p < 0.001$). It was labelled as SCD with social deprivation (SD) because of the young age, the high rate of subjects with universal healthcare coverage, the over-representation of men and the comorbidities associated with this cluster. Lower socioeconomic status, depression, anxiety, social isolation, and psychological stress have all been linked to an increase in cardiovascular mortality in diverse populations [Mensah et al., 2005, Rozanski et al., 1999]. As for HIV-infected patients³⁷, there is biologic plausibility that the following mechanisms may be contributing to the significantly heightened risk of sudden cardiac death in HIV to varying degrees: VA, myocardial fibrosis and scar, prolonged corrected QT interval (both as a direct effect of HIV on repolarization as well as a result of concurrent medications/antiretroviral therapies), substance abuse, structural heart disease, and premature atherosclerosis. Further studies are needed to assess the relative contribution of each of these mechanisms and risk factor.

SCD occurred in a public place more frequently for the SD, NARF and ND clusters (39.4% $p < 0.001$, 32.1% $p < 0.001$ and 31.0% $p < 0.001$ respectively) as compared to the total population (19.8%) (Table 1a). Bystander presence was almost the same in all clusters except for the SD and ND clusters in which bystanders were less present when cardiac arrest occurred ($p < 0.001$). Bystander CPR was almost the same in all clusters except for the MAW and OD clusters where CPR was less performed by bystanders.

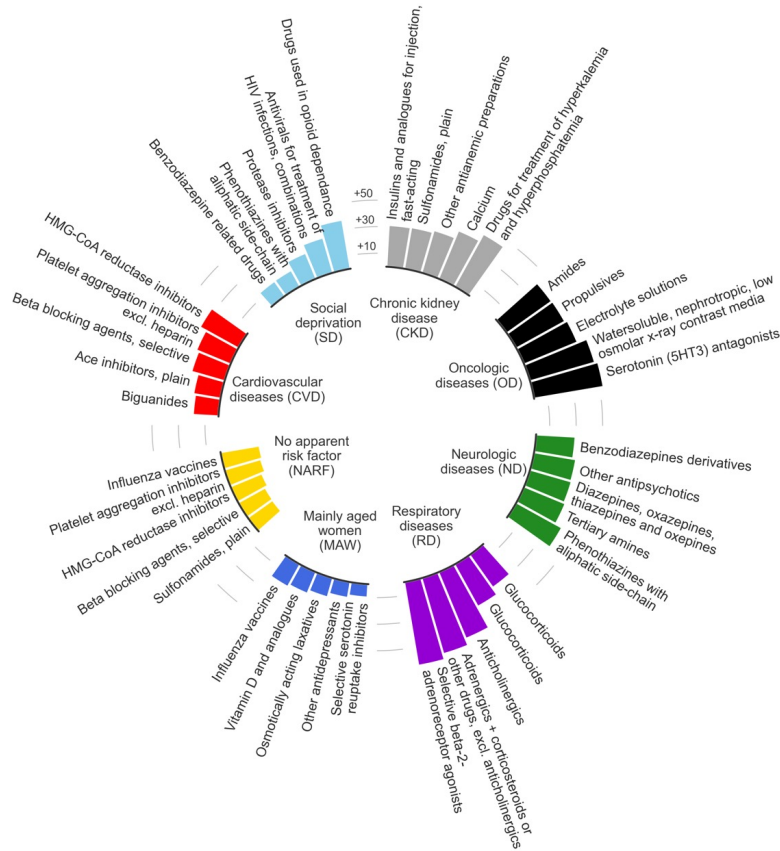


FIGURE 3.3: Main outpatient drugs in each cluster (derivation cohort)

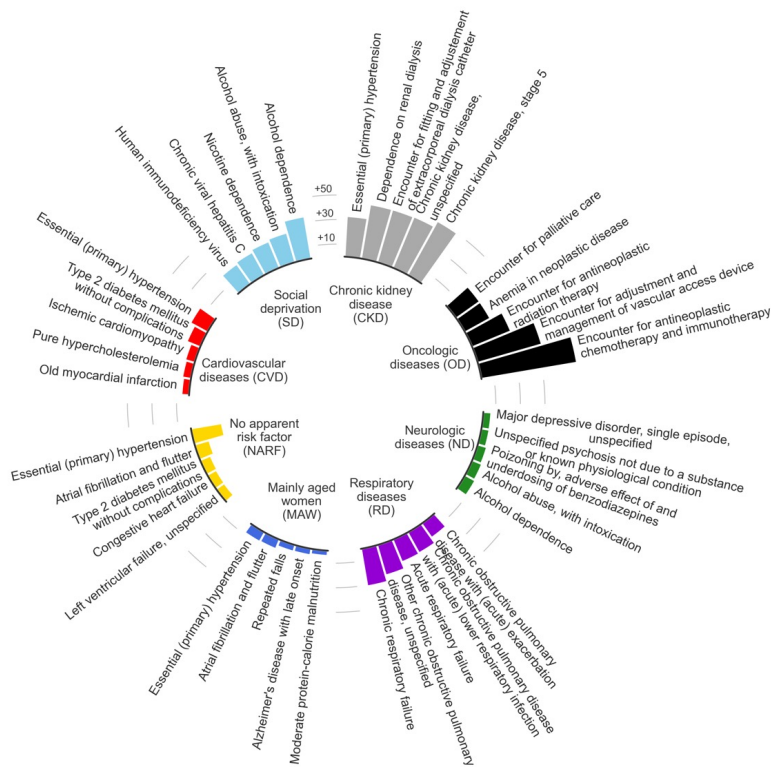


FIGURE 3.4: Main hospital diagnoses in each cluster (derivation cohort)

	Total	Cardiovascular disorders (CVD)	Mainly aged women (MAW)	No apparent risk factor (NARF)	Neurologic disorders (ND)	Respiratory diseases (RD)	Oncologic diseases (OC)	Chronic kidney disease (CKD)	Social deprivation (SD)
N (%)	12,189	3,638 (30%)	3,307 (27%)	2,645 (22%)	862 (7%)	688 (6%)	518 (4%)	279 (2%)	252 (2%)
Sociodemographic characteristics									
Age, y, mean (SD)	69.5 (18)	75.2 (12)	81.3 (12)	54.8 (17)	54.4 (16)	71.2 (16)	62.9 (12)	65.6 (16)	51.0 (11)
Men, n (%)	7,323 (60.1%)	2,697 (74.1%)	1,257 (38.0%)	1,709 (64.6%)	492 (57.1%)	436 (63.4%)	322 (62.2%)	202 (72.4%)	208 (82.5%)
UHC, n (%)	745 (6.1%)	119 (3.3%)	78 (2.4%)	296 (11.2%)	108 (12.5%)	36 (5.2%)	41 (7.9%)	14 (5.1%)	53 (21.1%)
SCD characteristics									
SCD location: public area, n (%)	2,403 (19.8%)	641 (17.7%)	338 (10.3%)	847 (32.1%)	267 (31.0%)	108 (15.7%)	47 (9.1%)	56 (20.1%)	99 (39.4%)
Bystander, n (%)	8,528 (70.8%)	2,664 (74.0%)	2,314 (70.9%)	1,830 (70.2%)	531 (62.3%)	479 (70.4%)	355 (69.7%)	204 (73.9%)	151 (61.1%)
CPR bystander, n (%)	4,432 (50.4%)	1,401 (51.2%)	1,092 (45.9%)	1,039 (55.7%)	292 (50.5%)	251 (51.4%)	154 (41.5%)	107 (50.0%)	96 (58.9%)
Initial rhythm: VT/VF, n (%)	1,802 (16.5%)	700 (21.3%)	271 (9.3%)	606 (25.0%)	61 (7.8%)	60 (9.7%)	31 (6.7%)	39 (15.6%)	34 (15.3%)
Outcome									
Transported alive, n (%)	2,714 (22.3%)	811 (22.3%)	452 (13.7%)	870 (32.9%)	188 (21.8%)	162 (23.5%)	68 (13.1%)	75 (26.9%)	88 (34.9%)
Coronary angiogram, n (%)	1,397 (49.6%)	487 (59.8%)	200 (44.3%)	478 (50.4%)	57 (28.8%)	74 (44.0%)	24 (34.8%)	40 (53.3%)	37 (41.6%)
Survival at hospital discharge, n (%)	621 (5.1%)	191 (5.3%)	91 (2.8%)	234 (8.8%)	22 (2.6%)	37 (5.4%)	15 (2.9%)	11 (3.9%)	20 (7.9%)

FIGURE 3.5: Socio-demographics, SCD characteristics and outcomes of patients in the total derivation cohort and in each cluster

Validation of the model

In order to assess the temporal validity of the results, we validated our model on the 2016-2020 SDEC population. 5,751,323 outpatient drugs and 979,639 hospital diagnoses were collected for the 11,485 SCD patients of the validation cohort up to 15 years before SCD. These patients were represented with the medical embedding space obtained from the derivation cohort. They were then classified into the 8 clusters described in the previous section. We evaluated the stability of the clusters by comparing the results obtained in the 2 cohorts. We found that the clusters were located exactly at the same place in the 2-dimensional space (Figure 3.6), except for the ND and SD clusters. These 2 clusters were on the left of the NARF cluster in the derivation cohort although they were found below in the validation cohort. We also found that baseline characteristics (Table 1b) and under or overrepresented medical codes (Figures 3b and 4b) were very similar to those obtained in the derivation cohort. As in the derivation cohort, survival was the highest in NARF cluster and the lowest in MAW cluster.

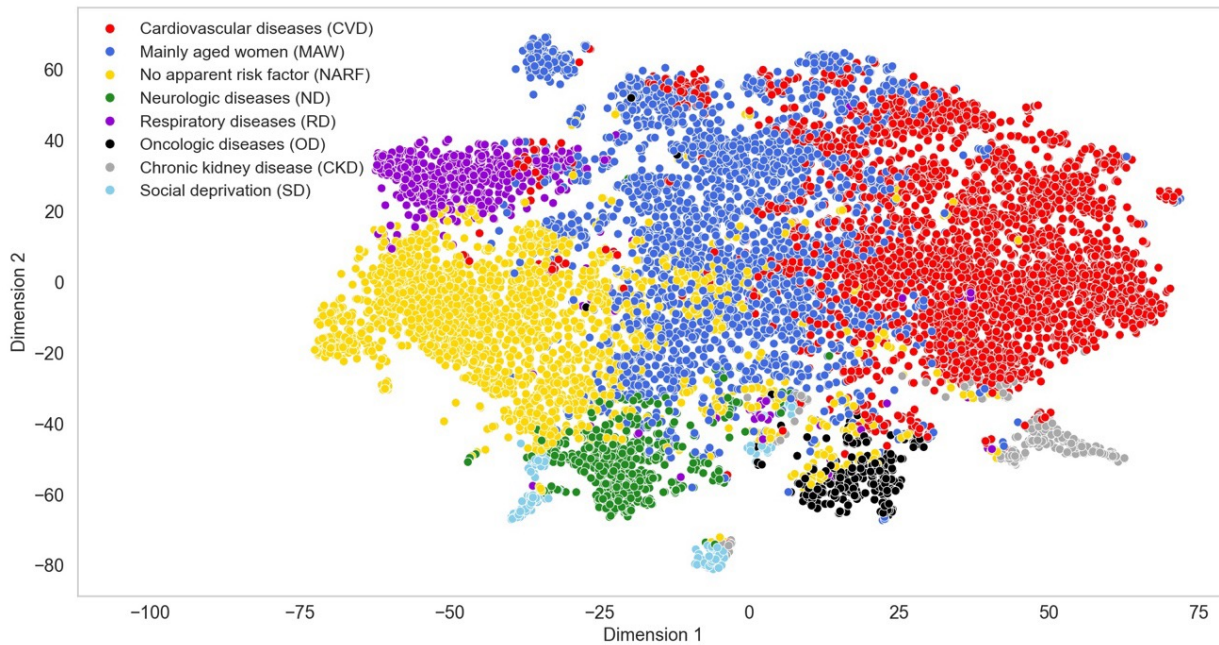


FIGURE 3.6: Visualization of the clusters (validation cohort)

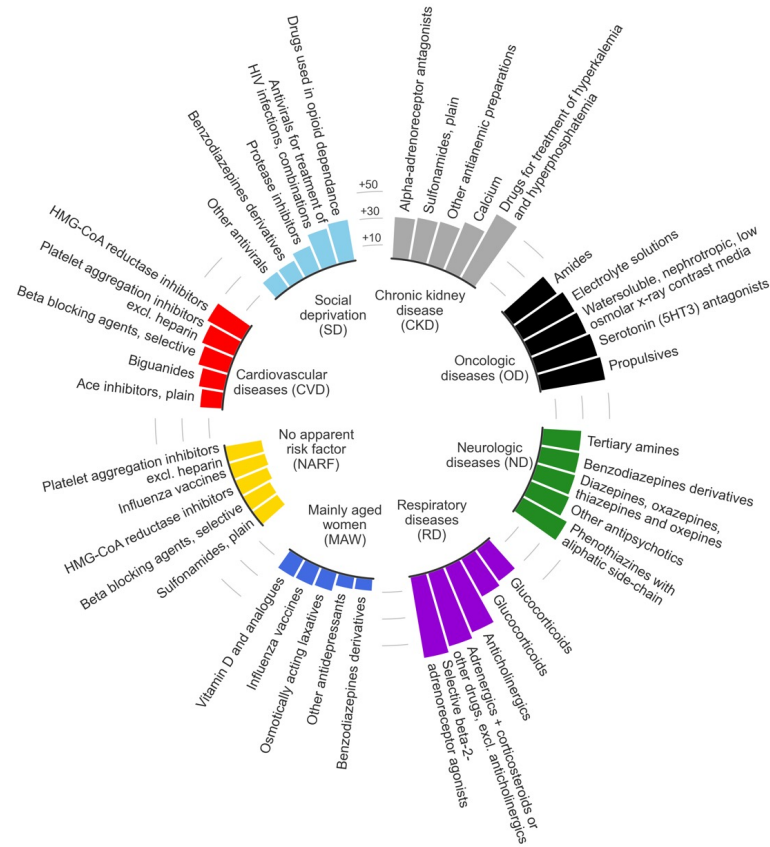


FIGURE 3.7: Main outpatient drugs in each cluster (validation cohort)

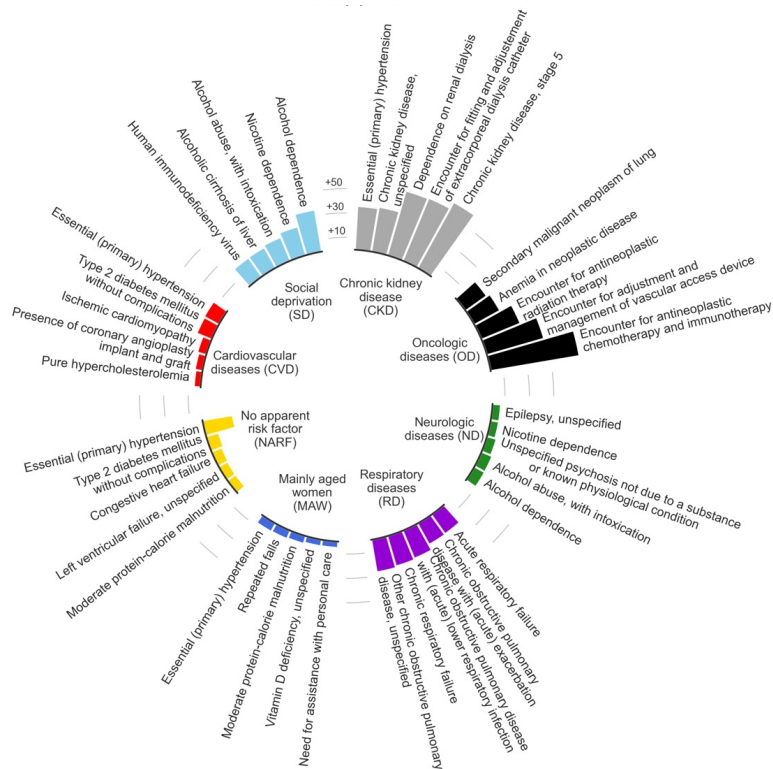


FIGURE 3.8: Main hospital diagnoses in each cluster (validation cohort)

	Total	Cardiovascular disorders (CVD)	Mainly aged women (MAW)	No apparent risk factor (NARF)	Neurologic disorders (ND)	Respiratory diseases (RD)	Oncologic diseases (OC)	Chronic kidney disease (CKD)	Social deprivation (SD)
N (%)	11,485	3,703 (32%)	3,097 (27%)	2,533 (22%)	603 (5%)	666 (6%)	360 (3%)	307 (3%)	216 (2%)
Sociodemographic characteristics									
Age, y, mean (SD)	71.9 (17)	77.2 (12)	82.2 (12)	58.9 (17)	56.7 (15)	71.3 (14)	64.1 (13)	66.6 (15)	53.1 (10)
Men, n (%)	6,930 (60.3%)	2,707 (73.1%)	1,224 (39.5%)	1,618 (63.9%)	345 (57.2%)	408 (61.3%)	230 (63.9%)	218 (71.0%)	180 (83.3%)
UHC, n (%)	742 (6.5%)	153 (4.1%)	98 (3.2%)	275 (10.9%)	64 (10.6%)	54 (8.1%)	24 (6.7%)	24 (7.8%)	50 (23.5%)
SCD characteristics									
SCD location: public area, n (%)	1,835 (16.0%)	543 (14.7%)	308 (10.0%)	641 (25.3%)	123 (20.4%)	81 (12.2%)	45 (12.6%)	42 (13.7%)	52 (24.2%)
Bystander, n (%)	6,941 (62.0%)	2,305 (63.8%)	1,806 (60.1%)	1,620 (65.3%)	286 (49.0%)	405 (62.2%)	220 (63.4%)	198 (65.8%)	101 (47.9%)
CPR bystander, n (%)	4,810 (71.0%)	1,573 (69.7%)	1,197 (68.4%)	1,196 (75.8%)	210 (75.3%)	263 (66.1%)	148 (68.8%)	151 (77.0%)	72 (72.0%)
Initial rhythm: VT/VF, n (%)	1,846 (21.3%)	636 (22.7%)	271 (12.7%)	694 (33.3%)	43 (9.2%)	72 (14.0%)	42 (16.1%)	58 (23.8%)	30 (17.3%)
Outcome									
Transported alive, n (%)	2,336 (20.3%)	675 (18.2%)	362 (11.7%)	820 (32.4%)	127 (21.1%)	158 (23.7%)	55 (15.3%)	84 (27.4%)	55 (25.5%)
Coronary angiogram, n (%)	1,365 (59.2%)	434 (65.8%)	186 (53.6%)	522 (62.8%)	58 (47.2%)	72 (45.9%)	24 (45.3%)	41 (52.6%)	28 (50.0%)
Survival at hospital discharge, n (%)	589 (5.1%)	165 (4.5%)	79 (2.6%)	265 (10.5%)	18 (3.0%)	18 (2.7%)	17 (4.7%)	17 (5.5%)	10 (4.6%)

FIGURE 3.9: Socio-demographics, SCD characteristics and outcomes of patients in the total validation cohort and in each cluster

3.4 Discussion

For the first time to our knowledge, AI combined with an agnostic approach allowed the identification of new clusters of SCD in the general population, using unselected cardiovascular and non-cardiovascular variables, providing finally a global picture of SCD subjects. Moreover, our approach permitted to identify a large group of relatively young subjects without known cardiovascular risk factors which were very difficult to identify until now. Considering the heterogeneity between the groups helps us to understand why it was so difficult until now to find the causes of SCD with the classical approach restricted to cardiovascular variables and conditions. The use of AI was necessary but overall it was the choice to open to unselected medical variables that permitted to provide this global picture. This was made possible thanks to a widely available, low-cost, exhaustive population-based data pertaining to the large Paris Sudden Death Expertise Center registry and the French National Health Insurance Database.

The validation of our model on a more recent population allows us to ensure that our AI technology did not inadvertently incorporate bias. However, careful attention should be paid to the social context in which the data have been collected; our results based on French data may not be generalized elsewhere. We believe further deliberations on the lesser known clusters identified in this study may eventually have practical implications, help in guiding management decisions, tailoring and targeting early treatment to patients who would benefit most, and most probably improve patient outcome. Further investigations within each cluster will require the involvement of other medical specialties in addition to cardiology. By extending far beyond cardiovascular pathology, our approach provides a global picture of SCD that might eventually lead to discover new pathways and help identifying high risk subjects who need specific individualized preventive strategies.

Chapter 4

Personalized Prediction Model of SCD in the General Population

Contents

4.1	Introduction	93
4.2	Methods	94
4.3	Results	97
4.4	Discussion	108
4.5	Conclusion	109

This Chapter comes from a joint work with Wulfran Bougouin, Frankie Beganton, Jean-Philippe Empana, Nicolas Chopin and Xavier Jouven, and will be submitted in The Lancet Journal in May 2023.

4.1 Introduction

Sudden Cardiac Death (SCD) remains a major public health challenge worldwide, accounting for 10% to 20% of all deaths in industrialized countries [Chugh et al., 2008, Zeppenfeld et al., 2022]. Resuscitation is difficult, and despite decades of research, prognosis remains poor, with survival after SCD below 10%. Considering disappointing results of recent therapeutic trials, a paradigm shift toward SCD prevention is essential from a public health perspective. Efficient tools for SCD prevention are available, such as implantable cardioverter defibrillators (ICD) which have proven their efficacy for both secondary and primary prevention. However, identifying the best candidates for ICD implantation remains challenging.

Some very-high risk patients (survivors of SCD, high-risk cardiomyopathies) are clearly identified for ICD implantation in both European and American guidelines [Zeppenfeld et al., 2022]. However, these patients only account for a small proportion of SCD burden, whereas most cases arise from the general population without any prior known heart disease [Myerburg and Junttila, 2012]. Identification of patients at risk of SCD as their first cardiac event remains a difficult challenge to address, with somehow disappointing results so far. Several prediction models for SCD in the general population have been proposed [Deo et al., 2016, Waks et al., 2016, Aro et al., 2017, Bogle et al., 2018, Holkeri et al., 2020], but they focused only on cardiac-related risk factors, and are not designed for individual-level prediction. Improving their broad applicability in the population would require a comprehensive assessment of medical history, including both cardiovascular and non-cardiovascular conditions. In addition, it should be combined with a new methodological approach to provide personalized and explainable risk score for each patient. Finally, optimal model should not

only offer high discrimination but also adequate calibration to identify accurately the top deciles of predicted risk, which could benefit for specific preventive strategies.

To this end, we developed and validated a population-based model of SCD prediction, using machine learning algorithms, and large-scale data analysis of electronic health records of every SCD occurred in Greater Paris (10% of the French population) during 10 years.

4.2 Methods

Study Design

This study was designed as a retrospective study for the development and validation of a clinical prediction model. Data collection and analyses were conducted in accordance with all relevant French regulatory requirements. Access to the French National Health Insurance database is regulated by the Committee of Expertise for Research, Studies and Evaluations in the field of Health (CEREES) and the French National Data Protection Agency (CNIL). The study protocol was approved in 2016 for SCD cases collected between 2011 and 2015 (CEREES approval N°12-336; CNIL authorization DR-2016-401) and approved in 2022 for SCD cases collected between 2016 and 2020 (CEREES approval N°2785673). The Paris Sudden Death registry was also approved by the CNIL (authorization DR-2012-445). In accordance with the regulations in force, informed patient consent was not required due to the retrospective and observational nature of the study. We used the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [Moons et al., 2015] for reporting the development and validation of the prediction model.

Population

Sudden Cardiac Death Cases

Every case of unexpected out-of-hospital cardiac arrest in persons older than 18 years that occurred between 16 May 2011 and 31 December 2020 in Paris (France) and its inner suburbs (Hauts-de-Seine, Seine-Saint-Denis, Val-de Marne) was collected throughout the Paris Sudden Death Expertise Center (SDEC). The SDEC registry is a multicenter population-based registry system covering a population of 6.7 million inhabitants (10% of the French population) [Bougouin et al., 2014, Maupain et al., 2016, Jabre et al., 2016, Bougouin et al., 2018, 2020]. It records prospectively and continuously information on the occurrence (Utstein criteria), management (pre- and in hospital) and patient outcomes (survival and neurological outcomes) of all SCD cases. Exclusion criteria of the SDEC registry are a prior terminal condition, no attempt at advanced cardiac life support by emergency medical system personnel or an obvious non-cardiac cause according to Utstein templates for resuscitation registries reporting data on cardiac arrest [Jacobs et al., 2004, Perkins et al., 2015]. Our study therefore included only cases who experienced SCD.

To ensure the completeness of collection in the area, an intensive and prospective epidemiological case-ascertainment programme was applied, involving all components of the emergency medical system. We performed a retrospective control on a sample of 3 intensive care units and the SDEC registry detected 99% of SCD cases admitted alive in this sample [Bougouin et al., 2014]. In addition, each case was independently reviewed by two investigators of the SDEC to ensure the accuracy of classification and to avoid over-estimation often encountered in retrospective data collection.

Controls

SCD cases included in this study were matched by age, sex and residence area with a control group which had no SCD and was randomly sampled from the French general population, using the French National Health Insurance database. One control was selected for each case of SCD through an individual case-control matching. The endpoint for the control group was the day on which SCD occurred among their corresponding cases. Exclusion criteria from the sampling procedure were individuals already included in the SDEC registry, aged under 18 years or who lived outside the area of interest.

Myocardial Infarction

SCD cases collected between 2016 and 2020 were also matched by age, sex and residence area with a group of controls who had myocardial infarction, with no SCD, and who were identified using the French National Health Insurance database. The endpoint for this group was the day on which myocardial infarction occurred. As for the general population, one control was selected for each case of SCD through an individual case-control matching.

Data Sources

For SCD cases and controls, we collected data from the French National Health Insurance database (SNDS) [Moulis et al., 2015, Bezin et al., 2017, Tuppin et al., 2017]. The French Universal Health Insurance System manages all reimbursements of healthcare for all people affiliated to a health insurance scheme in France, covering 98% of the population (67 million inhabitants). It provides information on all healthcare expenses, on an individual level, including outpatient visits, procedures and drugs as well as information from hospital discharge summaries and chronic conditions. Data acquisition is permanent, from birth to death, irrespective of wealth, age, or work status, resulting in one of the largest electronic health records databases in the world. The data are anonymized but individually linked, which allows an individual longitudinal follow-up of participants over time.

The SNDS database links 2 main existing databases, the nationwide claims database of the French National Healthcare System (SNIIRAM) and the National Hospital database (PMSI). The SNIIRAM database contains exhaustive and individual data on outpatient healthcare reimbursements. It includes data on ambulatory care with all reimbursed drugs from community pharmacies and all reimbursed medical interventions. It also includes long-term diseases and chronic conditions as well as information about occupational accidents and diseases. The PMSI database is the national hospital discharge database, concerning both French public- and private-sector hospitals. Main data includes dates of admission and discharge, type of diagnoses, duration of stay as well as procedures (medical acts and biology) and especially costly drugs administered in hospital.

Hospital diagnoses are coded according to the International Classification of Diseases, 10th revision (ICD-10), which is a classification tool developed by the World Health Organization for epidemiology, health management and clinical purposes. Drugs are coded according to the Anatomical Therapeutic Chemical (ATC) system, that classifies drugs according to the organ or system on which they act and their therapeutic, pharmacological, and chemical properties. Results relating to biological tests and other medical procedures are not recorded, and medical indications are not specified for the reimbursed medical cares. Demographic (age, sex) and socioeconomic (affiliate insurance scheme, universal healthcare coverage and state medical assistance) information are available in the SNDS database. The

universal healthcare coverage is obtained for all individuals whose income is below a specific threshold and was used as a proxy variable for social deprivation in this study.

The SNDS database has been described in detail previously [Moulis et al., 2015, Bezin et al., 2017, Tuppin et al., 2017, Revet et al., 2022] and has been used to conduct multiple studies in cardiovascular epidemiology Giral et al. [2019], Piot et al. [2022], Lecoeur et al. [2023]. More details are available at <https://www.health-data-hub.fr>.

Model Development and Validation

We developed a 3-month prediction model of SCD in the general population, using a supervised learning classification model. The prediction model was derived on SCD cases and matched controls collected between 2011 and 2015 (derivation cohort), and was then validated on SCD cases and matched controls collected between 2016 and 2020 (validation cohort). To evaluate to which extent our approach was specific for SCD and not only related to risk factors for coronary atherosclerosis development, our model was also validated on cases of myocardial infarction, as sensitivity analysis.

We assessed whether machine learning approaches outperform standard statistical methods, and compared the Logistic Regression model with 3 ensemble methods (Random Forest, Extreme Gradient Boosting and CatBoost) which aggregate multiple learning algorithms to obtain more accurate and robust prediction. We also applied a Soft Voting Classifier which averages the prediction generated by the 4 aforementioned models.

To deal with overfitting in model selection, hyper-parameters and model settings were chosen using a cross validation on the derivation cohort. The derivation cohort was splitted in 10 non-overlapping subsets of equal size, such that all sets contained the same proportion of SCD cases. The models were then trained 10 times, and each time one of the subsets was left out from training to be used as a test set. The final performance of the cross-validation was given by the average of the 10 estimates in the test sets. Once the best model was selected, we trained it on the full derivation cohort and applied it on the validation cohort. All the models were trained with the open source Python library Scikit-Learn (see <https://scikit-learn.org/>, version 1.1.3).

The prediction model is based on outpatient drugs and hospital diagnoses that occurred up to 5 years before SCD. To investigate whether non-cardiovascular variables could enhance predictive performance beyond standard risk factors of SCD, we compared 2 different strategies for variable inclusion. The first approach (CVD model) includes medical codes that attempt to represent traditional risk factors for cardiovascular diseases, based on an exhaustive literature review of SCD prediction models. We selected 8 variables as surrogate markers for coronary artery disease, stroke, diabetes, hypertension, smoking status, obesity, lipid disorders and chronic renal failure (see Supplementary material, Table A.1). Other main risk factors, including left ventricular ejection fraction, electrocardiogram signals and blood pressure were unfortunately not available in the SDNS database. The second approach (EHR model) includes all medical codes that occurred up to 5 years before SCD, without any prior selection. However, given the infrequent occurrence of some drugs or diagnoses in the medical history of participants, we grouped the codes according to the third level of the ATC and ICD-10 classification systems. For instance, acebutolol (ATC C07AB04) was labeled as selective beta blocking agents (ATC C07AB), and atherosclerotic heart disease of native coronary artery (ICD-10 I25.1) was labeled as chronic ischemic heart disease

(ICD-10 I25).

We performed 3 data preparation steps in the model development. For each medical code, we first summed up the number of times they occurred up to 5 years prior to the outcome. We then normalized the data (by scaling each variable to the range of 0 and 1), and finally performed a variable selection to achieve a ratio of 5:100 between variables and SCD cases, to prevent overfitting. This variable selection was based on importance weights using a Gradient Boosting classifier. A methodological overview of the study is provided in Figure 1.

Model Evaluation

The evaluation of the models was based on discrimination and calibration. For each model, we evaluated the area under the receiver operating characteristic (AUC), positive predictive value (PPV) and sensitivity. We applied a bootstrap method to build empirical confidence intervals of 95% for AUC using 1,000 samples (with replacement) and 30 iterations. We selected the one that showed the highest average AUC score in the cross-validation. The calibration (i.e., the similarity between predicted risks of SCD and the actual outcomes) was evaluated with the histogram of predicted risk. We adopted the usual approach for binary outcomes of plotting decile-binned predictions on the x-axis and number of observed SCD cases and controls in each bin on the y-axis. We also gauged the calibration visually by inspecting how the calibration curve aligned with the diagonal line that represented perfect calibration. To assess the robustness of our approach, we finally conducted several sensitivity analyses and stratified the performance on different subgroups of the population (regarding age, sex and social deprivation).

Model Explanation

Once the prediction model was trained and optimized, we used the Shapley additive explanations (SHAP) algorithm [Lundberg and Lee, 2017] to explain how the variables relate to the predicted risk at the individual level. SHAP is a model-agnostic representation of variable importance where the impact of each variable on a particular prediction is represented using Shapley values, inspired by cooperative game theory. A Shapley value measures how much a single variable, in the context of its interaction with other variables, contributes to each individual prediction. It is well suited to interpret complex models which are difficult for physicians to interpret, such as ensemble learning models, and has been already used in a wide range of medical applications [Lundberg et al., 2018, Thorsen-Meyer et al., 2020, Hyland et al., 2020]. We used the SHAP library (see <https://shap.readthedocs.io>, version 0.39.0), which provides a fast and exact method to estimate SHAP values for ensembles of trees.

4.3 Results

Baseline characteristics

The derivation cohort included 12,338 SCD cases (60.3% men, mean age 69.4 ± 17 years) collected between 2011 and 2015 in the SDEC registry. Among them, 2,453 (19.9%) SCD occurred in a public area. A bystander was present in 8,644 (70.1%) of cases and performed cardiopulmonary resuscitation (CPR) in 4,505 (50.5%) of cases. 1,854 (16.7%) had an initial shockable rhythm, 2,779 (22.5%) were transported alive to the hospital and 649 (5.3%) patients survived at hospital discharge. SCD cases were individually matched with 12,338

controls sampled in the French general population. This resulted in a total of 24,676 participants involved in the model development. The data collected up to 5 years prior to the outcome revealed that, on average, SCD cases were prescribed 300 outpatient drugs (against 130 for controls) and received 40 hospital diagnoses (against 7 for controls).

The validation cohort (2016-2020 SDEC registry) included 11,620 SCD cases (60.5% men, mean age 71.8 ± 16 years), 11,620 matched controls sampled in the French general population and 11,620 cases of myocardial infarction. Among SCD cases, 1,883 (16.2%) occurred in a public area. A bystander was present in 7,023 (62.0%) of cases and performed CPR in 4,871 (71.1%) of cases. 1,883 (21.5%) had an initial shockable rhythm, 2,388 (20.6%) were transported alive to the hospital and 612 (5.3%) survived at hospital discharge. Notably, we found that participants of the validation cohort were prescribed more outpatient drugs (+12%) and received more hospital diagnoses (+17%) as compared to the derivation cohort, both for SCD and controls from the general population. A flow chart of the study is provided in Supplementary material, Figure A.1.

Baseline characteristics of the populations are shown in Table 4.1 and described in detail in Supplementary material, Table A.2. With the exception of social deprivation, all characteristics differed significantly between SCD cases and controls. Among the SCD population, 11,798 (49%) were hospitalized for a cardiovascular disease, compared to 5,654 (23.6%) for controls, and 19,819 (83%) were prescribed cardiovascular drugs prior to the outcome, compared to 16,818 (70.2%) for controls. Notably, 2,960 (12.4%) SCD cases had experienced an acute coronary syndrome up to 5 years before the occurrence of SCD. Additional characteristics are available in Supplementary material, Table A.2.

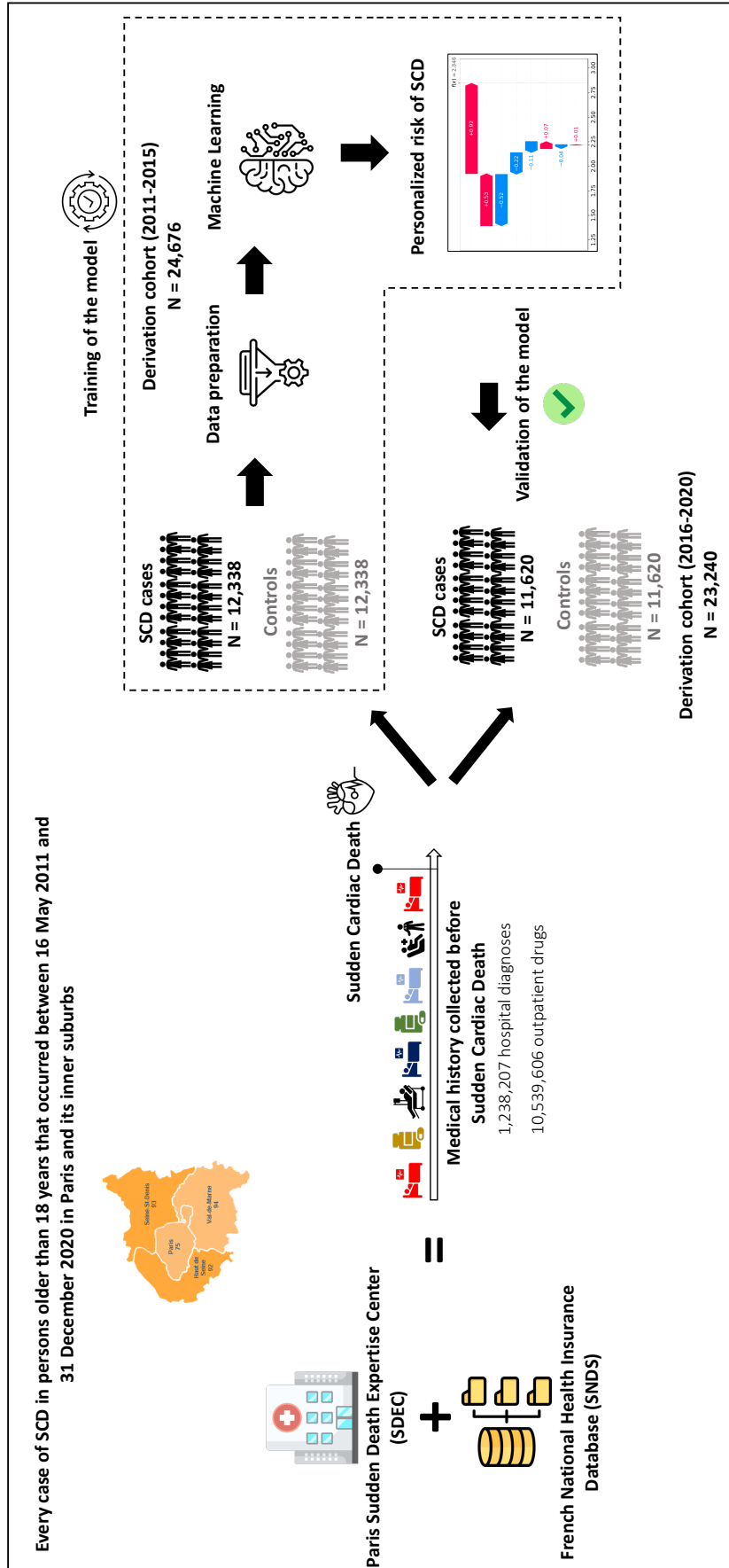


FIGURE 4.1: Methodological overview of the study

	Derivation cohort (2011 - 2015)		Validation cohort (2016 - 2020)	
	SCD N = 12,338	Controls N = 12,338	SCD N = 11,620	Controls N = 11,620
Socio-demographic characteristics				
Age, SD	69.4 (17)	69.8 (17)	71.8 (16)	72.3 (16)
Male, N (%)	7,439 (60.3%)	7,438 (60.3%)	7,028 (60.5%)	7,028 (60.5%)
Social deprivation	754 (6.1%)	644 (5.2%)	754 (6.5%)	733 (6.3%)
Cardiovascular risk factors				
Smoking abuse, N (%)	1,726 (14.0%)	669 (5.4%)	1,666 (14.3%)	445 (3.8%)
Obesity, N (%)	1,242 (10.1%)	557 (4.5%)	1,454 (12.5%)	427 (3.7%)
Dyslipidaemia, N (%)	2,091 (16.9%)	1,069 (8.7%)	1,966 (16.9%)	674 (5.8%)
Diabetes, N (%)	3,200 (25.9%)	2,025 (16.4%)	3,191 (27.5%)	1,833 (15.8%)
Hypertension, N (%)	5,444 (44.1%)	3,595 (29.1%)	5,322 (45.8%)	2,674 (23.0%)
Cardiovascular diseases				
Acute coronary syndrome, N (%)	1,547 (12.5%)	215 (1.7%)	1,413 (12.2%)	155 (1.3%)
Heart failure, N (%)	2,712 (22.0%)	800 (6.5%)	2,787 (24.0%)	663 (5.7%)
Cardiac arrhythmias and conduction disorders, N (%)	3,471 (28.1%)	1,845 (15.0%)	3,532 (30.4%)	1,431 (12.3%)
Cardiovascular drug used				
Angiotensin-converting enzyme inhibitors or angiotensin receptor blockers or aliskiren, N (%)	6,674 (54.1%)	5,072 (41.1%)	6,617 (56.9%)	4,542 (39.1%)
Diuretics, N (%)	6,342 (51.4%)	4,497 (36.4%)	6,048 (52.0%)	3,837 (33.0%)
Beta-blockers, N (%)	5,087 (41.2%)	3,142 (25.5%)	5,194 (44.7%)	2,837 (24.4%)
Aspirin, N (%)	5,476 (44.4%)	4,281 (34.7%)	5,258 (45.2%)	3,345 (28.8%)
Oral anticoagulants, N (%)	2,711 (22.0%)	1,982 (16.1%)	2,946 (25.4%)	1,726 (14.9%)
Insulin, N (%)	1,201 (9.7%)	569 (4.6%)	1,374 (11.8%)	501 (4.3%)
Oral antidiabetic agents, N (%)	2,879 (23.3%)	1,796 (14.6%)	2,948 (25.4%)	1,671 (14.4%)
Statins, N (%)	5,182 (42%)	3,782 (30.7%)	5,265 (45.3%)	3,267 (28.1%)
Other lipid-lowering agents, N (%)	1,527 (12.4%)	1,206 (9.8%)	1,154 (9.9%)	789 (6.8%)
Other comorbidities				
Chronic kidney disease, N (%)	1,846 (15.0%)	962 (7.8%)	1,928 (16.6%)	719 (6.2%)
Active cancer, N (%)	2,767 (22.4%)	2,306 (18.7%)	2,555 (22.0%)	1,660 (14.3%)

TABLE 4.1: Baseline characteristics of the populations

Variables used in the model

The aim of the study was to develop a 3-month prediction model of SCD in the general population, based on outpatient drugs and hospital diagnoses that occurred up to 5 years prior to the outcome. For this purpose, we conducted a longitudinal follow-up analysis and extracted a total of 9,460 medical codes from the SNDS database. After grouping the codes according to the third level of the ATC and ICD-10 classification systems, 196 groups of

drugs and 1,546 groups of hospital diagnoses were included in the model development. We then performed a variable selection to prevent overfitting in the cross-validation, and finally selected 188 medical codes to develop the model. A detailed flow chart of variable inclusion is provided in Supplementary material, Figure A.3.

Model performance

We found that the EHR model combined with the CatBoost algorithm offered the best performance based on cross-validation results. In the derivation cohort, the model achieved an AUC of 0.80 (95% CI 0.78 - 0.82) (see Figure 4.2), with a positive predictive value of 77% and a sensitivity of 68%. The calibration plot indicated a strong adequacy between the predicted risks and observed outcomes (see Supplementary material, Figure A.4). Notably, our model demonstrated excellent discrimination performance in the highest deciles of predicted risk, as depicted in Figure 4.3, which displays the histogram of SCD cases and controls for each decile of predicted risk. Our model detected 2,908 (24%) SCD cases with a predicted risk exceeding 90%, achieving a positive predictive value of 94% in this range. We also observed that most of controls are accurately identified in the lowest deciles, and their number linearly decreases in the high-risk subgroups.

In contrast, the CVD model combined with Logistic Regression displayed a poor predictive performance, as compared with the EHR + CatBoost model, with an AUC of 0.66 (0.63 - 0.68), a positive predictive value of 71% and a sensitivity of 45%. The use of machine learning algorithms with the same cardiovascular variables (CVD model) did not improve the results (Figure 4.2). Further information on model comparison is available in Supplementary Material, Table A.3.

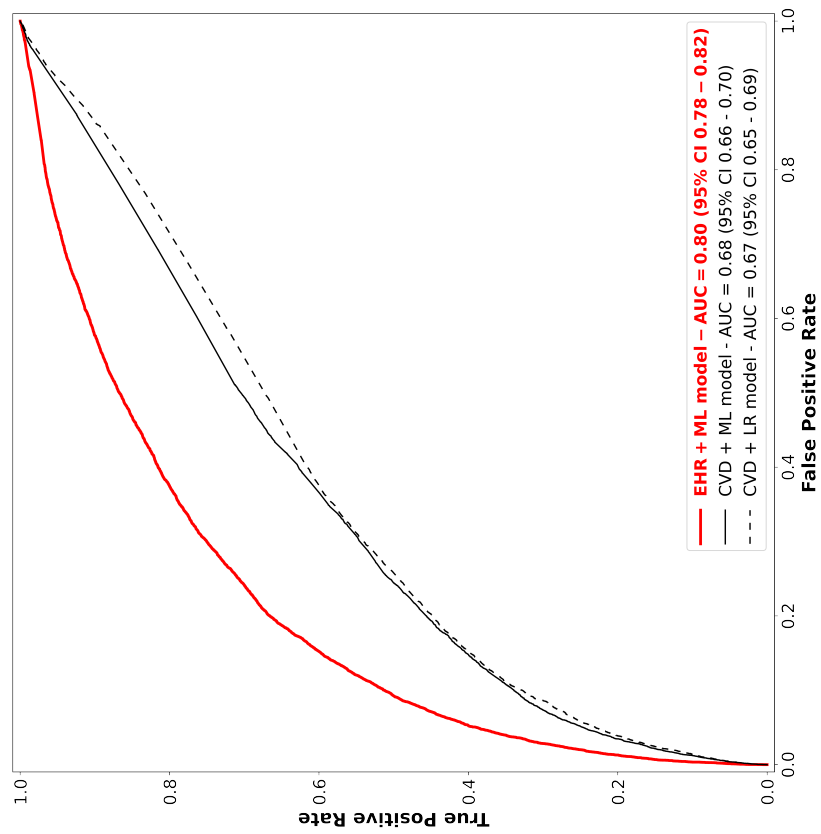
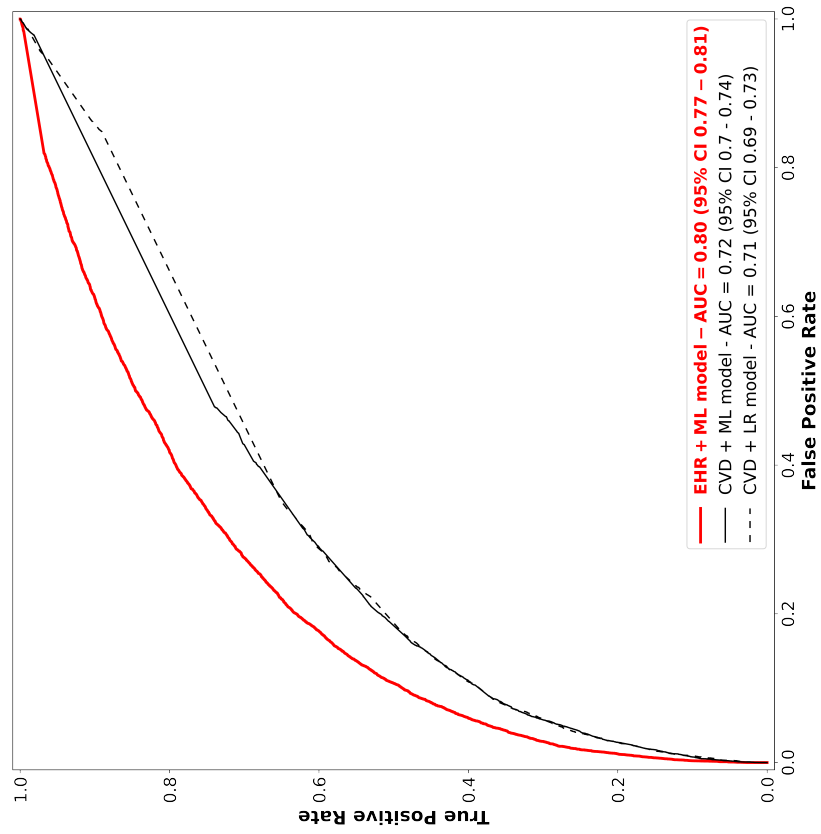
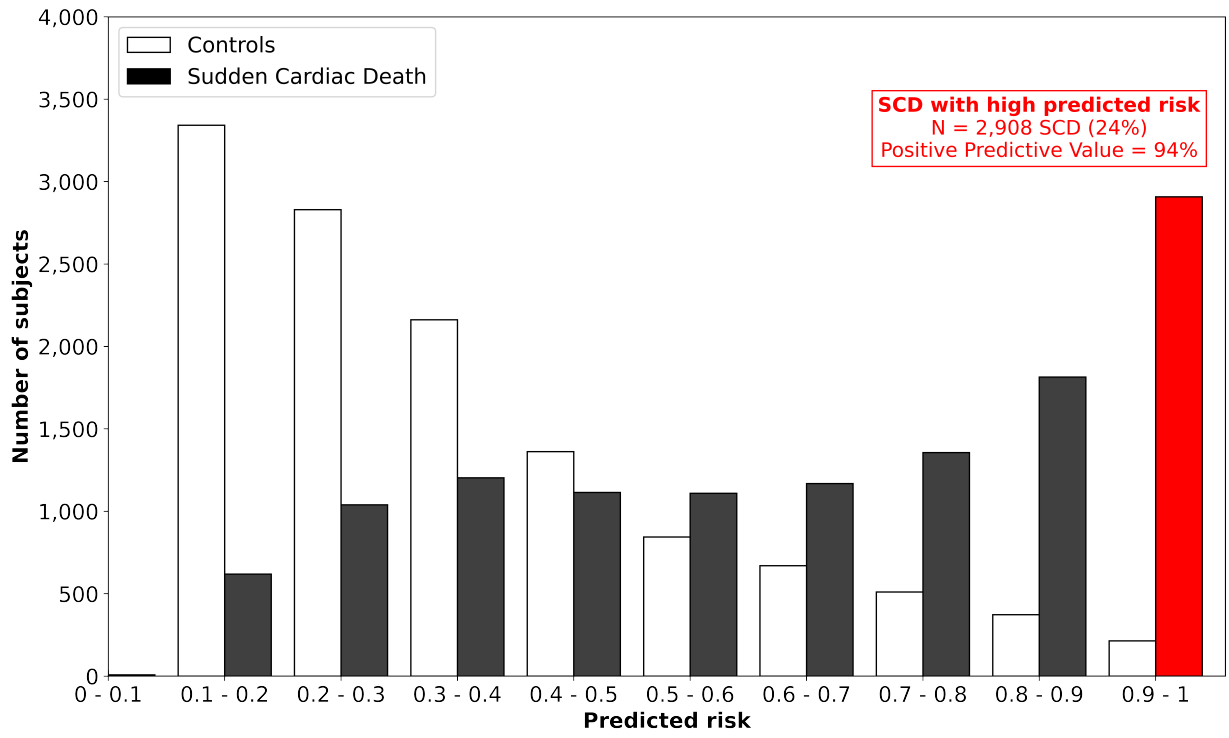
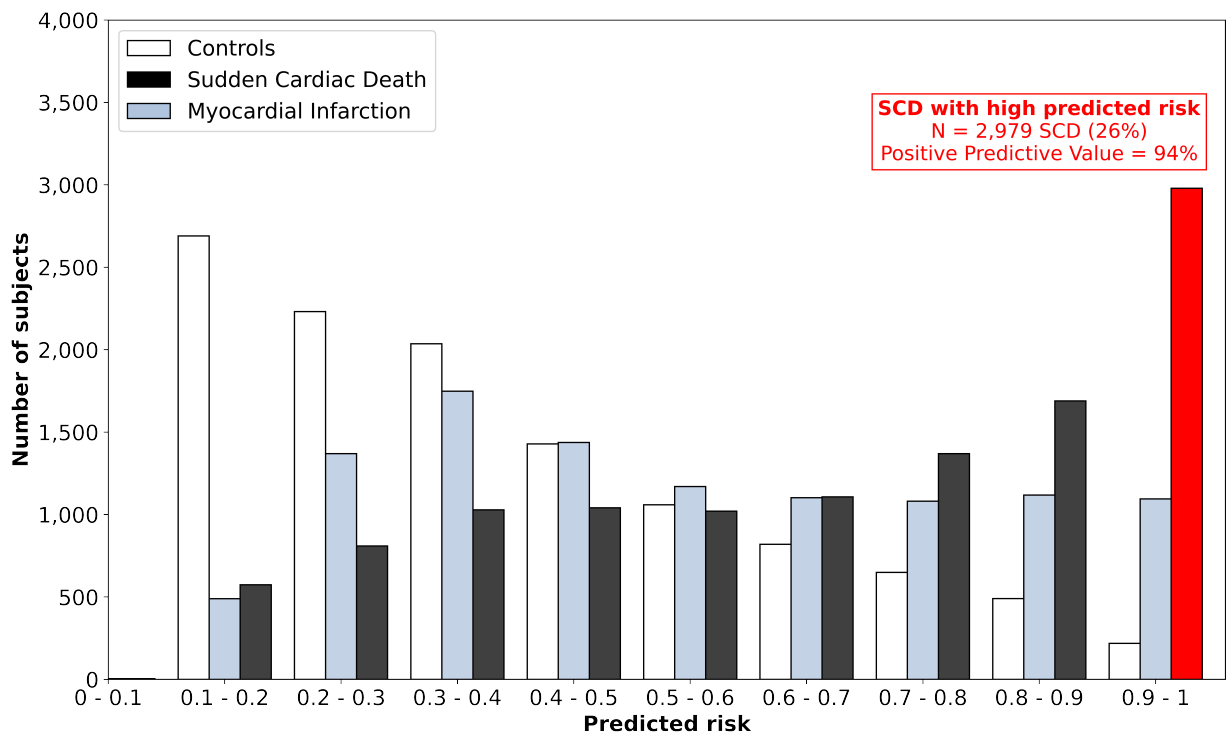


FIGURE 4.2: AUC curves



a. Derivation cohort (2011-2015)



b. Validation cohort (2016-2020)

FIGURE 4.3: Histogram of predicted risks

We then evaluated the performance of the model in the validation cohort and obtained an AUC of 0.80 (0.77 - 0.81) (Figure 2), which was consistent with the results of the derivation cohort. The sensitivity improved slightly to 71%, while the PPV decreased to 73%. The calibration plot also confirmed the robustness of the results (Supplementary material, Figure A.4), and the discrimination performance in the highest deciles of predicted risk remained

unchanged as compared to the derivation cohort (Figure 4.3). Myocardial infarction cases were mostly predicted in the low and middle deciles. 21% of them were identified in the two highest deciles, as compared to 40% for SCD. This result demonstrates the capacity of our model to accurately identify high risk patients with specific risk factors for SCD, as compared to acute coronary syndrome.

We conducted several sensitivity analyses and stratified the performance regarding age, sex and social deprivation index. The results are available in detail in Supplementary material, Table A.4. Notably, the model demonstrated a slightly better AUC and PPV for younger subjects (under 73 years old, median age of the population), while more SCD cases were identified in the oldest subjects (over 73 years old). The predictive performance remained independently associated with sex. However, we observed a slightly higher PPV for males and greater sensitivity for females. Finally, our model demonstrated superior performance for subjects with a lower social deprivation index. These insights can inform strategies to enhance the accuracy of our model in predicting SCD risk among diverse populations.

Model explanation

After training and validation, we identified the variables that have the most significant impact on the model's prediction. Among the selected variables, outpatient drugs were found to be the most influential predictors, explaining 60% of the model. Figure 4.4 illustrates the 10 most contributing groups of ATC and ICD-10 codes, which were ranked based on the sum of their importance values. The top 3 important groups are drugs related to the nervous system (15.2%), the cardiovascular system (13.1%) and alimentary tract and metabolism (11.2%), which collectively account for 39.4% of the model's overall prediction.

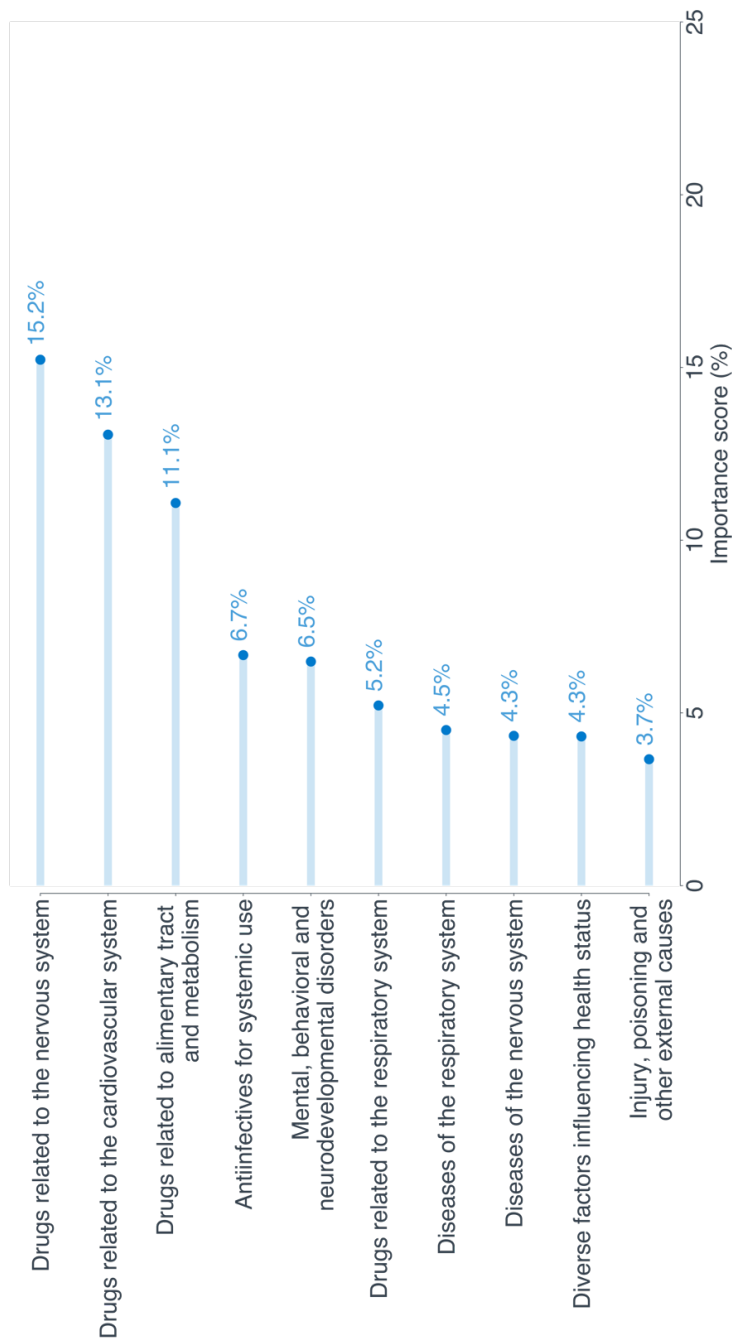


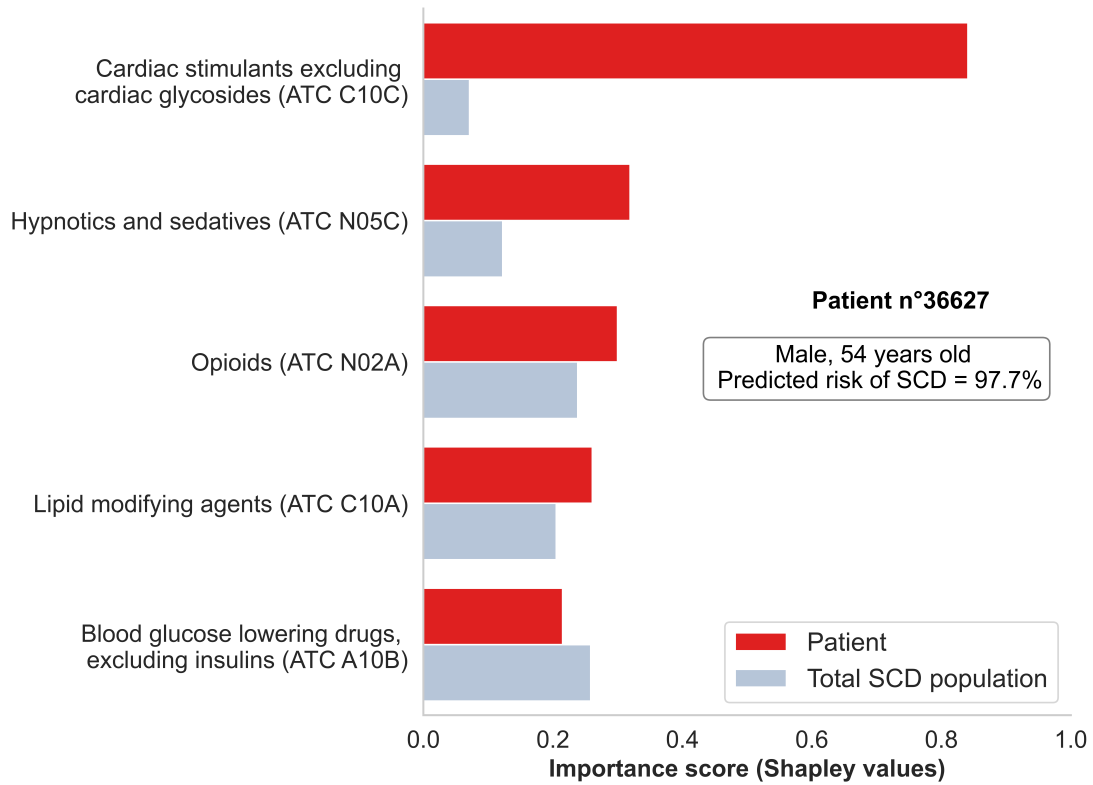
FIGURE 4.4: Importance of the variables

In addition, we conducted supplementary analysis to provide a detailed explanation of the prediction at the individual level. For each subject of the study, we computed the Shapley scores from the CatBoost model to evaluate the contribution of variables to his predicted risk. To illustrate this personalized approach, we present the Shapley scoring method for 2 SCD cases of the validation cohort in Figure 4.5.

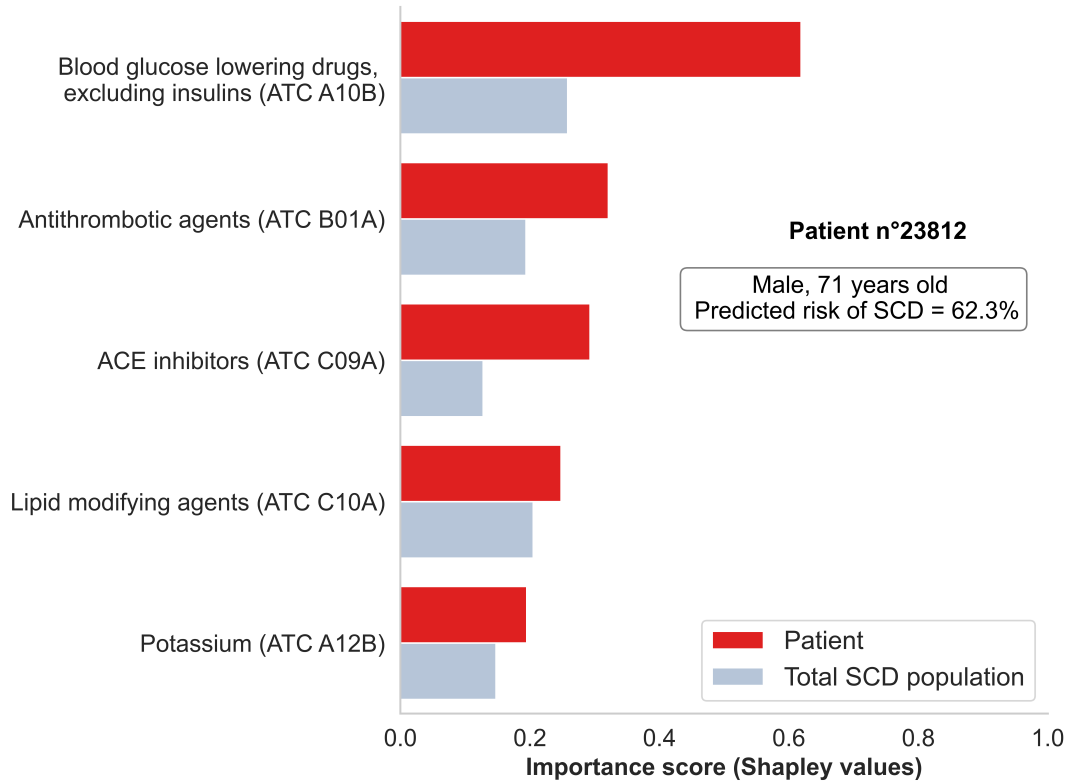
We first report the case of a 54-year-old male (Figure 5.a) who was hospitalized before SCD for diseases of the nervous system, including Parkinson’s disease, dystonia and

mononeuropathies of upper limb. He also had a history of inflammatory diseases of the genitourinary system (prostate inflammation and acute pyelonephritis) as well as cardiovascular disorders (hypertension and paroxysmal tachycardia). Our predictive model identified this patient with a very high level of risk (97.7%), which was mainly explained by cardiac stimulants, hypnotics and sedatives, lipid modifying agents and blood glucose lowering drugs.

The second case is a 71-year old male (Figure 5.b) who had not been hospitalized for the past 5 years prior to SCD. He was prescribed drugs mainly related to the alimentary tract and metabolism (blood glucose-lowering drugs, excluding insulins) and the cardiovascular system (beta-blocking agents, ACE inhibitors, lipid-modifying agents, and vasodilators). Our predictive model identified this subject with a moderate level of risk (62.3%), which was mainly explained by blood glucose-lowering drugs, antithrombotic, ACE inhibitors and potassium.



a. Most important risk factors for a high-level risk profil



b. Most important risk factors for a moderate-level risk profil

FIGURE 4.5: Individual explanations of the prediction model

4.4 Discussion

For the first time, we provide a personalized prediction model of SCD, which explains the predicted risk at the individual level, and accurately identifies the top deciles of risk in the general population. Our machine learning algorithm uses comprehensive assessment of medical history from 23,000 SCD cases, and achieved high discrimination and excellent calibration. Notably, the model identified a subgroup of patients with a predicted risk of SCD over 90% and a PPV of 94%, representing more than 25% of the total SCD population. These findings offer a new step towards personalized prevention in the general population, especially for high-risk individuals who may benefit from specific interventions.

To the best of our knowledge, this is the first research project using big data and machine learning techniques for SCD prediction in the general population. Recently, the PROFID study [Glen2021] proposed a new AI approach to predict SCD after myocardial infarction, by merging all important SCD clinical trials and cohorts, resulting in 225,000 participants from Europe, United States, and Israel. 85 variables, mostly cardiovascular, were included in the model, in addition to LVEF. However, this model can be only used to predict SCD among very-high risk patients, who are already well identified for preventive strategies. Previous research has attempted to predict SCD in the general population, with some success. Early cohorts from the 1970s combined baseline characteristics with family history, heart rate during rest and exercise, and ECG abnormalities as potential predictors of SCD compared to ischemic heart disease [Jouven et al., 1999, 2005]. More recent studies have developed generalizable risk scores for SCD in the general population using data from large cohorts, such as the Atherosclerosis Risk in Communities, the Cardiovascular Health Study and the Framingham Heart Study [Deo et al., 2016, Waks et al., 2016, Aro et al., 2017, Bogle et al., 2018, Holkeri et al., 2020]. These models showed good discrimination performance, with C-statistics ranging from 0.74 to 0.89, suggesting their potential utility in predicting SCD in the general population. However, they also predict the occurrence of non-sudden coronary death, even after considering electrical heterogeneity, and are not specific to SCD. One possible explanation for this limitation is that most risk factors considered in the models are mostly related to coronary atherosclerosis development, and are not specific to ventricular arrhythmic susceptibility. In addition, these models cannot provide personalized risk scores for each patient, as their methodological approach is not designed for individual risk assessment.

Current prediction models are often based on cardiovascular risk factors only, which may not provide a comprehensive picture of SCD in the general population. Most of these variables are also measured at one single occasion and do not integrate trajectories of risk factors and treatments over time. In contrast, our approach capitalizes on the use of one of the largest electronic health records databases in the world, which provides a comprehensive source of data. Through longitudinal follow-up, we were able to collect nearly 10,000 daily data points for each participant up to 5 years prior to SCD, including both cardiovascular and non-cardiovascular variables. This approach achieved to identify individuals at risk of SCD who may have otherwise gone undetected using traditional cardiovascular risk factors alone. This result could have significant implications for the general population, as our model can be easily deployed at scale and implemented in low-cost mass screening programs.

In addition, our findings demonstrate excellent calibration and high accuracy in predicting a significant proportion of cases in the last decile of predicted risk, which accounts for more than 25% of the total SCD population. These individuals could greatly benefit from

tailored interventions, such that ICD implantation, particularly those with unknown heart disease who are not identified by current risk stratification guidelines. Hence, our results suggest that incorporating our prediction model into routine clinical practice could improve the identification of optimal candidates for primary prevention of SCD.

Another major strength of our model is the use of Shapley values, which identify the variables that drive the predicted risk at the individual level. Most of machine learning models provide predictions without explanation and are difficult for physicians to trust and are given little insight into how they should respond. SHAP model provides a solution to mitigate this issue and to explain any prediction method, whereas classical approaches provide only global effects of the variables. This level of interpretability is critical, as it allows clinicians to tailor preventive measures for each patient, and better understand how to respond to their predicted risk. Our results offer therefore a promising avenue for future research in the field of personalized SCD prediction. However, it is important to note that we only provide an association between the identified modifiable risk factors and SCD, and caution is needed when it comes to identifying causation. Future prospective studies will be required to gain a more comprehensive understanding of the complex links between modifiable risk factors and SCD.

Several limitations should be acknowledged in our study. First, the investigation was conducted in France, and our findings may not be directly generalizable to other populations due to potential differences in environmental factors and healthcare systems. The French National Health Insurance database is a unique and comprehensive database which may be difficult to replicate in other countries. Therefore, caution is advised when extrapolating our results to other regions or ethnic groups, and further validation studies in diverse EHR collection systems are needed to determine the generalizability of our findings. The use of EHR as a source of data is another limitation. While EHR have demonstrated their potential as a valuable tool for research, they may not capture all relevant clinical information accurately. Incomplete or incorrect documentation, data entry errors, and variations in the quality of EHR systems may affect the accuracy and completeness of medical codes used in our analysis. Therefore, some risk factors included in the model may have been missed or misclassified, leading to potential underestimation or overestimation of their association with SCD. Finally, our study did not consider important risk factors such as smoking and physical activity, which were only approximated. Other potential risk factors, such as hyperlipidemia, were also only partially captured in the SNDS database, based on surrogate markers. This approach may have introduced bias or imprecision in the analysis. Future studies that include a more comprehensive assessment of lifestyle and behavioral factors are therefore needed to improve our prediction model.

4.5 Conclusion

We developed and validated a prediction model of SCD in the general population, using a unique population-based registry and large-scale data analysis of electronic health records. This personalized approach, based on machine learning algorithms and a comprehensive assessment of both cardiovascular and non-cardiovascular risk factors, make it promising to enhance preventive strategies and reduce the global burden of SCD in the population.

Chapter 5

Scalable Bayesian Bi-Level Variable Selection in Generalized Linear Models

Contents

5.1	Introduction	111
5.2	Model	112
5.3	The proposed algorithm	114
5.4	Numerical experiments	117
5.5	Conclusion	123

This Chapter comes from a joint work with Nicolas Chopin, and was submitted in the Foundations of Data Science Journal

5.1 Introduction

Motivation

While useful more generally, the approach developed in this paper was initially motivated by a public health dataset recording the medical history of a large number of individuals that may or may not have suffered from sudden cardiac death (SCD); this dataset will be described more fully later. One may use this data to determine whether consumption of medical drugs or hospitalization may increase the odds of an SCD event. Unfortunately, the number of potential drugs and diseases is very large, and their incidence in the studied population vary a lot. This makes it difficult to assess the impact of drugs and diseases that are rarely prescribed or observed. On the other hand, there are official nomenclatures for drugs and diseases, which can be classified into groups with similar properties. Hospital diagnoses are coded according to the International Classification of Diseases and drugs are coded according to the Anatomical Therapeutic Chemical system, that classifies them according to the organ or system on which they act and their therapeutic, pharmacological, and chemical properties. Therefore, there is clear medical interest in determining automatically whether there is enough information in the data to indicate that a particular drug or disease affects SCD, or, if not, whether the group it belongs to does.

This led us to develop a bi-level variable selection procedure, based on a binary regression (outcome variable is whether the individual had an SCD event) model, and which should work reliably for a fairly large number of individuals, variables and groups. In addition, we wanted this procedure to be Bayesian, in order to be able to obtain posterior

probabilities of inclusion (rather than simply 0/1 answers).

There are surprising few papers on Bayesian bi-level variable selection, and most of them focus on linear regression with Gaussian noise [Chen et al., 2016, Mallick and Yi, 2017, Cai et al., 2020]. For such a model, one may integrate out the regression coefficients (the prior provided is Gaussian) to obtain the marginal posterior distribution over a finite space (the inclusion of either individual variables or groups). Even so, designing a MCMC able to efficiently explore that finite space is challenging. Such discrete distributions tend to exhibit strongly separated modal regions, and a MCMC chain may fail to escape one of this region. We refer in particular to the numerical experiments of Schäfer and Chopin [2013] that show that various MCMC schemes may lead to unstable estimates because of this problem. Of course, this issue gets worse when the number of variables increases, making MCMC unable to scale properly with datasets with a large number of variables (and groups).

Proposed approach

Schäfer and Chopin [2013] designed a tempering SMC sampler for standard (one-level) variable selection for linear regressions, and showed it outperformed significantly MCMC, as explained above. We adapt this approach to our problem in three ways. First, we replace it by a waste-free SMC sampler, following Dau and Chopin [2022], as waste-free SMC tends to outperform standard SMC. Waste-free SMC amounts to resampling only a fraction of the particles, then moving them through numerous MCMC steps, and keeping all these intermediate. Second, we adapt the proposal mechanism within the MCMC step so as to accommodate the constraints specific to bi-level selection (namely, that a variable may be selected only if its group is selected). Third, we replace the intractable marginal likelihood (obtained by integrating out the regression coefficients) by either its LA (Laplace approximation), or by a cheaper approximation introduced by Rossell et al. [2021], called ALA (approximate LA). The reason why ALA is particularly attractive in our context is that it scales very well with respect to n (as we explain later). We assess in our numerical experiments the impact of the error introduced by ALA on the actual results. We note that Schäfer [2012] already showed in his PhD thesis that replacing the marginal likelihood by its LA within a SMC sampler (targeting a variable selection posterior) incurs only a negligible bias.

Plan

Section 5.2 describes the considered class of model, the bi-level variable selection problem, and the related notations. Section 5.3 describes the proposed algorithm, starting with a generic (waste-free) SMC sampler, and explaining how this generic algorithm may be adapted to bi-level variable selection. Section 5.4 assesses (statistically and numerically) the proposed approach through two numerical experiments, one on simulated data and one on the public health dataset mentioned in the introduction.

5.2 Model

Regression model

For the sake of concreteness, we consider the following binary regression model, although our approach could easily be generalised to other generalised linear models. We suppose that we have collected a dataset $\mathcal{D} = \{X, U, Z, y\}$ with sample size n , where $y \in \{0, 1\}^n$ is a vector of binary responses, $X = (x_{ij}) \in \mathbb{R}^{n \times p}$, $U = (u_{ij}) \in \mathbb{R}^{n \times q}$, and $Z = (z_{ij}) \in \mathbb{R}^{n \times r}$, are design matrices that contain, respectively, ‘individual variables’, ‘group variables’ (both

subject to variable selection later on), and extra variables that the user wants to include systematically (e.g. the intercept, socio-demographic effects such as sex, age, etc.).

Regarding the group structure, we assume that each of the p variables in X belongs to one (and only one) of the q groups; let $g(j)$ be the group of variable j . A group variable (in U) may represent different types of ‘group effects’. For instance, in a medical application, the variables in a group k may be the indicator that the patient took a certain drug in the last six months, and the group variable may be the indicator that a patient took any drug in that group in the same period. Alternatively, these variables could be the number of drug intakes for each drug; in that case, the group variable would be the number of intakes of drugs in that group. In either scenarios, the point is to determine whether one may measure a significant effect for each individual variable, *on top of* the group effect, or a significant effect for its group only, or neither.

To sum up, without variable selection, the distribution of each data point would be such that, for $i = 1, \dots, n$:

$$P(Y_i = 1|\beta) = F \left(\sum_{j=1}^p \beta_j^x x_{ij} + \sum_{k=1}^q \beta_k^u u_{ik} + \sum_{l=1}^r \beta_l^z z_{il} \right) \quad (5.1)$$

and $P(Y_i = 0|\beta) = 1 - P(Y_i = 1|\beta)$, where $\beta = (\beta^x, \beta^u, \beta^z)$ is the vector of regression parameters, F is the link function (e.g. $F = \Phi$, the unit Gaussian CDF for a probit model). We assign independent Gaussian priors to the regression coefficients: $p(\beta^z) \sim \mathcal{N}(0_r, \sigma^2 \mathbf{I}_r)$, $p(\beta^u) \sim \mathcal{N}(0_q, \sigma^2 \mathbf{I}_q)$ and $p(\beta^x) \sim \mathcal{N}(0_p, \sigma^2 \mathbf{I}_p)$.

Bi-level variable selection

We extend our model to perform selection of groups and variables simultaneously. Most of existing models lack flexibility as they impose only “all-in” or “all-out” selection for variables in the same group. That is, if a group is not selected by the model, variables belonging to this group will also not be selected. In this work, we propose a more general approach in order to capture sparsity at both the group and variable levels. To this end, we introduce $\theta = (\gamma, \eta)$, a set of two types of binary variables: γ_k indicates whether group k is active ($\gamma_k = 1$) or not ($\gamma_k = 0$), and η_j indicates whether individual variable j , which is in group $g(j)$, is active ($\eta_j = 1$) or not ($\eta_j = 0$). We consider a hierarchical structure such that the variable j is not selected if $\gamma_{g(j)} = 0$, that is $P(\eta_j = 1|\gamma_k = 0) = 0$ for $k = g(j)$. As compared to existing models, we propose to keep the flexibility of selecting variables within a group. For example, when a group of drugs is related to SCD, it does not necessarily mean that all drugs of this group are related to SCD. Therefore, we may want to not only remove unimportant groups effectively, but also identify important variables within important groups as well. Thus, we replace (5.1) by

$$P(Y_i = 1|\beta, \theta) = F \left(\sum_{j=1}^p \eta_j \beta_j^x x_{ij} + \sum_{k=1}^q \gamma_k \beta_k^u u_{ik} + \sum_{l=1}^r \beta_l^z z_{il} \right). \quad (5.2)$$

Let $p(\gamma)$ be the prior density of γ , which is a product of Bernoulli distributions with probabilities p_j^γ . For the predictors, we introduce a spike-and-slab prior defined by

$$P(\eta_j = 1|\gamma) = \begin{cases} p_j^\eta & \text{if } \gamma_{g(j)} = 1 \\ 0 & \text{otherwise.} \end{cases} \quad (5.3)$$

This bi-level structure implies that variable j may be selected only if the group it belongs to, $g(j)$, is selected.

To perform Bayesian bi-level variable selection, we aim to approximating the (marginal) posterior distribution of $\theta = (\gamma, \eta)$, i.e. $\pi(\theta) = p(\theta|\mathcal{D}) \propto p(\theta)L(\theta)$, where $p(\theta)$ is the prior described above, and $L(\theta)$ is the integrated likelihood obtained by integrating out β :

$$L(\theta) = \int L(\beta, \theta)p(\beta)d\beta, \quad L(\beta, \theta) = \left\{ \prod_{i=1}^N P(Y_i = y_i|\beta, \theta) \right\}.$$

5.3 The proposed algorithm

Tempering waste-free SMC

We propose a tempering waste-free Sequential Monte Carlo (SMC) sampler to approximate the joint posterior distribution $\pi(\theta) = p(\theta|\mathcal{D})$. SMC methods are iterative stochastic algorithms that approximate a sequence of probability distributions through successive importance sampling, resampling and Markov steps. In Bayesian modeling, this sequence can be used to interpolate between a distribution $p(\theta)$ which is easy to sample from (e.g. the prior distribution) and a distribution of interest $\pi(\theta)$ which may be difficult to simulate directly (i.e. the posterior distribution). The tempering approach in particular is based on a sequence of tempered distributions of the form

$$\forall t \geq 1, \pi_t(\theta) = \frac{p(\theta)L(\theta)^{\lambda_t}}{Z_t}$$

where $p(\theta)$ is the prior density, $L(\theta)$ the likelihood, $Z_t > 0$ is the normalising constant and $0 = \lambda_0 < \lambda_1 < \dots < \lambda_T = 1$ is a sequence increasing from 0 to 1. This geometric bridge smoothly interpolates between the initial distribution $p(\theta)$ and the target distribution $\pi(\theta) \propto p(\theta)L(\theta)$.

A typical application of such an approach is the simulation of a multimodal distribution π . Since simulating directly from such a distribution is difficult, we may use tempering SMC instead, to sample initially from a distribution p which covers the support of π , and to move progressively towards π through intermediate distributions that are progressively more and more multimodal. In this work, we combined the tempering approach with the waste-free SMC sampler proposed by Dau and Chopin [2022]. The main idea of this scheme is to resample only M ancestors from the N particles in the standard SMC sampler (with $M \ll N$). Each of the ancestors is then moved $P - 1$ times through a Markov kernel K_t . The M chains of length P are finally put together to form a new particle sample of size $N = MP$. Algorithm 4 describes the corresponding algorithm for a tempering sequence. At the final iteration T of the algorithm, one may approximate any expectation $\mathbb{E}_\pi \varphi(\theta)$ with $\sum_{n=1}^N W_T^n \varphi(\theta_T^n)$, where the W_T^n are the normalised weights at the final iteration T .

In practice, it is recommended to set the successive λ_t automatically, by choosing the next λ_t so that the ESS (effective sample size) of the weights equal a certain threshold. Another advantage of a SMC sampler such as Algorithm 4 is that it is easy to parallelise; in particular the evaluation of the likelihood of the N particles (which is typically the bulk of the computation) may be performed in parallel. We refer to Dau and Chopin [2022] for a more thorough discussion of the advantages of SMC samplers over MCMC, and the extra advantage brought by waste-free SMC (relative to standard SMC), in particular the greater

Algorithm 4: Tempering Waste-free SMC

Input : Prior distribution $p(\theta)$, likelihood function $\theta \rightarrow L(\theta)$, integers N, M, P such that $N = MP$, sequence $0 = \lambda_0 < \dots < \lambda_T = 1$, Markov kernels K_t that leave invariant $\pi_{t-1} \forall t \geq 1$

```

1 for  $t \leftarrow 0$  to  $T$  do
2   if  $t = 0$  then
3     for  $n \leftarrow 1$  to  $N$  do
4        $\theta_0^n \sim p(\theta)$ 
5   else
6      $A_t^{1:M} \sim \text{resample}(M, W_{t-1}^{1:N})$  (Draw IID variables such that
7        $P(A_t^m = n) = W_{t-1}^n$  for  $n = 1, \dots, N$ )
8     for  $m \leftarrow 1$  to  $M$  do
9        $\tilde{\theta}_t^{m,1} \leftarrow \theta_{t-1}^{A_t^m}$ 
10      for  $p \leftarrow 2$  to  $P$  do
11         $\tilde{\theta}_t^{m,p} \sim K_t(\tilde{\theta}_t^{m,p-1}, d\theta_t)$ 
12      Gather variables  $\tilde{\theta}_t^{m,P}$  so as to form a new sample  $\theta_t^{1:N}$ 
13      for  $n \leftarrow 1$  to  $N$  do
14         $w_t^n \leftarrow L(\theta_t^n)^{\lambda_t - \lambda_{t-1}}$ 
15      for  $n \leftarrow 1$  to  $N$  do
16         $W_t^n \leftarrow w_t^n / \sum_{m=1}^N w_t^m$ 
    
```

robustness relative to the choice of tuning parameters such as P and M .

For now, there are two points that need to be addressed in order to apply Algorithm 4 to our variable selection problem: first, we need to design Markov kernels K_t that leave invariant π_{t-1} at time t , and in particular that sample within the constrained support of π_{t-1} in our bi-level selection scenario (i.e. the fact that $\eta_j = 0$ as soon as $\gamma_{g(j)} = 0$). Second, we must find a way to evaluate, or approximate, the marginal likelihood $L(\theta)$. These two points are discussed in the next two sections.

π_{t-1} -invariant kernels

Consider a target distribution over binary vectors; that is $\pi(\gamma)$ with $\gamma \in \{0, 1\}^q$. Designing an efficient MCMC kernel that leaves invariant this target is challenging. One option is to use a Gibbs kernel, or a Metropolis kernel based on a local proposal, where only one component may be flipped at a time. But such kernels tend to mix poorly, and to get stuck in local modes.

The SMC sampler of Schäfer and Chopin [2013] used instead an independent Metropolis kernel based on a global proposal of the form:

$$q(\gamma) = q_1(\gamma_1) \prod_{k=1}^q q_k(\gamma_k | \gamma_{1:k-1}), \quad q_k(\gamma_k = 1 | \gamma_{1:k-1}) = \text{logistic} \left(b_{kk} + \sum_{i=1}^{k-1} b_{ki} \gamma_i \right). \quad (5.4)$$

that is, a sequence of nested logistic regressions. Given the chain rule decomposition above, it is easy to sample from this proposal distribution. In order to ensure that the resulting independent Metropolis sampler mixes well (and in particular that the acceptance rate is high), one needs to ensure that the proposal is as close as possible to the target. To

ensure this, Schäfer and Chopin [2013] set the parameters b_{ji} to the maximum likelihood estimators of the corresponding logistic regressions, based on the current (weighted) particle sample. The numerical experiments of Schäfer and Chopin [2013] show that a SMC sampler based on such global (properly calibrated) Metropolis steps may outperform significantly local MCMC chains.

Since Schäfer and Chopin [2013] considered standard (one-level) variable selection, they did not have to deal with constrained distribution (i.e. each vector $\gamma \in \{0, 1\}^p$ has positive probability). We adapt their approach to bi-level variable selection as follows. First, we extend the proposal in (5.4) as follows:

$$q(\theta) = q(\gamma, \eta) = q_1(\gamma_1) \prod_{k=1}^q q_k(\gamma_k | \gamma_{1:k-1}) \prod_{j=1}^p q_j(\eta_j | \gamma_{g(j)}). \quad (5.5)$$

where the conditional distributions of the η_j 's are set in the same way as in (5.4). Second, we set the conditional proposals of the η_j as follows:

$$q_j(\eta_j = 1 | \gamma_{g(j)}) = \begin{cases} c_j & \text{if } \gamma_{g(j)} = 1 \\ 0 & \text{otherwise} \end{cases}$$

where $c_j \in [0, 1]$ is a tuning parameter. We calibrate the c_j 's in the same way as for the coefficients b_{ji} in (5.4): by maximum likelihood estimation on the current particle sample.

This proposal respects the constraint that η_j must be zero as soon as $\gamma_{g(j)} = 0$. It is basic, and may be extended by correlating the η_j 's in the same group through a nested logistic regression of the same form as for the γ_k . In practice however, we did not observe much benefit in doing so, and stuck to this basic structure. Algorithm 5 summarizes how one may implement the considered type of Metropolis kernels.

Algorithm 5: Independent Metropolis kernel used to move the particles within Algorithm 3 at time t

Input : $\theta = (\gamma, \eta)$, tuning parameters (b_{ji}) and (c_j) (estimated from the current particle sample).

Output: A sample from $K_t(\theta, d\theta')$, where K_t leaves invariant π_{t-1} .

```

1  $\theta^p \sim q(\theta)$  (as defined in (5.5))
2  $u \sim \text{Uniform}[0, 1]$ 
3 if  $u \leq \pi_{t-1}(\theta^p)q(\theta)/\pi_{t-1}(\theta)q(\theta^p)$  then
4 |   return  $\theta^p$ 
5 else
6 |   return  $\theta$ 
    
```

Approximation of the marginal likelihood

The marginal likelihood $L(\theta) = \int L(\beta, \theta)p(\beta)d\beta$ is typically intractable (unless one considers a linear Gaussian regression model). A popular approximation to this quantity is the Laplace approximation (LA), which amounts to Taylor expanding the log of the integrand around its mode. Let β_θ denote the vector made of the components β_i such that $\theta_i = 1$, $h_\theta(\beta_\theta) = -\log\{L(\beta, \theta)p(\beta)\}$, and $\hat{\beta}_\theta = \arg \min_{\beta_\theta} h_\theta(\beta_\theta)$ (i.e. the MAP estimator given

θ), then:

$$\begin{aligned}\log L(\theta) &= \log \int \exp \{-h_\theta(\beta_\theta)\} d\beta_\theta \\ &\approx -h_\theta(\hat{\beta}_\theta) + \log \int \exp \left\{ -\frac{1}{2}(\beta_\theta - \hat{\beta}_\theta)^T \hat{H}_\theta(\beta_\theta - \hat{\beta}_\theta) \right\} d\beta_\theta \\ &= -h_\theta(\hat{\beta}_\theta) + \frac{d_\theta}{2} \log 2\pi - \frac{1}{2} \log |\hat{H}_\theta|\end{aligned}$$

where $|\hat{H}_\theta|$ is the determinant of the Hessian of function $\beta_\theta \rightarrow h_\theta(\beta_\theta)$ at $\beta_\theta = \hat{\beta}_\theta$, and $d_\theta = \dim \beta_\theta$.

Schäfer [2012] in his thesis gave numerical evidence that replacing the marginal likelihood with its Laplace approximation, within a SMC sampler for standard (one-level) variable selection, works well, in the sense that it leads to a negligible error (for approximating the posterior of θ). On the other hand, computing the Laplace approximation for many simulated θ -values is expensive; for each θ , one needs to run a Newton-Raphson optimiser to obtain $\hat{\beta}_\theta$ and \hat{H}_θ . Furthermore these operations have complexity $\mathcal{O}(n)$ in the sample size, and $\mathcal{O}(d_\theta^3)$ in the dimension.

Rossell et al. [2021] proposed a cheaper approximation, based on a Taylor expansion similar to Laplace, but around zero. Let $\mathbf{0}_\theta$ denote a vector of zeros of the same dimension as β_θ , then, the ALA (approximate Laplace approximation) is:

$$\begin{aligned}\log L(\theta) &\approx -h_\theta(\mathbf{0}_\theta) + \log \int \exp \left\{ -\beta_\theta^T g_\theta - \frac{1}{2} \beta_\theta^T H_\theta \beta_\theta \right\} d\beta_\theta \\ &= -h_\theta(\mathbf{0}_\theta) + \frac{1}{2} g_\theta^T H_\theta^{-1} g_\theta + \frac{d_\theta}{2} \log 2\pi - \frac{1}{2} \log |H_\theta|\end{aligned}$$

where g_θ and H_θ denote respectively the gradient and Hessian of function $\beta \rightarrow h_\theta(\beta)$ at point $\beta_\theta = \mathbf{0}_\theta$. Note that in practice, one simply need to compute the gradient g and Hessian H of minus log-likelihood at zero for the *full* model (i.e. θ is a vector of ones, all variables are included), to obtain g_θ and H_θ (e.g. g_θ contains the components i of g such that $\theta(i) = 1$, and H_θ is defined similarly).

Once quantities g and H have been computed in a preliminary step, the computation of ALA is $\mathcal{O}(1)$ in the sample size n . Its complexity remains cubic in the dimension, because of the determinant, however. Rossell et al. [2021] make it clear that ALA is not a consistent (in n) approximation of the marginal likelihood; they mention that it tends to be biased against truly active variables. That is, it tends to under-estimate the posterior probability that an active variable should be included. We refer to Rossell et al. [2021] for more discussion on this matter.

Still, ALA remains particularly attractive in our context, as our SMC sampler must perform many evaluations of the marginal likelihood. We will assess the impact of the approximation error of ALA by comparing two waste-free SMC samplers, one based on LA, and one based on ALA.

5.4 Numerical experiments

As explained above, our goal in this section is to assess numerically the performance of our tempering waste-free SMC sampler for bi-variable selection, when the marginal likelihood

is evaluated through either LA or ALA. We take the number of particles to be $N = 25,000$, and set $M = 125$, $P = 200$. Our algorithm was implemented using the particles Python library (see <https://github.com/nchopin/particles>). The results were obtained using a server with 64 Gb RAM and 8 cores.

Simulated data

We simulate data from our model (using the probit link function), using $g = 5$ groups, $r = 5$ systematically included covariates, a varying number p of individual variables, and a varying sample size n ; see below. The rows of the design matrices X , U , and Z are sampled independently from a Gaussian distribution $N(0, \Sigma)$, where $\Sigma_{ii} = 1$, and $\Sigma_{ij} = 0.5$. The corresponding regression parameters are set to $\beta^z = (0, 0, 1, 1, 1)$, $\beta^u = (0, 0, 1, 1, 1)$ and the components of β^x are all set to zero, except for the last variable of each active group, where it is set to one.

In a first scenario, we set $p = 50$ and let n vary from 100 to 2,500; while in a second scenario we fix $n = 1,500$ and let p vary from 10 to 250. We run our algorithm 10 times and uses the empirical standard deviation to draw confidence intervals.

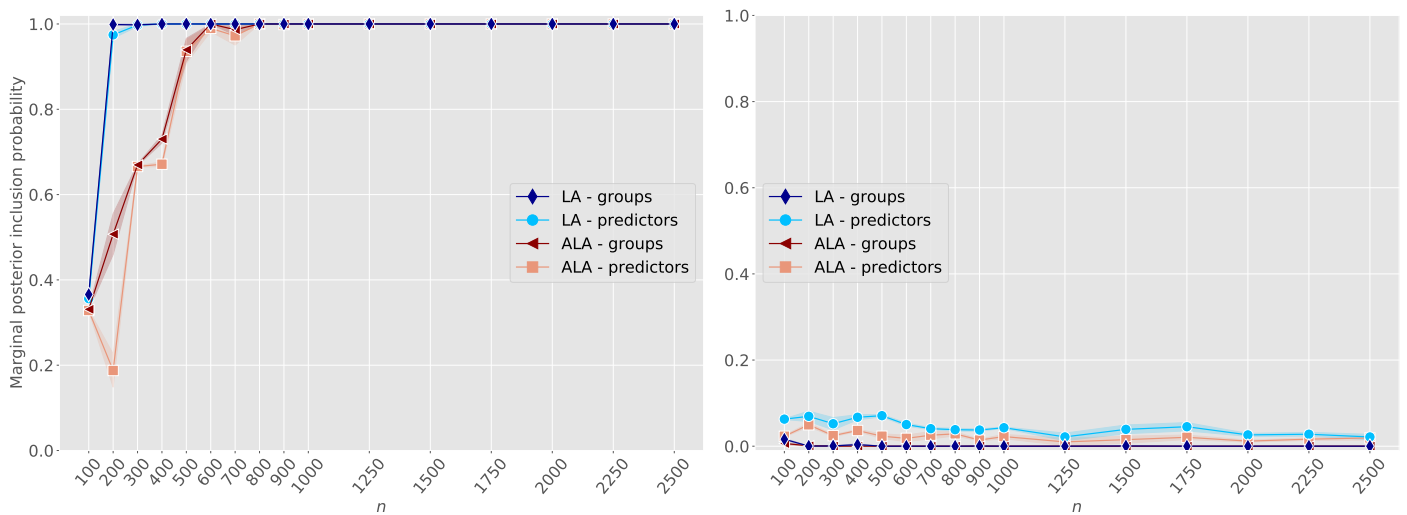


FIGURE 5.1: Comparison of ALA and LA for posterior inclusion probabilities of groups and predictors when n varies

Comparison of ALA and LA for posterior inclusion probabilities of groups and predictors when n varies from 100 to 2,500, with $p = 50$. Left: average posterior inclusion probabilities for truly active variables. Right: average posterior inclusion for truly inactive variables.

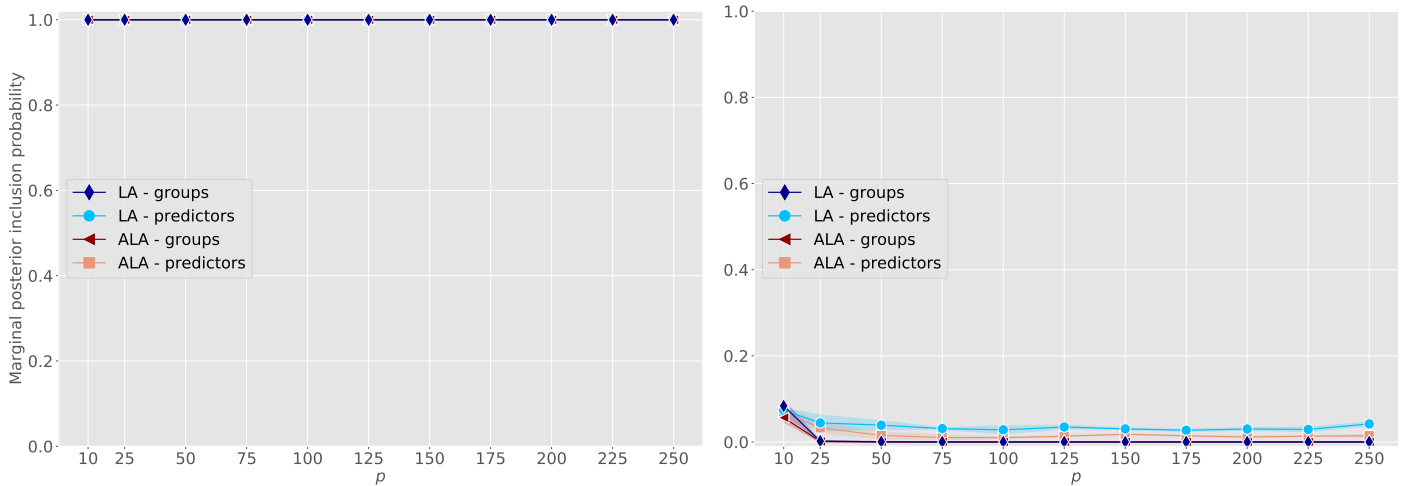


FIGURE 5.2: Comparison of ALA and LA for posterior inclusion probabilities of groups and predictors when p varies

Comparison of ALA and LA for posterior inclusion probabilities of groups and predictors when p varies from 10 to 25, with $n = 1,500$. Left: average posterior inclusion probabilities for truly active variables. Right: average posterior inclusion for truly inactive variables.

Figure 5.1 summarizes the results from the first scenario. Both LA and ALA discriminate properly truly active from inactive groups and variables when n is large enough. However, LA assigns larger inclusion probabilities for truly variables when $n \leq 500$. Figure 5.2 summarizes the results for the $n = 1,500$ case, when p varies from 10 to 25. LA and ALA performed equally and provided accurate estimates both for groups and variables.

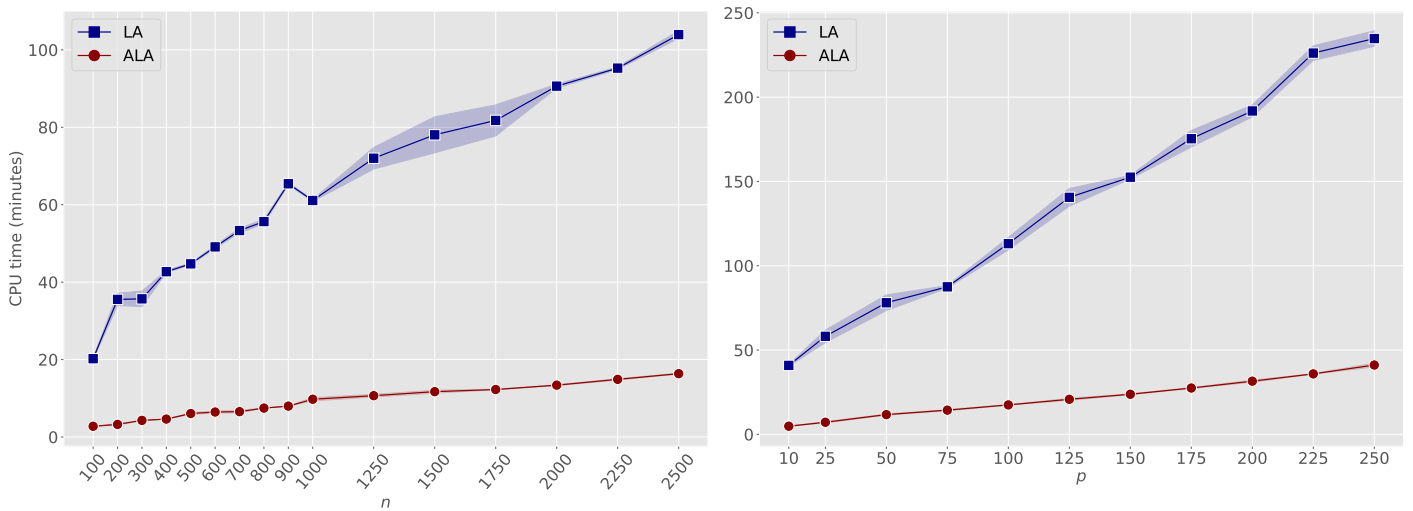


FIGURE 5.3: Comparison of ALA and LA for run time of waste-free SMC

.Left: average run time when n varies from 100 to 2,500 ($p = 50$). Right: average run time when p varies from 10 to 25 ($n = 1,500$).

Figure 5.3 compares the performance of ALA and LA in terms of computation time in both scenarios. ALA significantly reduces run times compared to LA, especially for larger n (mean run time = 16 min for ALA vs. 102 min for LA when $n = 2,500$ and $p = 50$) and p (mean run time = 39 min for ALA vs. 330 min for LA when $p = 250$ and $n = 1,500$). It is interesting to note that the CPU time still grows with n with ALA, although the computation of ALA is independent of n . The likely explanation is that when n grows, the prior and

the posterior differ more markedly, and thus more intermediate tempering distributions are required to bridge between the two. Still, the dependence on n of the CPU time remains mild compared to the LA-based sampler.

To sum up, one observes that ALA considerably reduces the CPU time of the sampler, in particular for large n (sample size) and p (number of variables). In return, as expected ALA tends to under-estimate the probability of inclusion of active variables, at least for n not sufficiently large.

Bi-level selection on the French National Healthcare Insurance database

To examine the performance of our SMC sampler on a big dataset, we study which factors are associated to sudden cardiac death (SCD) in a French epidemiological study. Sudden cardiac death is an unexpected death due to cardiac causes that occurs in a short time period (generally within 1 hour of symptom onset) in a person with known or unknown cardiac disease. Despite progress in epidemiology, clinical profiling and interventions, it remains a major public health problem worldwide, accounting for 10 to 20% of deaths in industrialised countries. The annual incidence of SCD is estimated 180,000 to 450,000 in the United States (Melissa et al. [2011]) and 275,000 in Europe (Empana et al. [2022]). The prognosis is terrible, with less than 10% surviving to hospital discharge, and significant functional and cognitive disabilities often persist among those who survive (Bougouin et al. [2014]). Therefore, identification of persons with an elevated risk of SCD is highly relevant from a clinical and public health perspective.

In this study, we implement bi-level variable selection to identify outpatient drugs and hospital diagnoses that could help to enhance risk prediction performance of SCD over many potential risk factors collected from electronic health records. We analyse the medical trajectories of $n_{\text{cases}} = 23,958$ cases of SCD collected between 2016 and 2020 throughout the Paris Sudden Death Expertise Center (Bougouin et al. [2014]), and $n_{\text{controls}} = 23,958$ controls sampled from the French general population. Cases and controls were matched with age, sex and residence area.

For the $n = n_{\text{cases}} + n_{\text{controls}} = 47,916$ individuals, we collected data from the French National Health Insurance (SNDS) database, which manages all reimbursements of healthcare for people affiliated to a health insurance scheme in France. It provides information on all healthcare expenses, on an individual level, including visits, procedures and reimbursed drugs relative to outpatient medical care claims, information from hospital discharge summaries and chronic conditions. Data acquisition is permanent, from birth to death, irrespective of wealth, age, or work status, resulting in one of the largest electronic health records databases in the world. The SNDS database has been described in detail previously and has been used to conduct multiple studies in cardiovascular epidemiology (Piot et al. [2022]). More details are available at <https://www.health-data-hub.fr/>.

We collected all outpatient drugs and hospital diagnoses that occurred up to 5 years before SCD; in this way we obtained $q = 36$ groups and $p = 337$ binary variables (0/1 whether the individual took a particular drug in the last 5 years, or a drug in the corresponding group). In the 36 groups, the minimum number of variables observed is 2 and the maximum is 27. No external variables were included in the study ($r = 0$). Figures 5.4 and 5.5 summarise the results of our ALA-based SMC sampler in terms of variable (and group) selection. We evaluate groups and variables selected by our model by comparing them with those described in the medical literature related to SCD. Overall, 16 out of 36 groups and 55

out of 337 variables are selected (Figure 5.4). Our bi-level variable selection scheme allows for a more flexible structure than "all-in all-out" methods and identifies 3 different "clusters" represented in Figure 5.5.

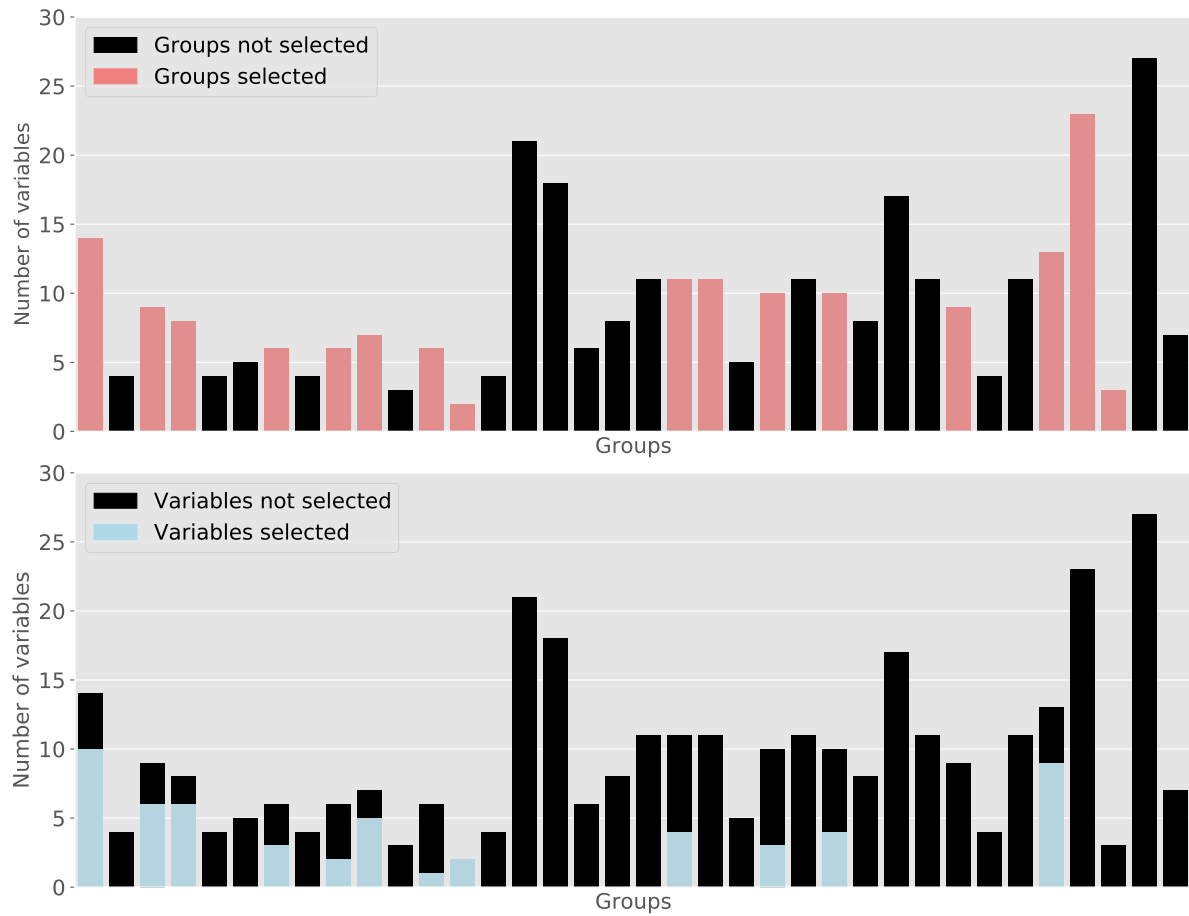


FIGURE 5.4: Groups and predictors selected by the ALA-based SMC sampler
 Top: selection of groups. Bottom: selection of predictors.

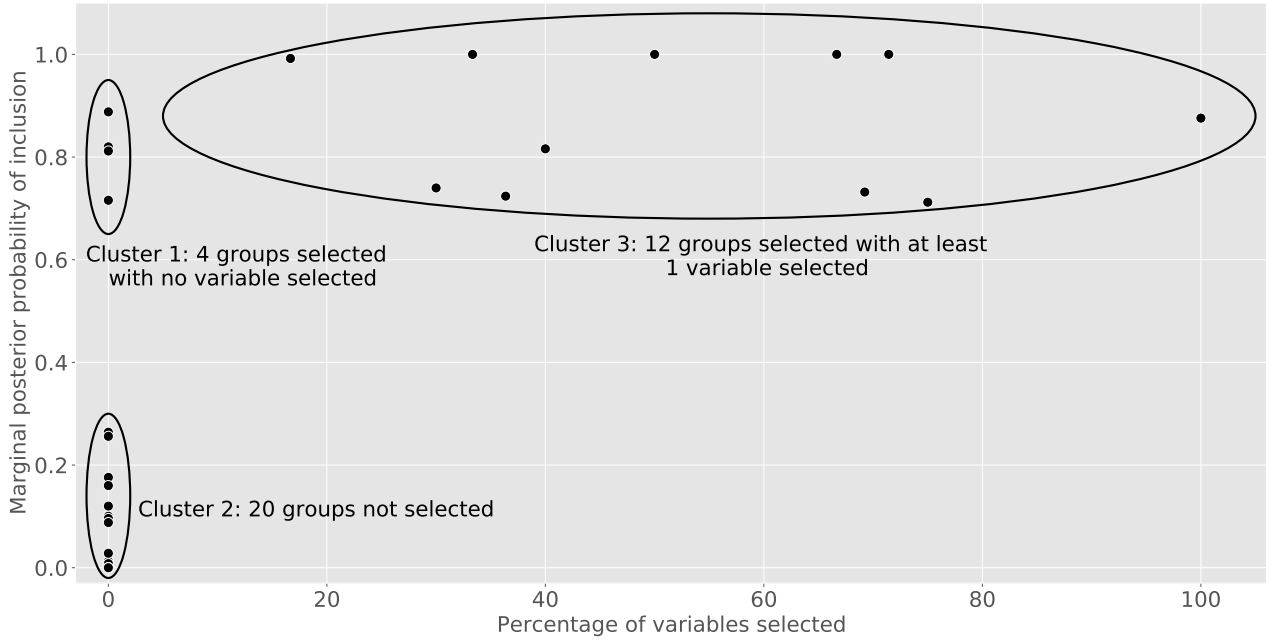


FIGURE 5.5: Bi-level variable selection scheme proposed by the ALA-based SMC sampler

In the first cluster (located in the upper left corner), 4 groups of hospital diagnoses are selected without any variable included. These groups correspond to diseases of the eye ($\pi(\gamma_k = 1) = 0.82$), diagnoses related to pregnancy, childbirth and the puerperium ($\pi(\gamma_k = 1) = 0.82$), injury and poisoning ($\pi(\gamma_k = 1) = 0.72$) and diagnoses for other special purposes ($\pi(\gamma_k = 1) = 0.89$). They are selected with high marginal posterior probabilities of inclusion, although none of their 46 corresponding variables are selected. This result suggests therefore that only global relationships exist between these groups and SCD, with no any precise effect of diseases or treatments.

In the second cluster (located in the lower left corner), 20 groups are not selected, as well as their 189 corresponding variables. They include diverse subgroups of diseases and treatments.

In the third cluster (located in the upper right corner), 12 groups are selected with at least 1 variable included. Among them, 3 well known groups of risk factors of SCD are identified. First, diseases and drugs associated to the cardiovascular system are selected (with $\pi(\gamma_k = 1) = 0.74$ and $\pi(\gamma_k = 1) = 1$ respectively), including 9 out of 19 variables. This result was expected, as cardiovascular conditions are known to be the most common pathology under SCD. Second, diseases and drugs related to the nervous system are selected (with $\pi(\gamma_k = 1) = 0.72$ and $\pi(\gamma_k = 1) = 1$ respectively), including 9 out of 18 variables. Several studies have suggested relationships between diseases of the nervous system and SCD (Japundzic-Zigon et al. [2018]). Indeed, some neurological disorders can cause damage to the heart and blood vessels (such as stroke or brain injury) or arrhythmia (such as epilepsy), increasing the risk of SCD. There are also neurological conditions that can cause SCD directly, such as long QT or Brugada syndromes, which affect the electrical activity of the heart. Third, a group related to treatments of the respiratory system is selected. A number of studies have also addressed the relationship between respiratory disorders and SCD. In particular, cumulating evidence associates chronic obstructive pulmonary diseases with an increased risk of SCD both in cardiovascular patient groups and in community-based

studies, independent of cardiovascular risk profile (Van den Berg et al. [2016]).

We ran our ALA-based SMC samplers 10 times to assess its numerical stability. Figure 5.6 describes the interquartile range of the marginal posterior probabilities of inclusion for variables. The mean run time was 61.8 hours (totalling to 7 days of total CPU time). We also launched 10 executions of our LA-based SMC sampler, but these executions had not completed after 30 days. We can see that, for this particular dataset, using ALA becomes crucial to make the approach usable for practitioners.

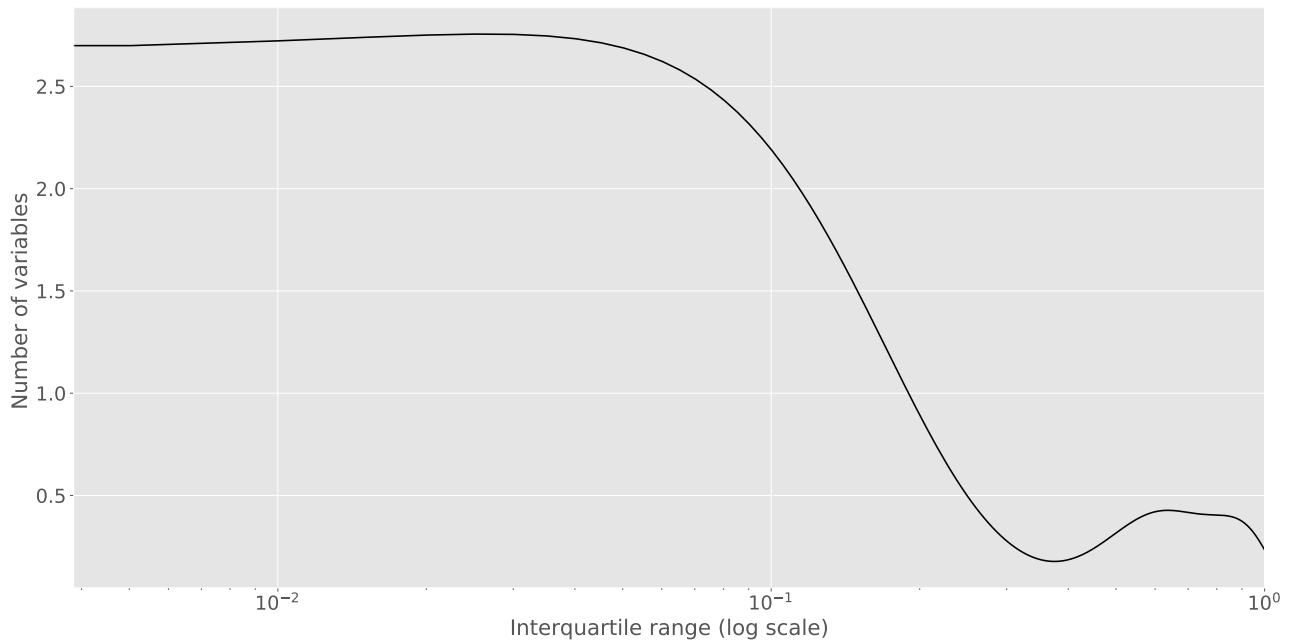


FIGURE 5.6: Kernel density estimate of the interquartile range (log scale) of the marginal posterior inclusion probabilities (variables) for the ALA-based SMC sampler.

5.5 Conclusion

Our bi-level variable selection approach based on a waste-free SMC sampler and the ALA approximation offers reliable performance for large-scale datasets within a reasonable computation time. Furthermore, our approach is more flexible than most of existing schemes, which impose only “all-in” or “all-out” selection for variables in the same group. This work could be therefore helpful in a wide range of applications, such as biomedical studies, where standard approaches provide information which may be difficult for physicians to interpret.

Appendices

Appendix A

Prediction Model

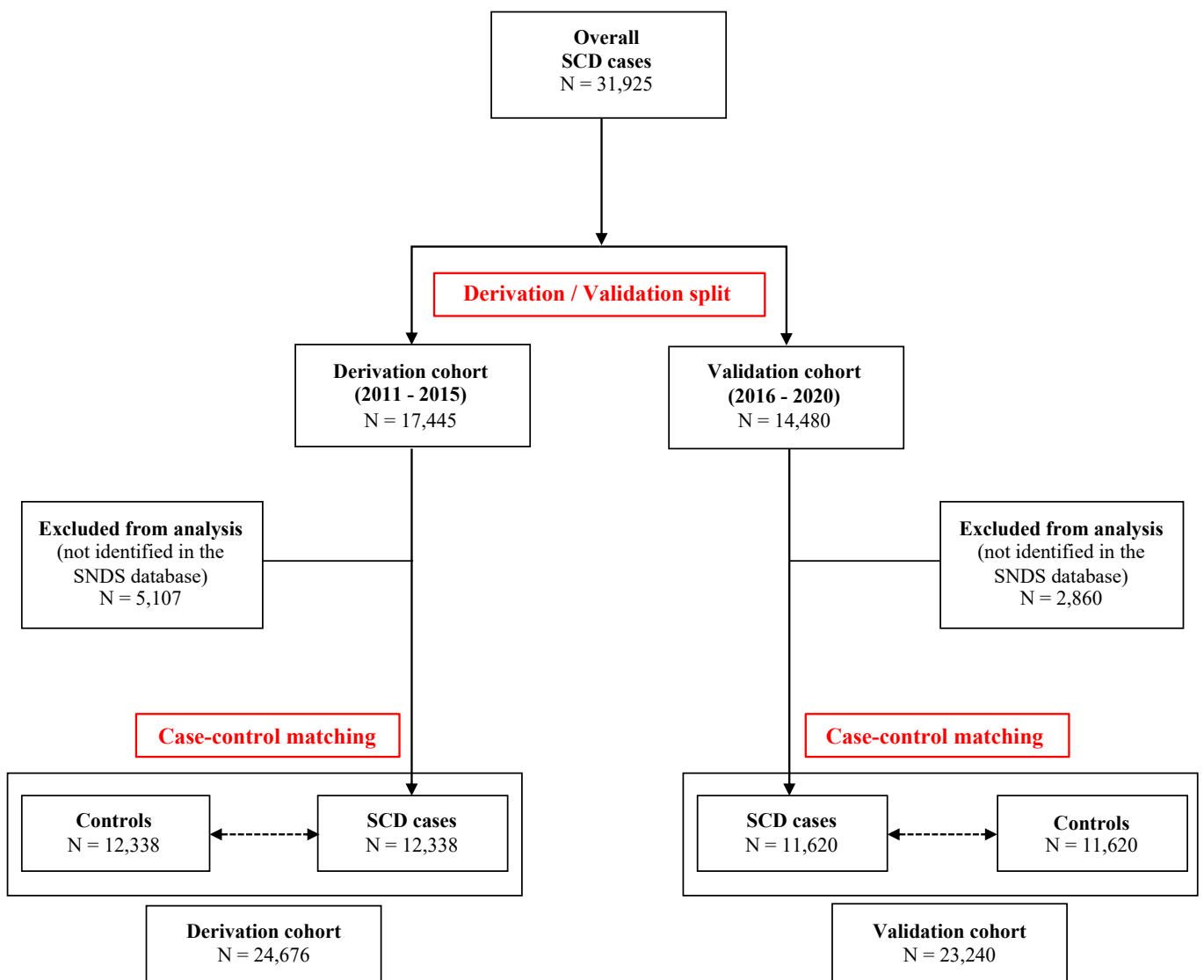


FIGURE A.1: Flow chart of the populations

TABLE A.1: Medical codes used for the CVD model

Covariate	ICD-10 / ATC codes	Literature review of SCD prediction models
Age	Available in the SNDS database but not included in the CVD model (used for the case-control matching)	Aro et al. 2017, Bogle et al. 2018, Deo et al. 2016, Waks et al. 2016, Holkeri et al. 2020
Sex	Available in the SNDS database but not included in the CVD model (used for the case-control matching)	Aro et al. 2017, Bogle et al. 2018, Deo et al. 2016, Waks et al. 2016, Holkeri et al. 2020
Race	Not available in the SNDS database	Deo et al. 2016, Waks et al. 2016
Left ventricular ejection fraction	Not available in the SNDS database	Aro et al. 2017
Electrocardiogram signals	Not available in the SNDS database	Aro et al. 2017, Deo et al. 2016, Waks et al. 2016, Holkeri et al. 2020
Coronary artery disease	ICD-10: angina pectoris (I20) ; acute myocardial infarction (I21) ; subsequent myocardial infarction (I22) ; certain current complications following acute myocardial infarction (I23) ; other acute ischaemic heart diseases (I24) ; chronic ischaemic heart disease (I25)	Waks et al. 2016, Holkeri et al. 2020
Stroke	ICD-10: subarachnoid haemorrhage (I60) ; intracerebral haemorrhage (I61) ; other nontraumatic intracranial haemorrhage (I62) ; cerebral infarction (I63) ; stroke, not specified as haemorrhage or infarction (I64)	Waks et al. 2016
Diabetes	ICD-10: type 1 diabetes mellitus (E10) ; type 2 diabetes mellitus (E11) ; malnutrition-related diabetes mellitus (E12) ; other specified diabetes mellitus (E13) ; unspecified diabetes mellitus (E14) ; diabetic mononeuropathy (G590) ; diabetic polyneuropathy (G632) ; myasthenic syndromes in endocrine diseases (G730) ; autonomic neuropathy in endocrine and metabolic diseases (G990) ; diabetic cataract (H280) ; diabetic retinopathy (H360) ; diabetic arthropathy (M142) ; glomerular disorders in diabetes mellitus (N083) ATC: drugs used in diabetes (A10)	Aro et al. 2017, Bogle et al. 2018, Deo et al. 2016, Waks et al. 2016, Holkeri et al. 2020
Hypertension	ICD-10: essential (primary) hypertension (I10) ; hypertensive heart disease (I11) ; hypertensive renal disease (I12) ; hypertensive heart	Aro et al. 2017, Bogle et al. 2018, Deo et al. 2016, Waks et al. 2016

	and renal disease (I13) ; secondary hypertension (I15) ; hypertensive encephalopathy (I674) ATC: antihypertensives (C02)	
Smoking abuse	ICD-10: nicotine dependence (F17) ; toxic effect of tobacco and nicotine (T652) ; exposure to tobacco smoke (Z587) ; tobacco abuse counselling (Z716) ; tobacco use (Z720) ATC: drugs used in nicotine dependence (N07BA)	Bogle et al. 2018, Deo et al. 2016
Obesity	ICD-10: overweight and obesity (E66) ; mechanical complication of gastrointestinal prosthetic devices, implants and grafts for obesity (T8550) ATC: antiobesity, preparations, excluding diet products (A08)	Bogle et al. 2018, Holkeri et al. 2020
Lipid disorders	ICD-10: disorders of lipoprotein metabolism and other lipidemias (E78) ATC: lipid modifying agents (C10)	Bogle et al. 2018, Deo et al. 2016, Holkeri et al. 2020
Systolic blood pressure	Not available in the SNDS database	Bogle et al. 2018, Deo et al. 2016, Holkeri et al. 2020
Diastolic blood pressure	Not available in the SNDS database	Bogle et al. 2018
Serum potassium	Not available in the SNDS database	Deo et al. 2016
Serum albumin	Not available in the SNDS database	Deo et al. 2016
Chronic renal failure	ICD-10: type 1 diabetes mellitus with renal complications (E102) ; type 2 diabetes mellitus with renal complications (E112) ; other specified diabetes mellitus with renal complications (E132) ; unspecified diabetes mellitus with renal complications (E142) ; hypertensive renal disease (I12) ; hypertensive heart and renal disease with renal failure (I131) ; hypertensive heart and renal disease with both (congestive) heart failure and renal failure (I132) ; chronic nephritic syndrome : diffuse membranous glomerulonephritis (N032) ; chronic nephritic syndrome : diffuse mesangial proliferative glomerulonephritis (N033) ; chronic nephritic syndrome : diffuse endocapillary proliferative	Deo et al. 2016

	<p>glomerulonephritis (N034) ; chronic nephritic syndrome : diffuse mesangiocapillary glomerulonephritis (N035) ; chronic nephritic syndrome : dense deposit disease (N036) ; chronic nephritic syndrome : diffuse crescentic glomerulonephritis (N037) ; unspecified nephritic syndrome : diffuse membranous glomerulonephritis (N052) ; unspecified nephritic syndrome : diffuse mesangial proliferative glomerulonephritis (N053) ; unspecified nephritic syndrome : diffuse endocapillary proliferative glomerulonephritis (N054) ; unspecified nephritic syndrome : diffuse mesangiocapillary glomerulonephritis (N055) ; unspecified nephritic syndrome : dense deposit disease (N056) ; unspecified nephritic syndrome : diffuse crescentic glomerulonephritis (N057) ; glomerular disorders in diabetes mellitus (N083) ; chronic kidney disease (N18) ; unspecified kidney failure (N19) ; renal osteodystrophy (N250) ; care involving dialysis (Z49) ; kidney transplant status (Z940) ; dependence on renal dialysis (Z992)</p>	
--	---	--

FIGURE A.2: Medical codes used for baseline characteristics of the populations

Diagnoses / treatments	ICD-10 / ATC codes	Literature review
Cardiovascular risk factors		
Smoking abuse	<p>ICD-10: nicotine dependence (F17) ; toxic effect of tobacco and nicotine (T652) ; exposure to tobacco smoke (Z587) ; tobacco abuse counselling (Z716) ; tobacco use (Z720)</p> <p>ATC: drugs used in nicotine dependence (N07BA)</p>	Oger et al. 2022, Lecoeur et al. 2022, Schapiro Dufour et al. 2019
Obesity	<p>ICD-10: overweight and obesity (E66) ; mechanical complication of gastrointestinal prosthetic devices, implants and grafts for obesity (T8550)</p> <p>ATC: antiobesity, preparations, excluding diet products (A08)</p>	Lecoeur et al. 2022, Goulabchand et al. 2021, Boucheron et al. 2021, Mohammadi et al. 2021, Schapiro Dufour et al. 2019, Rachas et al. 2022
Dyslipidaemia	<p>ICD-10: disorders of lipoprotein metabolism and other lipidemias (E78)</p>	Lecoeur et al. 2022
Diabetes	<p>ICD-10: type 1 diabetes mellitus (E10) ; type 2 diabetes mellitus (E11) ; malnutrition-related diabetes mellitus (E12) ; other specified diabetes mellitus (E13) ; unspecified diabetes mellitus (E14) ; diabetic mononeuropathy (G590) ; diabetic polyneuropathy (G632) ; myasthenic syndromes in endocrine diseases (G730) ; autonomic neuropathy in endocrine and metabolic diseases (G990) ; diabetic cataract (H280) ; diabetic retinopathy (H360) ; diabetic arthropathy (M142) ; glomerular disorders in diabetes mellitus (N083)</p> <p>ATC: drugs used in diabetes (A10)</p>	Lecoeur et al. 2022, Giral et al. 2019, Oger et al. 2022, Goulabchand et al. 2021, Zerah et al. 2021, Mohammadi et al. 2021, Schapiro Dufour et al. 2019, Pugnet et al. 2016
Hypertension	<p>ICD-10: essential (primary) hypertension (I10) ; hypertensive heart disease (I11) ; hypertensive renal disease (I12) ; hypertensive heart and renal disease (I13) ; secondary hypertension (I15) ; hypertensive encephalopathy (I674)</p> <p>ATC: antihypertensives (C02)</p>	Goulabchand et al. 2021, Mohammadi et al. 2021, Gabet et al. 2019, Schapiro Dufour et al. 2019, Pugnet et al. 2016, Lecoeur et al. 2022
Cardiovascular diseases		

<p>Acute coronary syndrome</p>	<p>ICD-10: unstable angina (I200) ; acute myocardial infarction (I21) ; subsequent myocardial infarction (I22) ; certain current complications following acute myocardial infarction (I23) ; other acute ischaemic heart diseases (I24)</p>	<p>Lam et al. 2022, Boucheron et al. 2021, Mohammedi et al. 2021, Gabet et al. 2019, Blin et al. 2018, Bezin et al. 2017, Rachas et al. 2022</p>
<p>Heart failure</p>	<p>ICD-10: hypertensive heart disease with (congestive) heart failure (I110) ; hypertensive heart and renal disease with (congestive) heart failure (I130) ; hypertensive heart and renal disease with both (congestive) heart failure and renal failure (I132) ; hypertensive heart and renal disease, unspecified (I139) ; heart failure (I50) ; pulmonary oedema (J81) ; chronic passive congestion of liver (K761)</p>	<p>Boucheron et al. 2021, Zerah et al. 2021, Gabet et al. 2019, Schapiro Dufour et al. 2019, Lam et al. 2022, Rachas et al. 2022</p>
<p>Peripheral arterial diseases</p>	<p>ICD-10: atherosclerosis (I70) ; aortic aneurysm and dissection (I71) ; other aneurysm and dissection (I72) ; other peripheral vascular diseases (I73) ; arterial embolism and thrombosis (I74) ; other disorders of arteries and arterioles (I77) ; diseases of capillaries (I78) ; disorders of arteries, arterioles and capillaries in diseases classified elsewhere (I79)</p>	<p>Goulabchand et al. 2021, Pugnet et al. 2016</p>
<p>Cardiac arrhythmias and conduction disorders</p>	<p>ICD-10: atrioventricular and left bundle-branch block (I44) ; other conduction disorders (I45) ; paroxysmal tachycardia (I47) ; atrial fibrillation and flutter (I48) ; other cardiac arrhythmias (I49)</p>	<p>Lam et al. 2022, Rachas et al. 2022</p>
<p>Valvular disease</p>	<p>ICD-10: rheumatic mitral valve diseases (I05) ; rheumatic aortic valve diseases (I06) ; rheumatic tricuspid valve diseases (I07) ; multiple valve diseases (I08) ; nonrheumatic mitral valve disorders (I34) ; nonrheumatic aortic valve disorders (I35) ; nonrheumatic tricuspid valve disorders (I36) ; pulmonary valve disorders (I37) ; endocarditis, valve unspecified (I38) ; endocarditis and heart valve disorders in diseases classified elsewhere (I39)</p>	<p>Zerah et al. 2021, Rachas et al. 2022</p>
<p>Pulmonary embolism</p>	<p>ICD-10: pulmonary embolism (I26) ; primary pulmonary hypertension (I270)</p>	<p>Lecoeur et al. 2022, Gouverneur et al. 2022</p>

Acute stroke	ICD-10: subarachnoid haemorrhage (I60) ; intracerebral haemorrhage (I61) ; other nontraumatic intracranial haemorrhage (I62) ; cerebral infarction (I63) ; stroke, not specified as haemorrhage or infarction (I64)	Lam et al. 2022, Gouverneur et al. 2022, Goulabchand et al. 2021, Mohammedi et al. 2021, Gabet et al. 2019
Cardiovascular drug used		
Angiotensin-converting enzyme inhibitors or angiotensin receptor blockers or aliskiren	ATC: ACE inhibitors, plain (C09A) ; ACE inhibitors, combinations (C09B) ; angiotensin II receptor blockers, plain (C09C) ; angiotensin II receptor blockers, combinations (C09D) ; aliskiren (C09XA02)	Giral et al. 2019
Diuretics	ATC: reserpine and diuretics (C02LA01) ; diuretics (C03) ; oxprenolol and thiazides (C07BA02) ; beta blocking agents, selective, and thiazides (C07BB) ; pindolol and other diuretics (C07CA03) ; timolol, thiazides and other diuretics (C07DA06) ; amlodipine and diuretics (C08GA02) ; ACE inhibitors and diuretics (C09BA) ; angiotensin II receptor blockers and diuretics (C09DA) ; aliskiren and hydrochlorothiazide (C09XA52) ; acetazolamide (S01EC01)	Giral et al. 2019
Beta-blockers	ATC: beta blocking agents (C07)	Giral et al. 2019
Calcium channel blockers	ATC: beta blocking agents and calcium channel blockers (C07FB) ; calcium channel blockers (C08) ; ACE inhibitors and calcium channel blockers (C09BB) ; angiotensin II receptor blockers and calcium channel blockers (C09DB) ; atorvastatin and amlodipine (C10BX03)	Giral et al. 2019
Antiarrhythmic agents	ATC: antiarrhythmics, class I and III (C01B)	Giral et al. 2019
Aspirin	ATC: acetylsalicylic acid (B01AC06) ; carbasalate calcium (B01AC08) ; platelet aggregation inhibitors excluding heparin, combinations (B01AC30) ; pravastatin and acetylsalicylic acid (C10BX02) ; acetylsalicylic acid (N02BA01)	Giral et al. 2019

Other antiplatelet agents	ATC: platelet aggregation inhibitors excluding heparin (B01AC) except for aspirin	Giral et al. 2019
Oral anticoagulants	ATC: vitamin K antagonists (B01AA) ; direct thrombin inhibitors (B01AE) ; direct factor Xa inhibitors (B01AF) ; other antithrombotic agents (B01AX)	Giral et al. 2019
Heparin	ATC: heparin group (B01AB) ; other antithrombotic agents (B01AX)	Giral et al. 2019
Insulin	ATC: insulins and analogues for injection, fast-acting (A10AB) ; insulins and analogues for injection, intermediate-acting (A10AC) ; insulins and analogues for injection, intermediate- or long-acting combined with fast-acting (A10AD) : insulins and analogues for injection, long-acting (A10AE)	Giral et al. 2019
Oral antidiabetic agents	ATC: drugs used in diabetes (A10) except for insulin and benfluorex (A10BX06)	Giral et al. 2019
Statins	ATC: HMG CoA reductase inhibitors (C10AA) ; combinations of various lipid modifying agents (C10BA) ; lipid modifying agents in combination with other drugs (C10BX)	Giral et al. 2019
Other lipid-lowering agents	ATC: fibrates (C10AB) ; bile acid sequestrants (C10AC) ; nicotinic acid and derivatives (C10AD) ; other lipid modifying agents (C10AX)	Giral et al. 2019
Comorbidities and lifestyle habits		

<p>Alcohol abuse</p>	<p>ICD-10: alcohol-induced pseudo-Cushing's syndrome (E244) ; alcohol related disorders (F10) ; alcoholic cardiomyopathy (I426), degeneration of nervous system due to alcohol (G312) ; alcoholic polyneuropathy (G621) ; alcoholic myopathy (G721) ; alcoholic gastritis (K292) ; alcoholic liver disease (K70) ; alcohol-induced acute pancreatitis (K852) ; alcohol-induced chronic pancreatitis (K860) ; toxic effect of alcohol (T51) ; intentional self-poisoning by and exposure to alcohol (X65) ; alcohol deterrents (Y573) ; evidence of alcohol involvement determined by blood alcohol level (Y90) ; alcohol involvement, not otherwise specified (Y919) ; alcohol rehabilitation (Z502) ; alcohol abuse counseling and surveillance (Z714) ; alcohol use (Z721)</p> <p>ATC: baclofen (M03BX01) ; drugs used in alcohol dependence (N07BB)</p>	<p>Oger et al. 2022, Zerah et al. 2021, Lecoeur et al. 2022, Schapiro Dufour et al. 2019, Krajden et al. 2010</p>
<p>Drug use</p>	<p>ICD-10: opioid related disorders (F11) ; cannabis related disorders (F12) ; sedative, hypnotic or anxiolytic related disorders (F13) ; cocaine related disorders (F14) ; hallucinogen related disorders (F16) ; other psychoactive substance related disorders (F19) ; findings of drugs and other substances, not normally found in blood (R78) ; poisoning by, adverse effect of and underdosing of narcotics and psychodysleptics (T40) ; poisoning by, adverse effect of and underdosing of psychostimulants (T436) ; drug rehabilitation (Z503) ; drug abuse counseling and surveillance (Z715) ; drug use (Z722) ; personal history of drug abuse (Z8641)</p>	<p>Lecoeur et al. 2022, Krajden et al. 2010</p>

<p>Chronic pulmonary disease</p>	<p>ICD-10: other specified pulmonary heart diseases (I278) ; pulmonary heart disease, unspecified (I279) ; bronchitis, not specified as acute or chronic (J40) ; simple and mucopurulent chronic bronchitis (J41) ; unspecified chronic bronchitis (J42) ; emphysema (J43) ; other chronic obstructive pulmonary disease (J44) ; asthma (J45) ; status asthmaticus (J46) ; bronchiectasis (J47) ; coalworker pneumoconiosis (J60) ; pneumoconiosis due to asbestos and other mineral fibres (J61) ; pneumoconiosis due to dust containing silica (J62) ; pneumoconiosis due to other inorganic dusts (J63) ; unspecified pneumoconiosis (J64) ; pneumoconiosis associated with tuberculosis (J65) ; airway disease due to specific organic dust (J66) ; hypersensitivity pneumonitis due to organic dust (J67) ; chronic respiratory conditions due to chemicals, gases, fumes and vapours (J684) ; chronic and other pulmonary manifestations due to radiation (J701) ; chronic drug-induced interstitial lung disorders (J703) ; chronic respiratory failure (J961)</p>	<p>Giral et al. 2019, Goulabchand et al. 2021, Rachas et al. 2022</p>
<p>Chronic kidney disease</p>	<p>ICD-10: type 1 diabetes mellitus with renal complications (E102) ; type 2 diabetes mellitus with renal complications (E112) ; other specified diabetes mellitus with renal complications (E132) ; unspecified diabetes mellitus with renal complications (E142) ; hypertensive renal disease (I12) ; hypertensive heart and renal disease with renal failure (I131) ; hypertensive heart and renal disease with both (congestive) heart failure and renal failure (I132) ; chronic nephritic syndrome : diffuse membranous glomerulonephritis (N032) ; chronic nephritic syndrome : diffuse mesangial proliferative glomerulonephritis (N033) ; chronic nephritic syndrome : diffuse endocapillary proliferative glomerulonephritis (N034) ; chronic nephritic syndrome : diffuse mesangiocapillary glomerulonephritis (N035) ; chronic nephritic syndrome : dense deposit disease (N036) ; chronic nephritic syndrome : diffuse crescentic glomerulonephritis (N037) ; unspecified nephritic syndrome : diffuse membranous glomerulonephritis (N052) ; unspecified nephritic syndrome : diffuse mesangial proliferative glomerulonephritis (N053) ; unspecified nephritic syndrome : diffuse endocapillary proliferative glomerulonephritis (N054) ; unspecified nephritic syndrome : diffuse mesangiocapillary glomerulonephritis (N055) ; unspecified nephritic syndrome : dense deposit disease (N056) ; unspecified nephritic syndrome : diffuse crescentic glomerulonephritis (N057) ; glomerular disorders in diabetes mellitus (N083) ; chronic kidney disease (N18) ; unspecified kidney failure (N19) ; renal osteodystrophy (N250) ; care involving dialysis (Z49) ; kidney transplant status (Z940) ; dependence on renal dialysis (Z992)</p>	<p>Lecoeur et al. 2022, Giral et al. 2019, Oger et al. 2022, Goulabchand et al. 2021, Zerah et al. 2021, Mohammedi et al. 2021, Gabet et al. 2019, Schapiro Dufour et al. 2019</p>

Active cancer	ICD-10: malignant neoplasms (C00-C97) ; in situ neoplasms (D00-D09) ; neoplasms of uncertain or unknown behaviour (D37-48) ; neoplasm of unspecified behavior of digestive system (D49)	Giral et al. 2019
----------------------	--	-------------------

	Derivation cohort (2011 - 2015)		Validation cohort (2016 - 2020)	
	SCD N = 12,338	Controls N = 12,338	SCD N = 11,620	Controls N = 11,620
Cardiovascular diseases				
Peripheral arterial diseases, N (%)	1,471 (11.9%)	688 (5.6%)	1,506 (13.0%)	461 (4.0%)
Valvular disease, N (%)	1,036 (8.4%)	405 (3.3%)	1,047 (9.0%)	305 (2.6%)
Pulmonary embolism, N (%)	630 (5.1%)	252 (2.0%)	636 (5.5%)	165 (1.4%)
Acute stroke, N (%)	910 (7.4%)	813 (6.6%)	925 (8.0%)	589 (5.1%)
Cardiovascular drug used				
Calcium channel blockers, N (%)	4,606 (37.3%)	3,966 (32.1%)	4,771 (41.1%)	3,451 (29.7%)
Antiarrhythmic agents, N (%)	1,805 (14.6%)	1,050 (8.5%)	1,692 (14.6%)	843 (7.3%)
Other antiplatelet agents, N (%)	2,086 (16.9%)	870 (7.1%)	1,813 (15.6%)	642 (5.5%)
Heparin, N (%)	3,117 (25.3%)	2,970 (24.1%)	2,759 (23.7%)	2,097 (18.0%)
Comorbidites and lifestyle habits				
Alcohol abuse, N (%)	1,716 (13.9%)	587 (4.8%)	1,431 (12.3%)	373 (3.2%)
Drug use, N (%)	318 (2.6%)	43 (0.3%)	306 (2.6%)	37 (0.3%)
Chronic pulmonary disease, N (%)	2,205 (17.9%)	784 (6.4%)	2,096 (18.0%)	499 (4.3%)

TABLE A.2: Baseline characteristics of the populations

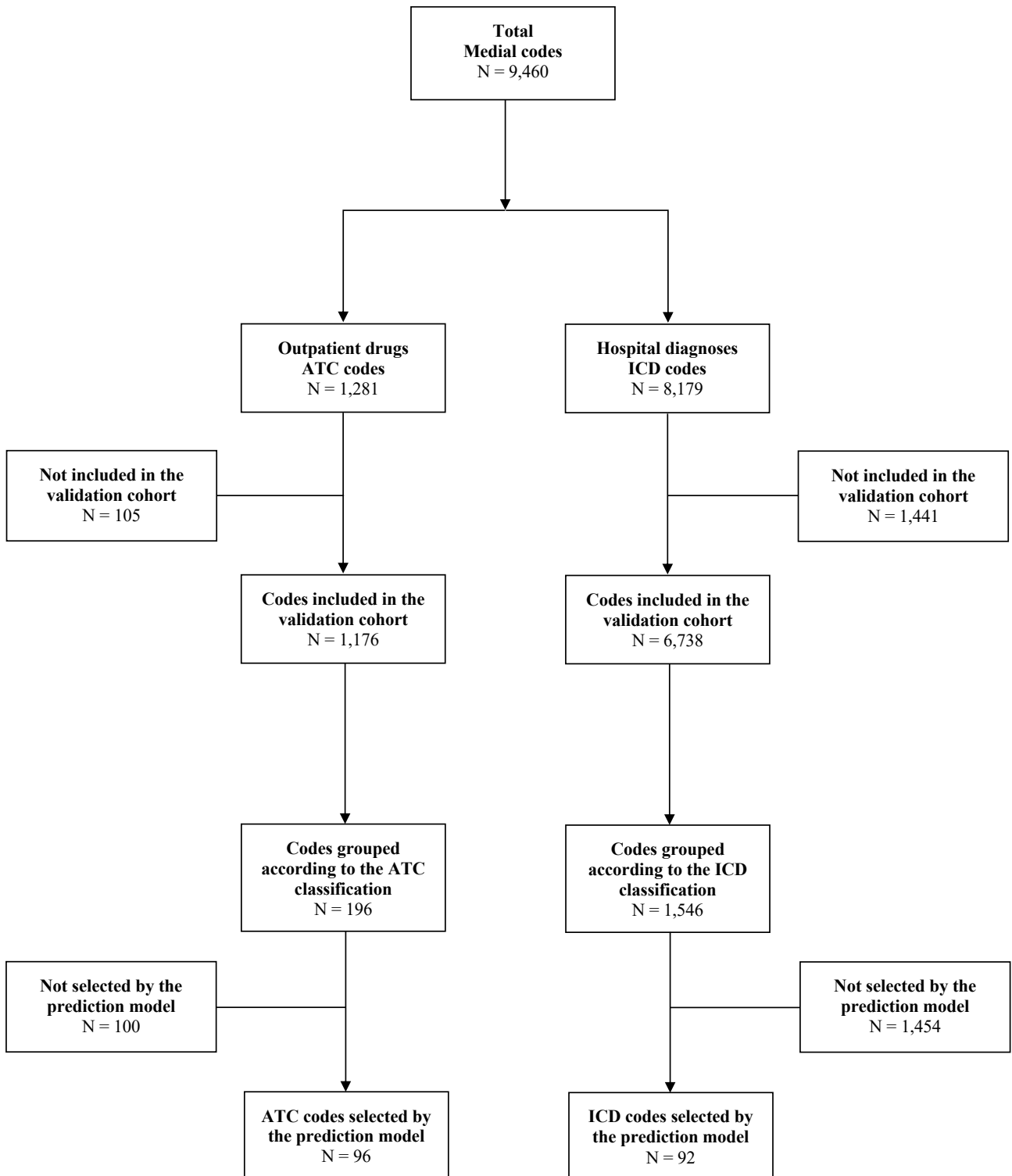


FIGURE A.3: Flow chart of the variables

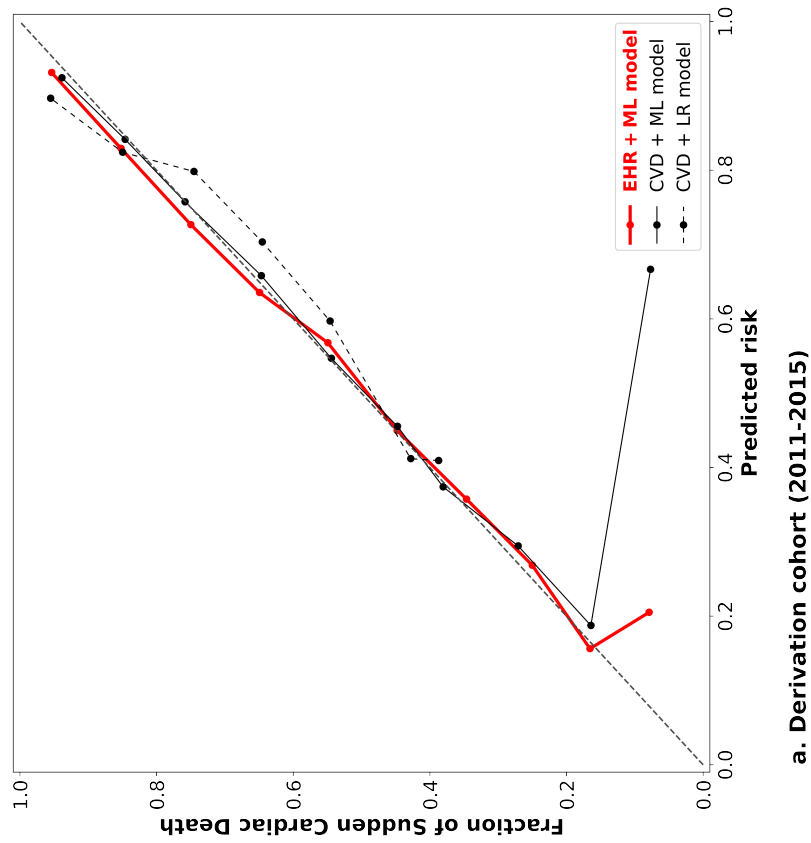
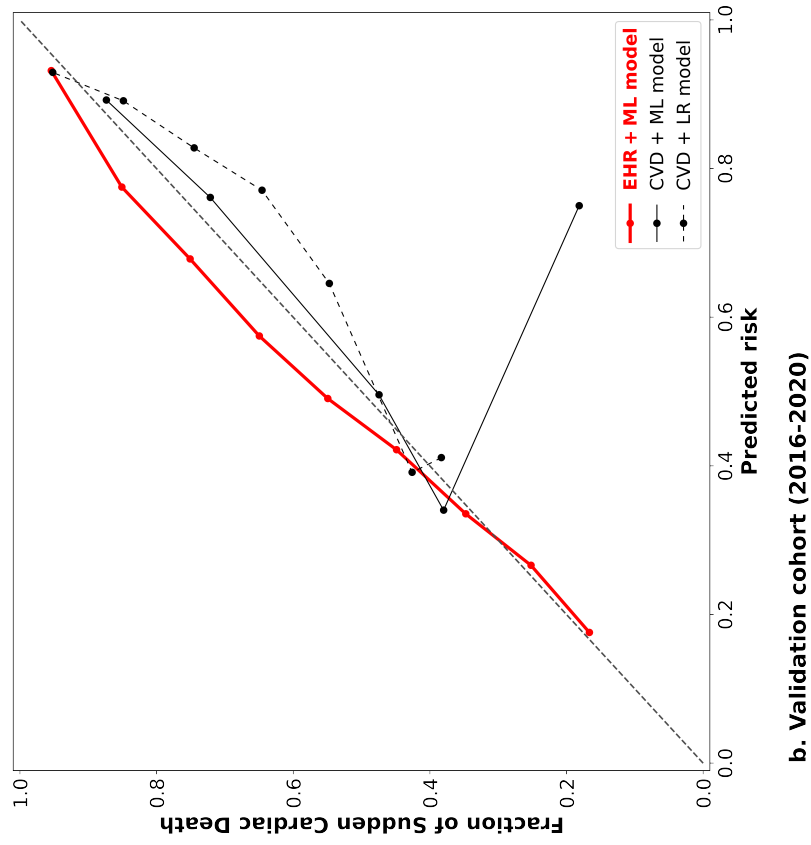


FIGURE A.4: Calibration plots

Model	Cohort	AUC (95% CI)	Sensitivity (%)	Positive Predictive Value (%)
EHR + CatBoost	Derivation	0.8 (0.78 - 0.82)	68%	77%
	Validation	0.8 (0.77 - 0.81)	71%	73%
EHR + XGBoost	Derivation	0.8 (0.78 - 0.81)	70%	75%
	Validation	0.78 (0.76 - 0.8)	72%	71%
EHR + Random Forest	Derivation	0.8 (0.78 - 0.81)	68%	76%
	Validation	0.78 (0.76 - 0.8)	69%	73%
EHR + soft Voting Classifier	Derivation	0.8 (0.79 - 0.82)	70%	76%
	Validation	0.79 (0.77 - 0.8)	72%	71%
EHR + Decision Tree	Derivation	0.75 (0.74 - 0.77)	55%	77%
	Validation	0.73 (0.72 - 0.75)	54%	74%
EHR + Logistic Regression	Derivation	0.78 (0.76 - 0.8)	67%	74%
	Validation	0.77 (0.75 - 0.78)	70%	70%
CVD + CatBoost	Derivation	0.65 (0.63 - 0.67)	44%	71%
	Validation	0.69 (0.67 - 0.71)	47%	76%
CVD + Logistic Regression	Derivation	0.66 (0.64 - 0.68)	45%	71%
	Validation	0.7 (0.68 - 0.72)	47%	75%

TABLE A.3: Comparison of the predictive performances

Variable	Value	Cohort	AUC (95% CI)	Sensitivity (%)	Positive Predictive Value (%)
Age (median)	< 73.3	Derivation (N = 12,754)	0.81 (0.79 - 0.83)	64%	83%
		Validation (N = 11,250)	0.8 (0.77 - 0.82)	63%	82%
	> 73.3	Derivation (N = 11,922)	0.79 (0.77 - 0.81)	74%	72%
		Validation (N = 11,990)	0.77 (0.75 - 0.8)	77%	67%
Sex	Male	Derivation (N = 14,877)	0.8 (0.78 - 0.82)	66%	79%
		Validation (N = 14,056)	0.79 (0.76 - 0.81)	67%	75%
	Female	Derivation (N = 9,799)	0.8 (0.78 - 0.83)	71%	75%
		Validation (N = 9,184)	0.78 (0.75 - 0.81)	73%	70%
Social deprivation	Yes	Derivation (N = 1,386)	0.84 (CI 0.77 - 0.9)	73%	82%
		Validation (N = 1,475)	0.82 (0.74 - 0.88)	66%	79%
	No	Derivation (N = 23,284)	0.8 (0.78 - 0.82)	68%	77%
		Validation (N = 21,759)	0.78 (0.76 - 0.8)	70%	72%

TABLE A.4: Sensitivity analyses

Bibliography

- K Zeppenfeld, J Tfelt-Hansen, M de Riva, BG Winkel, ER Behr, NA Blom, P Charron, D Corrado, N Dagues, C de Chillou, L Eckardt, T Friede, KH Haugaa, M Hocini, PD Lambiase, E Marijon, JL Merino, P Peichl, SG Priori, T Reichlin, J Schulz-Menger, C Sticherling, S Tzeis, A Verstrael, and M Volterrani. 2022 esc guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death. *Eur Heart J*, 43(40):3997–4126, 2022.
- P Coumel. The atrial fibrillation thromboembolism syndrome: implications of the arrhythmogenic substrate. *Journal of the American College of Cardiology*, 32(2):296–304, 1999.
- I Jacobs, V Nadkarni, J Bahr, RA Berg, JE Billi, L Bossaert, P Cassan, A Coovadia, K D’Este, J Finn, H Halperin, A Handley, J Herlitz, R Hickey, A Idris, W Kloeck, GL Larkin, ME Mancini, P Mason, G Mears, K Monsieurs, W Montgomery, P Morley, G Nichol, J Nolan, K Okada, J Perlman, M Shuster, PA Steen, F Sterz, J Tibballs, S Timerman, T Truitt, and D Zideman. Cardiac arrest and cardiopulmonary resuscitation outcome reports: update and simplification of the utstein templates for resuscitation registries: a statement for healthcare professionals from a task force of the international liaison committee on resuscitation (american heart association, european resuscitation council, australian resuscitation council, new zealand resuscitation council, heart and stroke foundation of canada, interamerican heart foundation, resuscitation councils of southern africa). *Circulation*, 110(21):3385–97, 2004.
- RO Cummins, JP Ornato, WH Thies, and Pepe PE. Improving survival from sudden cardiac arrest: the "chain of survival" concept. a statement for health professionals from the advanced cardiac life support subcommittee and the emergency cardiac care committee, american heart association. *Circulation*, 83(5):1832–47, 1991.
- T Kitamura, K Kiyohara, T Sakai, T Iwami, C Nishiyama, K Kajino, T Nishiuchi, Y Hayashi, Y Katayama, K Yoshiya, and T Shimazu. Epidemiology and outcome of adult out-of-hospital cardiac arrest of non-cardiac origin in osaka: a population-based study. *BMJ Open*, 4(12), 2018.
- JT Gräsner, J Herlitz, RW Koster, F Rosell-Ortiz, L Stamatakis, and L Bossaert. Quality management in resuscitation—towards a european cardiac arrest registry (eureca). *Resuscitation*, 82(8):989–94, 2016.
- SS Chugh, K Reinier, C Teodorescu, A Evanado, E Kehr, M Al Samara, R Mariani, K Gunson, and J Jui. Epidemiology of sudden cardiac death: clinical and research implications. *Prog Cardiovasc Dis*, 51(3):213–28, 2008.
- SG Priori, C Blomström-Lundqvist, A Mazzanti, N Blom, M Borggrefe, J Camm, PM Elliott, D Fitzsimons, R Hatala, G Hindricks, P Kirchhof, K Kjeldsen, KH Kuck, A Hernandez-Madrid, N Nikolaou, TM Norekvål, C Spaulding, and DJ Van Veldhuisen. Esc guidelines for the management of patients with ventricular arrhythmias and the prevention

- of sudden cardiac death: The task force for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death of the european society of cardiology (esc). *Eur Heart J*, 36(41):2793–2867, 2015.
- SS Chugh, J Jui, K Gunson, EC Stecker, BT John, B Thompson, N Ilias, C Vickers, V Dogra, M Daya, J Kron, ZJ Zheng, G Mensah, and J McAnulty. Current burden of sudden cardiac death: multiple source surveillance versus retrospective death certificate-based review in a large u.s. community. *J Am Coll Cardiol*, 44(1268-75):6, 2009.
- D Mozaffarian, EJ Benjamin, AS Go, DK Arnett, MJ Blaha, M Cushman, S de Ferranti, JP Després, HJ Fullerton, VJ Howard, MD Huffman, SE Judd, BM Kissela, DT Lackland, JH Lichtman, LD Lisabeth, S Liu, RH Mackey, DB Matchar, DK McGuire, ER Mohler, CS Moy, P Muntner, ME Mussolino, K Nasir, RW Neumar, G Nichol, L Palaniappan, DK Pandey, MJ Reeves, CJ Rodriguez, PD Sorlie, J Stein, A Towfighi, TN Turan, SS Virani, JZ Willey, D Woo, RW Yeh, and MB Turner. Heart disease and stroke statistics—2015 update: a report from the american heart association. *Circulation*, 131(4), 2015.
- JP Empana, I Lerner, E Valentin, F Folke, B Böttiger, G Gislason, M Jonsson, M Ringh, F Beanton, W Bougouin, E Marijon, M Blom, H Tan, and X Jouven. Incidence of sudden cardiac death in the european union: a systematic review and meta-analysis. *J Am Coll Cardiol*, 79(18):1818–1827, 2022.
- G Nichol, E Thomas, CW Callaway, J Hedges, JL Powell, TP Aufderheide, T Rea, R Lowe, T Brown, J Dreyer, D Davis, A Idris, and I Stiell. Regional variation in out-of-hospital cardiac arrest incidence and outcome. *JAMA*, 300(12):1423–31, 2008.
- PA Meaney, BJ Bobrow, ME Mancini, J Christenson, AR de Caen, F Bhanji, BS Abella, ME Kleinman, DP Edelson, RA Berg, TP Aufderheide, V Menon, and M Leary. Cardiopulmonary resuscitation quality: improving cardiac resuscitation outcomes both inside and outside the hospital: a consensus statement from the american heart association. *Circulation*, 128(417-35):4, 2013.
- X Jouven, M Desnos, C Guerot, and P Ducimetière. Predicting sudden death in the population: the paris prospective study i. *Circulation*, 99(15):1978–83, 1999.
- BG Winkel, AG Holst, J Theilade, IB Kristensen, JL Thomsen, GL Ottesen, H Bundgaard, JH Svendsen, S Haunsø, and J Tfelt-Hansen. Nationwide study of sudden cardiac death in persons aged 1-35 years. *Eur Heart J*, 32(8):983–90, 2011.
- SS Chugh, K Reinier, T Singh, A Uy-Evanado, C Socoteanu, D Peters, R Mariani, K Gunson, and J Jui. Determinants of prolonged qt interval and their contribution to sudden death risk in coronary artery disease: the oregon sudden unexpected death study. *Circulation*, 119(5):663–70., 2004.
- MJ Ackerman, SG Priori, S Willems, C Berul, R Brugada, H Calkins, AJ Camm, PT Ellinor, M Gollob, R Hamilton, RE Hershberger, DP Judge, H Le Marec, WJ McKenna, E Schulze-Bahr, C Semsarian, JA Towbin, H Watkins, A Wilde, C Wolpert, and DP Zipes. Hrs/ehra expert consensus statement on the state of genetic testing for the channelopathies and cardiomyopathies. *Heart Rhythm*, 8(8):1308–39, 2011.
- X Jouven, JP Empana, PJ Schwartz, M Desnos, D Courbon, and P Ducimetière. Heart-rate profile during exercise as a predictor of sudden death. *Circulation*, 352(19):1951–8, 2005.
- CM Albert, MA Mittleman, CU Chae, IM Lee, CH Hennekens, and JE Manson. Triggering of sudden death from cardiac causes by vigorous exertion. *N Engl J Med*, 343(19):1355–61, 2000.

- BJ Maron. Sudden cardiac death in hypertrophic cardiomyopathy. *J Cardiovasc Transl Resl*, 2(4):368–80, 2009.
- C Basso, D Corrado, FI Marcus, A Nava, and G Thiene. Arrhythmogenic right ventricular cardiomyopathy. *Lancet*, 373(9671):1289–300, 2009.
- CR Bezzina, R Pazoki, A Bardai, RF Marsman, JSSG de Jong, MT Blom, BP Scicluna, JW Jukema, NR Bindraban, P Lichtner, A Pfeufer, NH Bishopric, DM Roden, T Meitinger, SS Chugh, RJ Myerburg, X Jouven, S Kääh, LRC Dekker, HL Tan, MWT Tanck, and Wilde AAM. Genome-wide association study identifies a susceptibility locus at 21q21 for ventricular fibrillation in acute myocardial infarction. *Nat Genet*, 42(8):688–691, 2010.
- AJ Moss, W Zareba, WJ Hall, H Klein, DJ Wilber, DS Cannom, JP Daubert, SL Higgins, MW Brown, and ML Andrews. Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. *N Engl J Med*, 346(12):877–83, 2002.
- SJ Pocock, D Wang, MA Pfeffer, S Yusuf, JJ McMurray, KB Swedberg, J Ostergren, EL Michelson, KS Pieper, and CB Granger. Predictors of mortality and morbidity in patients with chronic heart failure. *JAMA*, 27(1):65–75, 2006.
- FM Kusumoto, Barrett C Schoenfeld, MH, JR Edgerton, KA Ellenbogen, MR Gold, NF Goldschlager, RM Hamilton, JA Joglar, RJ Kim, R Lee, JE Marine, CJ McLeod, KR Oken, KK Patton, CN Pellegrini, KA Selzman, A Thompson, and PD Varosy. Acc/aha/hrs guideline on the evaluation and management of patients with bradycardia and cardiac conduction delay: A report of the american college of cardiology/american heart association task force on clinical practice guidelines and the heart rhythm society. *Circulation*, 140(9):382–482, 2019.
- K Narayanan, K Reinier, A Uy-Evanado, C Teodorescu, H Chugh, E Marijon, Gunson K, J Jui, and SS Chugh. Frequency and determinants of implantable cardioverter defibrillator deployment among primary prevention candidates with subsequent sudden cardiac arrest in the community. *Circulation*, 128(16):1733–8, 2013.
- RJ Myerburg, KM Kessler, and A Castellanos. Sudden cardiac death. structure, function, and time-dependence of risk. *Circulation*, 85, 1992.
- RJ Myerburg. Sudden cardiac death: exploring the limits of our knowledge. *J Cardiovasc Electrophysiol*, 12(3):369–81, 2001.
- R Deo, FL Norby, R Katz, N Sotoodehnia, S Adabag, CR DeFilippi, B Kestenbaum, LY Chen, SR Heckbert, AR Folsom, RA Kronmal, S Konety, KK Patton, D Siscovick, MG Shlipak, and A Alonso. Development and validation of a sudden cardiac death prediction model for the general population. *Circulation*, 134(11):806–16, 2016.
- JW Waks, CM Sitlani, EZ Soliman, M Kabir, E Ghafoori, ML Biggs, CA Henrikson, N Sotoodehnia, T Biering-Sørensen, SK Agarwal, DS Siscovick, WS Post, SD Solomon, AE Buxton, ME Josephson, and LG Tereshchenko. Global electric heterogeneity risk score for prediction of sudden cardiac death in the general population: The atherosclerosis risk in communities (aric) and cardiovascular health (chs) studies. *Circulation*, 133(23):2222–34, 2016.
- AL Aro, Reinier, Rusinaru K, Uy-Evanado C, Darouian A, Phan N, Mack D, J WJ, Jui, EZ Soliman, Tereshchenko LG, and SS Chugh. Electrical risk score beyond the left ventricular ejection fraction: prediction of sudden cardiac death in the oregon sudden unexpected

- death study and the atherosclerosis risk in communities study. *Eur Heart J*, 38(40):3017–3025, 2017.
- BM Bogle, H Ning, JJ Goldberger, S Mehrotra, and DM Lloyd-Jones. A simple community-based risk-prediction score for sudden cardiac death. *Am J Med*, 131(5):532–539, 2018.
- A Holkeri, A Eranti, MAE Haukilahti, T Kerola, TV Kenttä, JT Tikkanen, O Anttonen, K Nononen, T Seppänen, H Rissanen, M Heliövaara, P Knekt, MJ Junttila, HV Huikuri, and AL Aro. Predicting sudden cardiac death in a general population using an electrocardiographic risk score. *Heart*, 106(6):427–433, 2020.
- P Rajpurkar, E Chen, O Banerjee, and EJ Topol. Ai in health and medicine. *Nat Med*, 28(1):31–38, 2022.
- C Krittanawong, HUH Virk, S Bangalore, Z Wang, KW Johnson, R Pinotti, H Zhang, S Kaplin, B Narasimhan, T Kitai, U Baber, JL Halperin, and WHW Tang. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci Rep*, 10(1), 2020.
- V Houshyarifar and M Chehel Amirani. An approach to predict sudden cardiac death (scd) using time domain and bispectrum features from hrv signal. *Biomed Mater Eng*, 27(2-3):275–85, 2016.
- E Ebrahimzadeh, A Foroutan, M Shams, R Baradaran, L Rajabion, M Joulani, and F Fayaz. An optimal strategy for prediction of sudden cardiac death through a pioneering feature-selection approach from hrv signal. *Comput Methods Programs Biomed*, pages 19–36, 2019.
- JP Amezcua-Sanchez, M Valtierra-Rodriguez, H Adeli, and CA Perez-Ramirez. A novel wavelet transform-homogeneity model for sudden cardiac death prediction using ecg signals. *J Med Syst*, 42(10), 2018.
- F Meng, Z Zhang, X Hou, Z Qian, Y Wang, Y Chen, Y Wang, Y Zhou, Z Chen, X Zhang, J Yang, J Zhang, J Guo, K Li, L Chen, R Zhuang, H Jiang, W Zhou, S Tang, Y Wei, and J Zou. Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: study protocol for a retrospective multicentre registry in china. *BMJ Open*, 9(5), 2019.
- DH Jang, J Kim, YH Jo, JH Lee, JE Hwang, SM Park, DK Lee, I Park, D Kim, and H Chang. Developing neural network models for early detection of cardiac arrest in emergency department. *Am J Emerg Med*, 38(1):43–49, 2020.
- J Kim, Chaen M, HJ Chang, YA Kim, and E Park. Predicting cardiac arrest and respiratory failure using feasible artificial intelligence with simple trajectories of patient data. *J Clin Med*, 8(9), 2019.
- JM Kwon, Y Lee, Y Lee, S Lee, and J Park. An algorithm based on deep learning for predicting in-hospital cardiac arrest. *J Am Heart Assoc*, 7(13), 2018.
- Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlali, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, R. Andrew Taylor, Harlan M. Krumholz, and Dragomir Radev. Neural natural language processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46:100511, 2022.
- E Choi, A Schuetz, WF Stewart, and J Sun. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc*, 24(2):361–370, 2017.

- C Xiao, E Choi, and J Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc*, 25(10): 1419–1428, 2018.
- W Bougouin, L Lamhaut, E Marijon, D Jost, F Dumas, N Deye, F Beganton, JP Empana, E Chazelle, A Cariou, and X Jouven. Characteristics and prognosis of sudden cardiac death in greater paris: population-based approach from the paris sudden death expertise center (paris-sdec). *Intensive Care Med*, 40(6):846–54, 2014.
- C Maupain, W Bougouin, L Lamhaut, N Deye, JL Diehl, G Geri, MC Perier, F Beganton, E Marijon, X Jouven, A Cariou, and F Dumas. The cahp (cardiac arrest hospital prognosis) score: a tool for risk stratification after out-of-hospital cardiac arrest. *Eur Heart J*, 37(42): 3222–3228, 2016.
- P Jabre, W Bougouin, F Dumas, P Carli, C Antoine, L Jacob, B Dahan, F Beganton, JP Empana, E Marijon, N Karam, A Loupy, C Lefaucheur, D Jost, A Cariou, F Adnet, TD Rea, and X Jouven. Early identification of patients with out-of-hospital cardiac arrest with no chance of survival and consideration for organ donation. *Ann Intern Med*, 165(11):770–778, 2016.
- W Bougouin, F Dumas, N Karam, C Maupain, E Marijon, L Lamhaut, D Jost, G Geri, F Beganton, O Varenne, C Spaulding, and X Jouven. Should we perform an immediate coronary angiogram in all patients after cardiac arrest?: Insights from a large french registry. *JACC Cardiovasc Interv*, 11(3):249–256, 2018.
- W Bougouin, F Dumas, L Lamhaut, E Marijon, P Carli, A Combes, R Pirracchio, N Aissaoui, N Karam, N Deye, G Sideris, F Beganton, D Jost, A Cariou, and X Jouven. Extracorporeal cardiopulmonary resuscitation in out-of-hospital cardiac arrest: a registry study. *Eur Heart J*, 41(21):1961–1971, 2020.
- E Marijon, N Karam, D Jost, D Perrot, B Frattini, C Derkenne, A Sharifzadehgan, V Waldmann, F Beganton, K Narayanan, A Lafont, W Bougouin, and X Jouven. Out-of-hospital cardiac arrest during the covid-19 pandemic in paris, france: a population-based, observational study. *Lancet Public Health*, 5(8):e437–e43, 2020.
- G Moulis, M Lapeyre-Mestre, A Palmaro, G Pugnet, JL Montastruc, and L Sailler. Les bases de données françaises de l’assurance maladie: quel intérêt pour la recherche médicale ? *Rev Med Interne*, 36(6):411–7, 2015.
- J Bezin, M Duong, R Lassalle, C Droz, A Pariente, P Blin, and N Moore. The national health-care system claims databases in france, sniiram and egb: Powerful tools for pharmacoepidemiology. *Pharmacoepidemiol Drug Saf*, 26(8):954–962, 2017.
- P Tuppin, J Rudant, P Constantinou, C Gastaldi-Ménager, A Rachas, L de Roquefeuil, G Maura, H Caillol, A Tajahmady, J Coste, C Gissot, A Weill, and A Fagot-Campagna. Value of a national administrative database to guide public decisions: From the système national d’information interrégimes de l’assurance maladie (sniiram) to the système national des données de santé (snds) in france. *Rev Epidemiol Sante Publique*, 65:S149–S167, 2017.
- A Revet, G Moulis, JP Raynaud, E Bui, and M Lapeyre-Mestre. Use of the french national health insurance information system for research in the field of mental health: Systematic review and perspectives. *Fundam Clin Pharmacol*, 36(1):16–34, 2022.

- P Tuppin, S Rivière, A Rigault, S Tala, J Drouin, L Pestel, P Denis, C Gastaldi-Ménager, C Gissot, Y Juillièrè, and A Fagot-Campagna. Prevalence and economic burden of cardiovascular diseases in france in 2013 according to the national health insurance scheme database. *Arch Cardiovasc Dis*, 109(6-7):399 –411, 2016.
- A Weill, M Dalichampt, F Raguideau, P Ricordeau, PO Blotière, J Rudant, F Alla, and M Zureik. Low dose oestrogen combined oral contraception and risk of pulmonary embolism, stroke, and myocardial infarction in five million french women: cohort study. *BMJ*, 353:i2002, 2016.
- P Giral, A Neumann, A Weill, and J Coste. Cardiovascular effect of discontinuing statins for primary prevention at the age of 75 years: a nationwide population-based cohort study in france. *Eur Heart J*, 40(43):3516 –3525, 2019.
- SF Feldman, T Lesuffleur, V Olié, C Gastaldi-Ménager, Y Juillièrè, and P Tuppin. French annual national observational study of 2015 outpatient and inpatient healthcare utilization by approximately half a million patients with previous heart failure diagnosis. *Arch Cardiovasc Dis*, 114(1):17 –32, 2021.
- O Piot, P Defaye, J Lortet-Tieulent, JC Deharo, J Beisel, A Vainchtock, C Leboucher, E Marjion, and S Boveda. Healthcare costs in implantable cardioverter-defibrillator recipients: A real-life cohort study on 19,408 patients from the french national healthcare database. *Int J Cardiol*, 348:39 –44, 2022.
- E Lecoeur, O Domeng, A Fayol, AS Jannot, and JS Hulot. Epidemiology of heart failure in young adults: a french nationwide cohort study. *Eur Heart J*, 44(5):383 –392, 2023.
- A Rachas, C Gastaldi-Ménager, P Denis, P Barthélémy, P Constantinou, J Drouin, D Lastier, T Lesuffleur, C Mette, M Nicolas, L Pestel, S Rivière, A Tajahmady, C Gissot, and A Fagot-Campagna. The economic burden of disease in france from the national health insurance perspective: The healthcare expenditures and conditions mapping used to prepare the french social security funding act and the public health act. *Med Care*, 60(9):655 –664, 2022.
- T Mikolov, K Chen, G Corrado, and J Dean. Efficient estimation of word representations in vector space. *Proceedings of the Workshop at ICLR*, 2013.
- H Steinhaus. Sur la division des corp materiels en partie. *Bull. Acad. Polon. Sci*, 804(1):801, 1956.
- SZ Selim and MA Ismail. K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. *IEEE Trans Pattern Anal Mach Intell.*, 6(1):81 –87, 1984.
- D Arthur and S Vassilvitskii. *k-means++: the advantages of careful seeding*. Society for Industrial and Applied Mathematics, 2007.
- A Dorogush, V Ershov, and A Gulin. Catboost: gradient boosting with categorical features support. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6638 –6648, 2018.
- Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *31st Conference on Neural Information Processing Systems*, 2017.
- HD Dau and N Chopin. Waste-free sequential monte carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):114–148, 2022.

- D Rossell, O Abril, and A Bhattacharya. Approximate Laplace approximations for scalable model selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 83(4):853–879, 2021.
- GI Fishman, SS Chugh, JP Dimarco, CM Albert, ME Anderson, RO Bonow, AE Buxton, PS Chen, M Estes, X Jouven, R Kwong, DA Lathrop, AM Mascette, JM Nerbonne, B O'Rourke, RL Page, DM Roden, DS Rosenbaum, N Sotoodehnia, NA Trayanova, and ZJ Zheng. Sudden cardiac death prediction and prevention: report from a national heart and lung and blood institute and heart rhythm society workshop. *Circulation*, 122(22):2335–48, 2010.
- LE Hinkle and HT Thaler. Clinical classification of cardiac deaths. *Circulation*, 65(3):457–64, 1982.
- RJ Myerburg and A Castellanos. *Sudden cardiac death*. Saunders Elsevier, 2009.
- M Hayashi, W Shimizu, and CM Albert. The spectrum of epidemiology underlying sudden cardiac death. *Circ Res*, 116(12):1887–1906, 2015.
- ML Weisfeldt, CM Sitlani, JP Ornato, T Rea, TP Aufderheide, D Davis, J Dreyer, P Hess, J Jui, and J Maloney. Survival after application of automatic external defibrillators before arrival of the emergency medical system: evaluation in the resuscitation outcomes consortium population of 21 million. *Journal of the American College of Cardiology*, 55(16):1713–1720, 2010.
- RJ Myerburg and MJ Junttila. Sudden cardiac death caused by coronary heart disease. *Circulation*, 125(8):1043–1052, 2012.
- X Jouven, W Bougouin, K Narayanan, and E Marijon. Sudden cardiac death and sports. *European Heart JOURNAL*, 38(4):232–234, 2017.
- A Malhotra and S Sharma. Outcomes of cardiac screening in adolescent soccer players. *The New England Journal of Medicine*, 379(21):2084, 2018.
- GD Perkins, IG Jacobs, VM Nadkarni, RA Berg, F Bhanji, D Biarent, LL Bossaert, SJ Brett, D Chamberlain, AR de Caen, CD Deakin, JC Finn, JT Grasner, MF Hazinski, T Iwami, RW Koster, SH Lim, HM Ma, BF McNally, PT Morley, LJ Morrison, KG Monsieurs, W Montgomery, G Nichol, K Okada, MEH Ong, AH Travers, and JP Nolan. Cardiac arrest and cardiopulmonary resuscitation outcome reports: update of the utstein resuscitation registry templates for out-of-hospital cardiac arrest. *Circulation*, 132(13):1286–1300, 2015.
- F Adnet and F Lapostolle. International ems systems: France. *Resuscitation*, 63(1):7–9, 2004.
- V Satopaa, J Albrecht, D Irwin, and B Raghavan. *Finding a 'Kneedle' in a Haystack: Detecting Knee Points in System Behavior*. IEEE, 2011.
- L Van der Maaten and G Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- MAE Haukilahti, L Holmström, J Vähätalo, T Kenttä, J Tikkanen, L Pakanen, ML Kortelainen, J Perkiömäki, H Huikuri, RJ Myerburg, and MJ Junttila. Sudden cardiac death in women. *Circulation*, 139(8):1012–1021, 2019.
- A Kumar, DM Avishay, CR Jones, JD Shaikh, R Kaur, M Aljadah, A Kichloo, N Shiwalkar, and S Keshavamurthy. Sudden cardiac death: epidemiology and pathogenesis and management. *Rev Cardiovasc Med*, 22(1):147–158, 2021.

- JP Empana, X Jouven, RN Lemaitre, N Sotoodehnia, T Rea, TE Raghunathan, G Simon, and DS Siscovick. Clinical depression and risk of out-of-hospital cardiac arrest. *Archives of Internal Medicine*, 166(2):195–200, 2006.
- CM Albert, CU Chae, KM Rexrode, JE Manson, and I Kawachi. Phobic anxiety and risk of coronary heart disease and sudden cardiac death among women. *Circulation*, 111(4):480–7, 2005.
- W Whang, Laura D K, I Kawachi, KM Rexrode, CH Kroenke, RJ Glynn, H Garan, and CM Albert. Depression and risk of sudden cardiac death and coronary heart disease in women: results from the nurses' health study. *Journal of the American College of Cardiology*, 53(11):950–958, 2009.
- P Weeke, A Jensen, F Folke, GH Gislason, JB Olesen, EL Fosbøl, M Wissenberg, FK Lippert, EF Christensen, SL Nielsen, E Holm, JK Kanters, HE Poulsen, L Køber, and C Torp-Pedersen. Antipsychotics and associated risk of out-of-hospital cardiac arrest. *Clinical Pharmacology and Therapeutics*, 96(4):490–497, 2014.
- TM. McMillan and GM Teasdale. Death rate is increased for at least 7 years after head injury: a prospective study. *Brain*, 130(10):2520–2527, 2007.
- Marten Van den Berg, Bruno Stricker, Guy Brusselle, and Lies Lahousse. Chronic obstructive pulmonary disease and sudden cardiac death: A systematic review. *Trends in Cardiovascular Medicine*, 26(7):606–613, 2016.
- ED Robin and N Lewiston. Unexpected and unexplained sudden death in young asthmatic subjects. *Chest*, 96(4):790–793, 1989.
- E Markatis, A Afthinos, E Antonakis, and IC Papanikolaou. Cardiac sarcoidosis: diagnosis and management. *Rev Cardiovasc Med*, 21(3):321–338, 2020.
- A Porta-Sanchez, C Gilbert, D Spears, E Amir, J Chan, K Nanthakumar, and P Thavendiranathan. Incidence and diagnosis and and management of qt prolongation induced by cancer therapies: A systematic review. *Journal of the American Heart Association*, 6(12):e007724, 2017.
- O Itzhaki Ben Zadok, I Nardi Agmon, V Neiman, A Eisen, G Golovchiner, T Bental, N Schamroth-Pravda, E Kadmon, G R Goldenberg, A Erez, R Kornowski, and A Barsheshet. Implantable cardioverter defibrillator for the primary prevention of sudden cardiac death among patients with cancer. *The American Journal of Cardiology*, 191:32–38, 2023.
- L Di Lullo, R Rivera, V Barbera, A Bellasi, M Cozzolino, D Russo, A De Pascalis, D Banerjee, F Floccari, and C Ronco. Sudden cardiac death and chronic kidney disease: From pathophysiology to treatment strategies. *International Journal of Cardiology*, 217:16–27, 2016.
- GA Mensah, AH Mokdad, ES Ford, KJ Greenlund, and JB Croft. State of disparities in cardiovascular health in the united states. *Circulation*, 111(10):1233–1241, 2005.
- A Rozanski, JA. Blumenthal, and J Kaplan. Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. *Circulation*, 99(16):2192–2217, 1999.

- K. G. M. Moons, D. G. Altman, J. B. Reitsma, J. P. A. Ioannidis, P. Macaskill, E. W. Steyerberg, A. J. Vickers, D. F. Ransohoff, and G. S. Collins. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): Explanation and elaboration. *Annals of Internal Medicine*, 162(1):W1 –73, 2015.
- S.M Lundberg, B Nair, M.S Vavilala, M Horibe, M.J Eisses, T Adams, D.E Liston, D.K Low, S.F Newman, J Kim, and S.I Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749 –760, 2018.
- H.C Thorsen-Meyer, A.B Nielsen, A.P Nielsen, B.S Kaas-Hansen, P Toft, J Schierbeck, T Strøm, Chmura P.J, M Heimann, L Dybdahl, L Spangsege, P Hulsen, K Belling, S Brunak, and A Perner. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health*, 2(4):e179 –e191, 2020.
- SL Hyland, M Faltys, M Hüser, X Lyu, T Gumbsch, C Esteban, C Bock, M Horn, M Moor, B Rieck, M Zimmermann, D Bodenham, K Borgwardt, G Rätsch, and TM Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med*, 26(3):364 –373, 2020.
- Ray-Bing Chen, Chi-Hsiang Chu, Shinsheng Yuan, and Ying Nian Wu. Bayesian sparse group selection. *J. Comput. Graph. Statist.*, 25(3):665 –683, 2016.
- Himel Mallick and Nengjun Yi. Bayesian group bridge for bi-level variable selection. *Comput. Statist. Data Anal.*, 110:115 –133, 2017.
- Mingxuan Cai, Mingwei Dai, Jingsi Ming, Heng Peng, Jin Liu, and Can Yang. Bivas: a scalable bayesian method for bi-level variable selection with applications. *J. Comput. Graph. Statist.*, 29(1):40 –52, 2020.
- Christian Schäfer and Nicolas Chopin. Sequential Monte Carlo on large binary sampling spaces. *Stat. Comput.*, 23(2):163 –184, 2013.
- Christian Schäfer. *Monte Carlo methods for sampling high-dimensional binary vectors*. PhD thesis, Université Paris Dauphine, 2012.
- Kong Melissa, Fonarow Gregg, Peterson Eric, Curtis Anne, Hernandez Adrian, Sanders Gillian, Thomas Kevin, Hayes David, and Al-Khatib Sana. Systematic review of the incidence of sudden cardiac death in the united states. *Journal of the American College of Cardiology*, 57(7):794–801, 2011.
- Nina Japundzic-Zigon, Olivera Sarenac, Maja Lozic, Marko Vasic, Tatjana Tasic, Dragana Bajic, Vladimir Kanjuh, and David Murphy. Sudden death: Neurogenic causes, prediction and prevention. *European Journal of Preventive Cardiology*, 25(1):29–39, 2018.

List of talks

- | | |
|---------------------------|--|
| October 16, 2020 | Journée Européenne de Sensibilisation à l'Arrêt Cardiaque 2020, Hôpital Européen Georges Pompidou, Paris |
| October 15, 2021 | Journée Européenne de Sensibilisation à l'Arrêt Cardiaque 2021, Hôpital Européen Georges Pompidou, Paris |
| May 9, 2022 | Groupe de travail Intelligence Artificielle et Rythmologie, Académie Nationale de Médecine, Paris |
| May 13, 2022 | CREST Research Day, ENSAE, Palaiseau |
| September 16, 2022 | Institut Hors Murs Sciences Cardiovasculaires, Université Paris-Cité, Paris |
| October 14, 2022 | Journée Européenne de Sensibilisation à l'Arrêt Cardiaque 2022, Hôpital Européen Georges Pompidou, Paris |



Titre : Prédiction de la Mort Subite de l'Adulte et Identification des Facteurs de Risque Associés grâce au Machine Learning

Mots clés : Mort Subite de l'Adulte, Apprentissage Statistique, Clustering, Prédiction Individualisée, Sélection de Variables

Résumé :

La mort subite de l'adulte est définie comme une mort inattendue sans cause extracardiaque évidente, survenant avec un effondrement rapide en présence d'un témoin, ou en l'absence de témoin dans l'heure après le début des symptômes. Son incidence est estimée à 350,000 personnes par an en Europe et 300,000 personnes aux Etats-Unis, ce qui représente 10 à 20% des décès dans les pays industrialisés. Malgré les progrès réalisés dans la prise en charge, le pronostic demeure extrêmement sombre. Moins de 10% des patients sortent vivants de l'hôpital après la survenue d'une mort subite. Les défibrillateurs automatiques implantables offrent une solution thérapeutique efficace chez les patients identifiés à haut risque de mort subite. Leur identification en population générale demeure donc un enjeu de santé publique majeur, avec des résultats jusqu'à présent décevants. Cette thèse propose des outils statistiques pour répondre à ce problème, et améliorer notre compréhension de la mort subite en population générale. Nous analysons les données du Centre d'Expertise de la Mort Subite et les bases médico-administratives de l'Assurance Maladie, pour développer trois travaux principaux. La première partie de la thèse vise à identifier de nouveaux sous-groupes de mort

subite pour améliorer les modèles actuels de stratification du risque, qui reposent essentiellement sur des variables cardiovasculaires. Nous utilisons des modèles d'analyse du langage naturel et de clustering pour construire une nouvelle représentation pertinente de l'historique médical des patients. La deuxième partie vise à construire un modèle de prédiction de la mort subite, capable de proposer un score de risque personnalisé et explicable pour chaque patient, et d'identifier avec précision les individus à très haut risque en population générale. Nous entraînons pour cela un algorithme de classification supervisée, combiné avec l'algorithme SHapley Additive exPlanations, pour analyser l'ensemble des consommations de soin survenues jusqu'à 5 ans avant l'événement. La dernière partie de la thèse vise à identifier le niveau optimal d'information à sélectionner dans des bases médico-administratives de grande dimension. Nous proposons un algorithme de sélection de variables bi-niveaux pour des modèles linéaires généralisés, permettant de distinguer les effets de groupe des effets individuels pour chaque variable. Cet algorithme repose sur une approche bayésienne et utilise une méthode de Monte Carlo séquentiel pour estimer la loi *a posteriori* de sélection des variables.

Title : Exploring Risk Factors and Prediction Models for Sudden Cardiac Death with Machine Learning

Keywords : Sudden Cardiac Death, Machine Learning, Clustering, Personalized prediction, Variable Selection

Abstract :

Sudden cardiac death (SCD) is defined as a sudden natural death presumed to be of cardiac cause, heralded by abrupt loss of consciousness in the presence of witness, or in the absence of witness occurring within an hour after the onset of symptoms. Despite progress in clinical profiling and interventions, it remains a major public health problem, accounting for 10 to 20% of deaths in industrialised countries, with survival after SCD below 10%. The annual incidence is estimated 350,000 in Europe, and 300,000 in the United States. Efficient treatments for SCD management are available. One of the most effective options is the use of implantable cardioverter defibrillators (ICD). However, identifying the best candidates for ICD implantation remains a difficult challenge, with disappointing results so far. This thesis aims to address this problem, and to provide a better understanding of SCD in the general population, using statistical modeling. We analyze data from the Paris Sudden Death Expertise Center and the French National Healthcare System Database to develop three main works. The first part of the thesis aims to identify

new subgroups of SCD to improve current stratification guidelines, which are mainly based on cardiovascular variables. To this end, we use natural language processing methods and clustering analysis to build a meaningful representation of medical history of patients. The second part aims to build a prediction model of SCD in order to propose a personalized and explainable risk score for each patient, and accurately identify very-high risk subjects in the general population. To this end, we train a supervised classification algorithm, combined with the SHapley Additive exPlanation method, to analyze all medical events that occurred up to 5 years prior to the event. The last part of the thesis aims to identify the most relevant information to select in large medical history of patients. We propose a bi-level variable selection algorithm for generalized linear models, in order to identify both individual and group effects from predictors. Our algorithm is based on a Bayesian approach and uses a Sequential Monte Carlo method to estimate the posterior distribution of variables inclusion.

