



HAL
open science

Statistical Learning from Multimodal Genetic and Neuroimaging data for prediction of Alzheimer's Disease

Pascal Lu

► **To cite this version:**

Pascal Lu. Statistical Learning from Multimodal Genetic and Neuroimaging data for prediction of Alzheimer's Disease. Statistics [math.ST]. Sorbonne Université, 2019. English. NNT: . tel-02433613v1

HAL Id: tel-02433613

<https://inria.hal.science/tel-02433613v1>

Submitted on 9 Jan 2020 (v1), last revised 21 Sep 2021 (v2)

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.

SORBONNE UNIVERSITÉ
ÉCOLE DOCTORALE D'INFORMATIQUE, DE TÉLÉCOMMUNICATION ET D'ÉLECTRONIQUE
(ED130)
ARAMIS LAB À L'INSTITUT DU CERVEAU ET DE LA MOELLE ÉPINIÈRE (UMR 7225)

Thèse de doctorat de Sorbonne Université

Spécialité: Informatique

Présentée par

Pascal LU

Pour obtenir de le grade de

Docteur de l'Université Sorbonne Université

Soutenue le 26 novembre 2019

Statistical Learning from Multimodal Genetic and Neuroimaging
data for prediction of Alzheimer's Disease

Apprentissage Statistique à partir de Données Multimodales de
Génétique et de Neuroimagerie pour la prédiction de la Maladie
d'Alzheimer

Rapporteurs

M. Christophe AMBROISE
Mme Agathe GUILLOUX

Professeur
Professeur

Université d'Évry Val-d'Essonne
Université d'Évry Val-d'Essonne

Examineurs

M. Jean-Daniel ZUCKER
M. Theodoros EVGENIOU

Directeur de recherche
Professeur

IRD, Sorbonne Université
INSEAD

Directeur de thèse

M. Olivier COLLIOT

Directeur de recherche

CNRS, ICM

Abstract

Alzheimer's Disease (AD) is nowadays the main cause of dementia in the world. It provokes memory and behavioural troubles in elderly people. The early diagnosis of Alzheimer's Disease is an active topic of research. Three different types of data play a major role when it comes to its diagnosis: clinical tests, neuroimaging and genetics. The two first data bring informations concerning the patient's current state. On the contrary, genetic data help to identify whether a patient could develop AD in the future. Furthermore, during the past decade, researchers have created longitudinal dataset on A and important advances for processing and analyse of complex and high-dimensional data have been made.

The first contribution of this thesis will be to study how to combine different modalities in order to increase their predictive power in the context of classification. We will focus on hierarchical models that capture potential interactions between modalities. Moreover, we will adequately modelled the structure of each modality (genomic structure, spatial structure for brain images), through the use of adapted penalties such as the ridge penalty for images and the group lasso penalty for genetic data.

The second contribution of this thesis will be to explore models for predict the conversion date to Alzheimer's Disease for mild cognitive impairment subjects. Such problematic has been enhanced by the TADPOLE challenge. We will use the framework provided by survival analysis. Starting from basic models such as the Cox proportional hazard model, the additive Aalen model, and the log-logistic model, we will develop other survival models for combining different modalities, such as a multilevel log-logistic model or a multilevel Cox model.

Résumé

De nos jours, la maladie d'Alzheimer est la principale cause de démence. Elle provoque des troubles de mémoires et de comportements chez les personnes âgées. Le diagnostic précoce de la maladie d'Alzheimer est un sujet actif de recherche. Trois différents types de données jouent un rôle particulier dans le diagnostic de la maladie d'Alzheimer: les tests cliniques, les données de neuroimagerie et les données génétiques. Les deux premières modalités apportent de l'information concernant l'état actuel du patient. En revanche, les données génétiques permettent d'identifier si un patient est à risque et pourrait développer la maladie d'Alzheimer dans le futur. Par ailleurs, durant la dernière décennie, les chercheurs ont créé des bases de données longitudinales sur la maladie d'Alzheimer et d'importantes recherches ont été réalisées pour le traitement et l'analyse de données complexes en grande dimension.

La première contribution de cette thèse sera d'étudier comment combiner différentes modalités dans le but d'améliorer leur pouvoir prédictif dans le contexte de la classification. Nous explorons les modèles multiniveaux permettant de capturer les potentielles interactions entre modalités. Par ailleurs, nous modéliserons la structure de chaque modalité (structure génétique, structure spatiale du cerveau) à travers l'utilisation de pénalités adaptées comme la pénalité ridge pour les images, ou la pénalité group lasso pour les données génétiques.

La deuxième contribution de thèse sera d'explorer les modèles permettant de prédire la date de conversion à la maladie d'Alzheimer pour les patients atteints de troubles cognitifs légers. De telles problématiques ont été mises en valeur à travers de challenge, comme TADPOLE. Nous utiliserons principalement le cadre défini par les modèles de survie. Partant de modèles classiques, comme le modèle d'hasard proportionnel de Cox, du modèle additif d'Aalen, et du modèle log-logistique, nous allons développer d'autres modèles de survie pour la combinaison de modalités, à travers un modèle log-logistique multiniveau ou un modèle de Cox multiniveau.

Remerciements

À mon directeur de thèse

Je tiens tout d'abord à remercier Olivier, mon directeur de thèse, pour m'avoir introduit au monde des données médicales, pour tout sa patience, ses remarques constructives tout au long de la thèse, et surtout sa grande disponibilité.

Aux membres du jury

Je souhaite remercier Theos que j'ai rencontré pendant ma première année de thèse. Je retiendrais de nos échanges qu'il n'y a pas d'intérêt de faire des modèles complexes lorsque les modèles les plus simples fonctionnent, et qu'il faut se poser la question si une équation mathématique a un sens physique.

Je tiens aussi à remercier Mme Agathe Guilloux et M. Christophe Ambroise, qui ont déjà eu l'opportunité d'être jury du comité de suivi et donc de suivre mes travaux à des moments très particuliers de la thèse, d'avoir aussi émis des remarques pertinentes et permis de rectifier le tir lorsque c'était nécessaire.

Enfin, je remercie M. Jean-Daniel Zucker d'avoir accepté de faire parti du jury de thèse. Je suis honoré de pouvoir présenter mes travaux devant vous et de bénéficier de votre expertise.

À l'équipe ARAMIS

L'équipe ARAMIS est une équipe sympathique et soudée avec qui j'ai passé trois bonnes années. En particulier, j'ai pris beaucoup de plaisir à intégrer le club des Na, avec SimoNa toujours là à me surveiller, TiziaNa pour sa bonne humeur, Giulia BassignaNa, CataliNa, SabriNa, JuliaNa, EliNa et Federica(Na). Je voudrais remercier aussi mes voisins de palliers Jorge, Ludovic, Jérémy, Hao et Manon, pour toutes les astuces geek Alexandre R. et Benoît, mais aussi Raphaël, Maxime, Igor, Alexandre B., avec lesquels j'ai gardé un bon souvenir de la soirée du gala de MICCAI 2017 à Québec, Wen, Marie-Constance, Fanny, Alexis, Arnaud(s), Adam, Baptiste, Vincent, Clémentine, Adam, Quentin, Paul, Dario, Emmanuelle, Ninon, Fabrizio, Benjamin, Stanley, Stéphane; Jean-Baptiste et Pietro les premières personnes que j'ai connu à ARAMIS. J'ai aussi une pensée pour Anne Bertrand qui nous a quitté en 2018.

À mes parents

Pour tous leurs encouragements pendant toutes ces années.

Contents

Abstract	3
Résumé	7
Acknowledgements	7
Contents	9
List of Abbreviations	13
Introduction	15
I State of the art	19
1 Alzheimer’s Disease	21
1.1 Introduction	21
1.2 Alzheimer’s Disease	21
1.2.1 Alzheimer’s Disease and the human brain	22
1.2.2 Diagnostic of Alzheimer’s Disease	22
1.3 Modalities involved to study Alzheimer’s Disease	23
1.3.1 Clinical and cognitive tests	23
1.3.2 Biological biomarkers	25
1.3.2.1 Anatomical MRI	25
1.3.2.2 PET scans	25
1.3.2.3 Lumbar puncture	26
1.3.3 Genetic data	26
1.4 Features	27
1.4.1 Neuroimaging data	27
1.4.2 Genetic data	27
1.5 The Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset	30
1.5.1 Presentation	30
1.5.2 Descriptive statistics of ADNI1 dataset	30
1.5.3 MCI patients follow-up	32
1.6 Conclusion	33
2 Statistical and machine learning approaches for Imaging Genetics	35
2.1 Introduction	35

2.2	Univariate and multivariate analyses between genetic and neuroimaging data	36
2.2.1	Univariate association between genotype and one phenotype trait	37
2.2.2	Multivariate analyses	39
2.2.2.1	Partial Least Squares	39
2.2.2.2	Linear regression between SNPs and neuroimaging data	40
2.2.2.3	Generative models	40
2.2.2.4	Pathways-based regression modelling	43
2.2.2.5	Epistasis effects and Random Forests on Distance Matrices	44
2.3	Combinaison of genetic and neuroimaging data for disease diagnosis	45
2.3.1	Dealing with high dimensional data	47
2.3.2	Multiple Kernel Learning (MKL)	48
2.3.3	Structured sparse regularisation	50
2.4	Conclusion	51
3	Survival analysis: from theory to application for Alzheimer’s Disease	53
3.1	Introduction	53
3.2	Background	53
3.2.1	Assumptions	53
3.2.2	Censoring and truncation	54
3.2.3	Survival and hazard function	55
3.2.4	Kaplan-Meier and Nelson-Aalen estimators	56
3.2.5	Adding the covariates for individual predictions	57
3.2.5.1	Cox proportional hazard model	57
3.2.5.2	Exponential model	58
3.2.5.3	Log-logistic model	59
3.2.6	Model fitting	59
3.2.6.1	Parametric models	59
3.2.6.2	Semi-parametric models	60
3.2.7	Hypotheses behind the Cox proportional hazard model	61
3.3	Measuring the predictive value of a survival model	62
3.3.1	Median survival time	62
3.3.2	Concordance-index (C-index)	62
3.3.3	Cumulative AUC(t)	63
3.3.4	Kullback-Leibler divergence and Brier score	63
3.4	Combinaison of multimodal data using the Cox PH framework	64
3.5	Illness-death models	66
3.6	Conclusion	67
II	Contributions	69
4	Multilevel Modeling with Structured Penalties for Classification from Imaging Genetics Data	71
4.1	Introduction	71

4.2	State of the art	72
4.3	Model set-up	73
4.3.1	Multilevel Logistic Regression with Structured Penalties	73
4.3.2	Minimization of $S(\mathbf{W}, \beta_T, \beta_T, \beta_0)$	74
4.3.3	Probabilistic formulation	76
4.4	Experimental results	77
4.4.1	Dataset	77
4.4.2	Results	77
4.4.3	Parameters	78
4.5	Conclusion	80
5	Contributions to the TADPOLE 2017-2022 challenge	83
5.1	Introduction	83
5.2	Project details	83
5.2.1	Overview	83
5.2.2	Datasets	84
5.2.3	Metrics	85
5.2.3.1	Multiclass Area Under the ROC Curve	85
5.2.3.2	Balanced Classification Accuracy	86
5.3	Contributions	87
5.3.1	Disease progression paths	87
5.3.2	Estimation of the conversion date to AD for MCI patients	88
5.4	Results and evaluation on D4	92
5.5	Conclusion	96
6	Multimodal survival models for predicting the conversion to Alzheimer's Disease	97
6.1	Introduction	97
6.2	Comparison between classification and survival analysis	97
6.2.1	Models	98
6.2.2	Effect of covariates	99
6.2.3	Assessment of the predictive value	100
6.2.4	Conclusion	101
6.3	A multilevel log-logistic model	102
6.3.1	Motivation	102
6.3.2	Parametrisation	104
6.3.3	Model fitting	106
6.3.4	Experiments and results	108
6.3.5	Discussion	111
6.4	Conclusion	112
7	Multilevel Cox Proportional hazard Model with Structured Penalties for Imaging Genetics data	113
7.1	Introduction	113
7.2	State of the art	114

7.3	Methods	115
7.3.1	Model set-up	115
7.3.2	Optimization	116
7.4	Experiments and results	117
7.4.1	Dataset	117
7.4.2	Evaluation	117
7.4.3	Effect of cross-product covariates on conversion	119
7.5	Conclusion	120
	Conclusion	123
	Appendix	127
	Scientific production	127
	List of Figures	129
	List of Tables	133
	Bibliography	135

List of Abbreviations

Aβ	Amyloid beta protein
AD	Alzheimer's disease
ADASCog	Alzheimer's disease assessment scale cognitive sub-scale
ADNI	Alzheimers Disease Neuroimaging Initiative
APOE	Apolipoprotein E
AUC	Area under the receiver operating characteristic curve
CN	Cognitively normal
CSF	Cerebrospinal fluid
CV	Cross-validation
DNA	Deoxyribonucleic acid
FDG	¹⁸ F 2-fluoro-2-deoxy-D-glucose
GWAS	Genome Wide Association Study
KM	Kaplan-Meier estimator
LogMem	Logical Memory test
MCI	Mild cognitive impairment
MMSE	Mini-mental state examination
MKL	Multiple Kernel Learning
MRI	Magnetic resonance imaging
NA	Nelson-Aalen estimator
PET	Positron emission tomography
PH	Proportional hazard
pMCI	Progressive mild cognitive impairment
RAVLT	Rey auditory verbal learning test
ROI	Region of interest
sMCI	Stable mild cognitive impairment
SNP	Single Nucleotide Polymorphisms
SVM	Support vector machine
T1	T1-weighted magnetic resonance imaging

Introduction

Personalised medicine in the context of Alzheimer's Disease

Personalised medicine aims at tailoring medical decisions, prevention and therapies to individual patients, based on their predicted risk of disease, evolution and response. In this approach, patients are characterised using rich multimodal measurements (genomics, medical imaging, biomarkers. . .). A central challenge is then to develop predictive models from these measurements. To that end, it is necessary to design new statistical learning approaches that can fully exploit the different types of data.

Neurodegenerative diseases, such as Alzheimer's disease (AD), are complex multifactorial diseases that represent major public health issues. In the context of these brain disorders, three types of data play a major role: genetics, neuroimaging and clinical tests. First, genetics allow identifying factors that modulate the risk of a given disease, its evolution and response to treatment. It involves measurement of increasing complexity, from series of Single Nucleotide Polymorphisms (SNPs) provided by microarrays to high-throughput sequencing approaches such as whole-exome or even whole-genome sequencing. Second, neuroimaging allows measuring, in the living patient, different types of anatomical and functional alterations, using a variety of imaging modalities: anatomical, functional and diffusion magnetic resonance imaging (MRI) and positron emission tomography (PET). Third, through clinical tests, the neurologist will assess the cognitive functions of the patient, such as memory, attention or executive functions. Example of such tests are the Mini-Mental State Exam (MMSE) and the ADAS-Cog test. Overall, both neuroimaging and clinical tests provide an accurate picture of the subject's state.

Combining multimodal genetic and imaging data

Genetic and imaging technologies have witnessed considerable development during the past 15 years. In the meantime, important advances have been made for processing and statistical analysis of these complex data. However, machine learning approaches that can adequately integrate neuroimaging and genetic data are currently lacking. The development of such approaches is particularly timely because massive datasets of patients with both imaging and genetic data are now available. One can cite for instance the Alzheimer's Disease Neuroimaging Initiative – ADNI (<http://adni.loni.usc.edu>), in the context of AD.

Methodological developments are challenging because of:

- (i) the high dimensionality of both types of data (around 10^5 or 10^6);

- (ii) the complex multivariate interactions between variables. How can we combine these modalities in order to get a higher predictive power?

A large part of the literature on combination of imaging and genetic data focused on association studies, i.e. studying the relationships between genetic data in a univariate or multivariate manner. In such approaches, genetic is the time-independent variable and imaging the time-dependent variable. Univariate approaches follow the paradigm of so-called Genome-Wide Association Studies (GWAS) and look for statistical associations between each individual SNP and each neuroimaging variable (for instance, average within a brain region or value at a given voxel). However, SNPs often have weak effects when taken separately. Similarly, anatomical and functional brain phenotypes are best described using combinations of imaging variables. For that reason, researchers have proposed various types of multivariate approaches based for instance on penalized multiple regression, partial least squares or multivariate generative models.

A different question is to design statistical models that can predict the patient state (i.e. current or future diagnosis) from multimodal genetic and imaging data. Even though it is clinically relevant, this question has been the subject of less work in the literature. One can cite approaches based on regularized and sparse classification techniques as well as multiple kernel learning. However, all these approaches put genetic and imaging data at the same level while they could play different roles in the prediction. Indeed, imaging (or clinical) data provide a snapshot of the patient state at a given time while genetics can modulate the evolution of the patient.

Statistical learning for prediction of Alzheimer's disease

In the past decade, there has been intense research on the development of statistical learning methods for predicting AD. Initial work has focused on automatic classification of patients with AD and control subjects, in order to assist diagnosis. Such initial research has focused on a single modality, using neuroimaging. A more challenging and useful aim is to predict the future state of patients. In particular, many papers have been devoted to predicting the future occurrence of AD in patients with mild symptoms, a state called Mild Cognitive Impairment (MCI). Again, most of the approaches have used neuroimaging data as input, more rarely clinical tests and even more rarely multimodal data such as imaging and genetic or clinical and genetic data. Moreover, they mostly tackled this question through classification approaches: one sets a fixed time window (for instance 3 years) and aims to discriminate between MCI patients who will progress to AD within this time window and those who will not. However, another clinically relevant question is to determine the date at which the patient will develop AD. For instance, the care of the patient and the information of the family (for instance for preparing institutionalization) can be quite different depending on whether dementia is expected to develop within one year or five years.

Objectives of this PhD

The main objective of this PhD is to propose a methodological approaches to combine different and genetic, neuroimaging and clinical modalities in order to predict the evolution of patients to Alzheimer's disease.

In the first part of this PhD, we aimed at predicting the current or future diagnosis of Alzheimer's Disease using classification methods. In the state of the art, most models combine different modalities using an additive framework. Is it the best way to combine genetics and neuroimaging data, although they do not provide the same level of information? Instead, we aimed to propose a multilevel framework that can capture interactions between variables.

In a second part, we aimed at predicting the conversion date to Alzheimer's Disease. Among the subjects who have Mild Cognitive Impairment (MCI), who are the patients who will be affected by AD? if the patient converts, what is the conversion date? Our developments were performed within the framework of survival analysis, which is well suited to that purpose. Survival models provide a regression framework which directly estimates the conversion date using only one time-point. Using cross-sectional data, instead of longitudinal data, is much more realistic from a medical point of view, although the results in predictions will be less accurate. We applied different survival models to our data, such as the Cox proportional hazard model, the log-logistic model, and the Aalen additive models. We also proposed modified versions of these models, using a multilevel framework.

The remainder of this document is organised as follows.

The first part is devoted to the background and state-of-the-art. Chapter 1 provides background information on AD and multimodal imaging, genetic and clinical data. It also introduces the ADNI database, a publicly available multicentric dataset that we will use for our experiments. Chapter 2 reviews existing statistical learning approaches for integration of imaging and genetic data. Chapter 3 provides an overview of survival analysis and existing applications to multimodal data and AD.

The second part of the document contains the contributions. Chapter 4 focuses on classification techniques for assisting diagnosis of AD and predicting the future occurrence of AD in patients with MCI. In particular, we propose a multilevel framework for combining imaging and genetic data. Chapter 5 presents our contribution to the TADPOLE international challenge on prediction of AD. We used a standard survival model and ranked high on one of the challenge outcomes. Chapter 6 is focused on multimodal survival models for prediction of AD. We first compare standard survival models and classification approaches. We then propose and evaluate an original multilevel log-logistic survival model. In Chapter 7, we introduce a multilevel Cox Proportional Hazard model which extends the approach proposed in Chapter 4 to the case of survival analysis.

Part I

State of the art

Chapter 1

Alzheimer's Disease

1.1 Introduction

In this chapter, we provide a concise description of Alzheimer's Disease (AD), from its known causes to its diagnosis. We will also describe the modalities that are involved in the study of AD, such as neuropsychological tests, neuroimaging data, lumbar puncture, and genetic data; and how these modalities are preprocessed for statistical studies. This chapter provides a short description of the ADNI dataset, that will be used throughout this work.

1.2 Alzheimer's Disease

Alzheimer's Disease (AD) is a type of dementia that provokes memory and behavioural troubles. Symptoms usually start slowly and get worst with time.

It is currently the main cause of dementia in the world, and has impacted many ageing populations in developed countries. Nowadays, it is one of the top priority in research, focus on understanding and treating AD. In 2015, estimations show that 46 millions of people were affected by AD, and studies predict that this number will increase in the coming years. Although one of the major risk of developing AD is age, and that AD usually affects elderly people, AD is not a normal process of ageing. Most people develop AD after 65 years old, but there are some patients that start to develop the disease at 45 years old.

AD is an evolutive disease whose symptoms get worst with time. Mechanisms of the disease start many years before the first clinical symptoms become noticeable. The preclinical stage starts with pathological changes, but without any visible symptoms [Dubois et al., 2016]. Then, the patient starts to have mild cognitive deficits, and memory troubles; this stage is named Mild Cognitive Impairment (MCI) [Dubois and Albert, 2004]. At an advanced stage, many functions such as memory, language, motor functions are affected, and the patient becomes dependant; the patient has dementia [McKhann et al., 1984, McKhann et al., 2011].

Nowadays, there is no treatment for AD, but treatments for symptoms of AD are available. These current treatments do not slow the disease progression, but can temporarily reduce the worsening of the symptoms and improve the patient's life quality. Finding the

best cure for AD or a cure that can slow its development is currently a top priority in the research community.

1.2.1 Alzheimer's Disease and the human brain

The human brain is made up of 100 billion nerve cells, called neurones. Each neurone is connected to several other neurones, and form all together a communication network. A group of neurones have a defined role, for instance there are groups for memory and learning, and others for interactions with the environment (vision, feeling, hearing). Neurones process and store the information, need energy, and communicate with each other.

Alzheimer's Disease prevents neurones from functioning normally. The main cause behind Alzheimer's Disease is not really known; what it is known is that some cells start by having malfunctions, and these malfunctions cause troubles and may affect other neurones. The damage spreads to other cells, and progressively, neurones are unable to work properly and eventually die. These damage are irreversible.

The causes behind the degradation and death of neurones are associated with two types of lesions throughout the cerebral cortex: amyloid plaques (found between neurones) and neurofibrillary degeneration (found inside neurones). They are both clumps of proteins that form during the normal ageing process. However, in AD, these proteins accumulate in a much larger quantity [Duyckaerts et al., 2009].

Amyloid plaques are small, dense deposits of a the β -amyloid protein. This one gradually agglutinates to form the plates. When it reaches high plaque levels, the β -amyloid protein becomes toxic to these neurones.

Neurofibrillary degeneration is due to the τ -protein that becomes abnormal inside the neurones. Neurones have a transport system that connects the cell body to the end of the axon, using filaments called microtubules. Nutrients and other essential materials travel along these microtubules. These microtubules are parallel thanks to the τ -protein. In patients affected by AD, these τ -protein do not work properly and consequently does not keep parallel the microtubules. The nerve located at the end of the axon will be the first to degenerate due to this lack of nutrients. Then, communication with neighbouring neurones will be diminished. At the end, the entire neurone degenerates.

It is the destruction and death of nerve cells that cause memory loss, personality disorders, difficulty performing daily tasks and other symptoms of Alzheimer's disease.

The pathological pathway of Alzheimer's Disease is described in figure 1.2.

1.2.2 Diagnostic of Alzheimer's Disease

The diagnosis of Alzheimer's Disease usually takes time. There are various elements that contribute to establishing the diagnosis of Alzheimer's disease [Dubois et al., 2007, Dubois et al., 2014]. The first element is that the patient and his entourage start by complaining of forgetfulness. At this moment, the patient will perform a consultation in a hospital. Results of clinical tests and neuropsychological assessment are critical to assess the diagnosis. Some clinical tests are specific to cognitive disorders in AD.

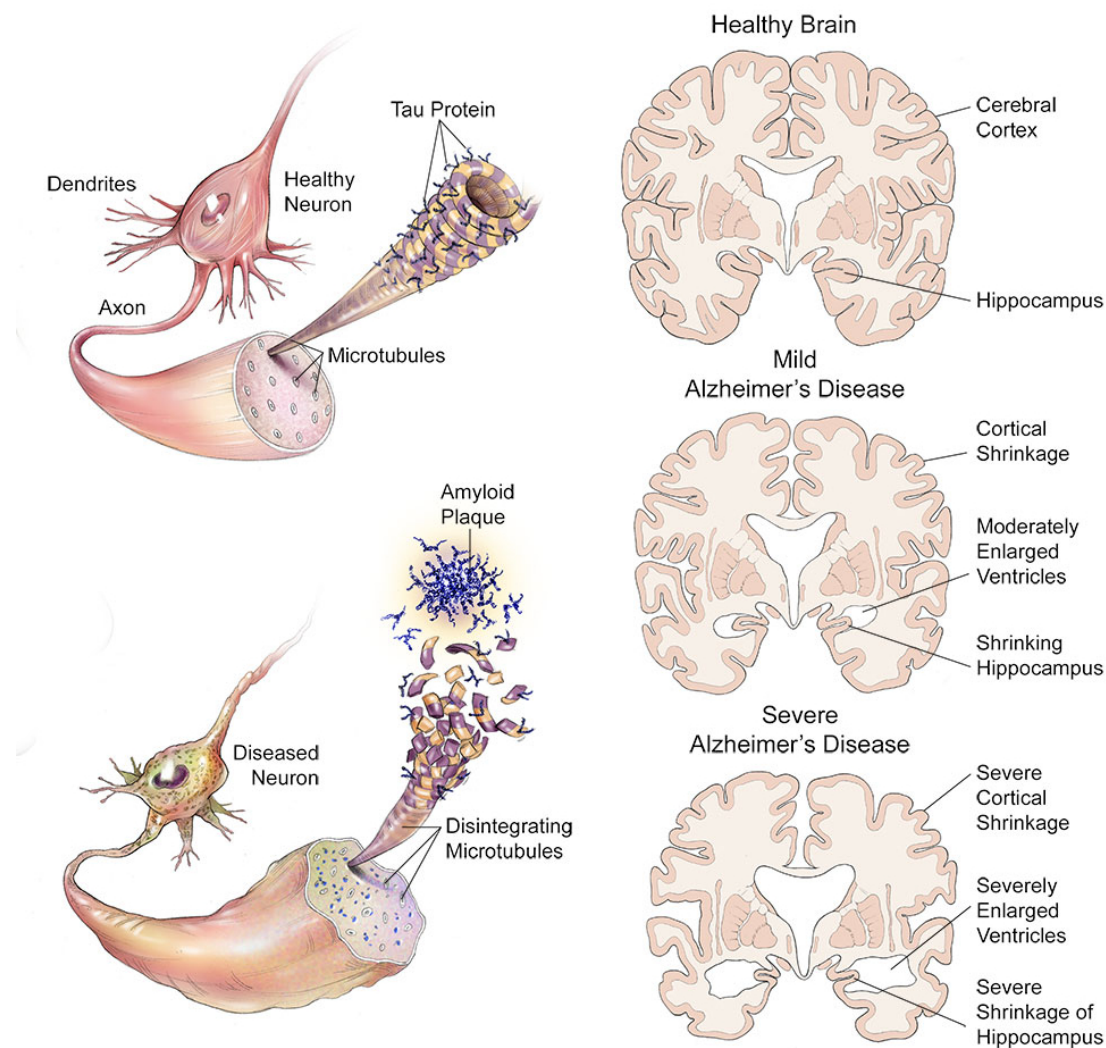


FIGURE 1.1: Progression of Alzheimer's Disease (reproduced from <https://www.brightfocus.org/alzheimers-disease/infographic/progression-alzheimers-disease>)

To complete the diagnosis, several analysis are made. The analysis of magnetic resonance imaging (MRI) for anatomical and functional structure of the brain and Positron emission tomography scan (PET scan) for neuronal cell metabolism are performed to measure the atrophy *in vivo*. The biological markers will support the neurologists' hypothesis [Hampel et al., 2014].

1.3 Modalities involved to study Alzheimer's Disease

1.3.1 Clinical and cognitive tests

The clinical and cognitive tests determine the patient's cognitive disorders through a series of questions. These tests evaluate the memory and other cognitive functions such as orientation in space and time, language, comprehension, attention and reasoning. These comprehensive tests distinguish between patients with Alzheimer's Disease at a very

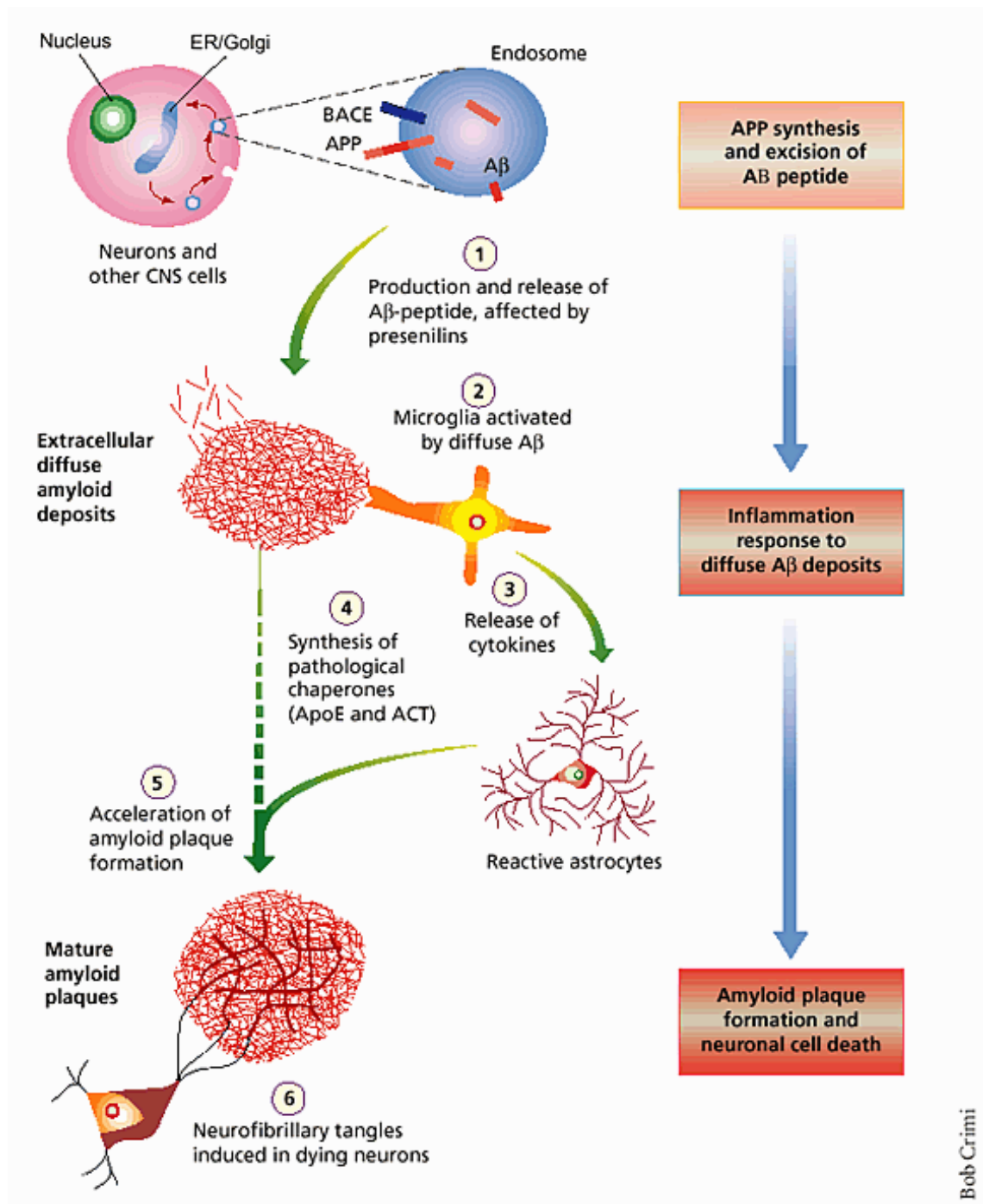


FIGURE 1.2: Simplified biological pathway leading to Alzheimer's Disease

early stage, healthy people and patients with dementia. The tests are adapted to the patient depending of his education level and its stage of evolution of the disease, and are performed by a neuropsychologist.

Examples of such tests are the ADAS-Cog (Alzheimer's Disease Assessment Scale-Cognitive Subscale test), RAVLT (Rey Auditory Verbal Learning Test), FAQ (Functional Assessment Questionnaire) and MMSE (Mini-Mental State Examination).

1.3.2 Biological biomarkers

On the contrary of clinical tests that aim at finding the existence of cognitive function disorders, biological biomarkers provide specific characteristic of the disease *in vivo*.

1.3.2.1 Anatomical MRI

Magnetic resonance imaging (MRI) can detect cortical atrophy and especially atrophy of brain regions. As the progression of atrophy for AD is well established, the atrophy of the hippocampus (the brain structure involved in memory), the medial temporal lobe [Scheltens et al., 1992], then the temporal neocortex, associative parietal areas and frontal regions are strongly correlated with AD.

MRI uses scanner with high spatial resolution and and show clearly the different tissue types. It also has the advantage to provide an accurate picture of the patient's state and is not invasive and expensive to acquire. That is why MRI are acquired in the clinical examination. Based on whole-brain data, AD vs CN classification is usually highly accurate.

Finally, it makes possible to eliminate other causes of dementia, such as the presence of vascular lesions, a hepatoma or a brain tumour.

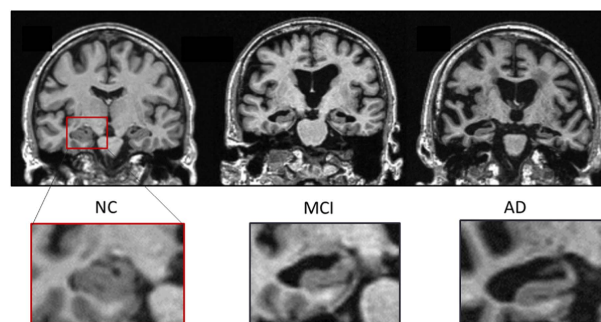


FIGURE 1.3: Coronal slides of the T1-weighted MRI for normal brain, MCI brain and AD brain. The bounding box represents the hippocampus region (reproduced from [Ahmed et al., 2017]).

1.3.2.2 PET scans

In the case where an MRI is contraindicated (for instance, for patients with a pacemaker), a PET scan may be prescribed.

PET scans (positron emission tomography) is a functional imaging technique and provides a representation of a given metabolic process through the detection of a positron-emitting isotope that is bound to a biologically active molecule. Depending on the molecule and the metabolic process, a specific phenomenon will be observed.

PET scans allows to visualise brain lesions characteristic of the disease, including amyloid plaques. An example of PET scans of the β -amyloid charge is shown on figure 1.4: on the left, the healthy subject, while on right a patient with AD. The red color corresponds to the highest concentration of β -amyloid protein, while the blue and green

colors correspond to little or no β -amyloid protein. The amount of red areas is much greater in Alzheimer's patients than in healthy men.

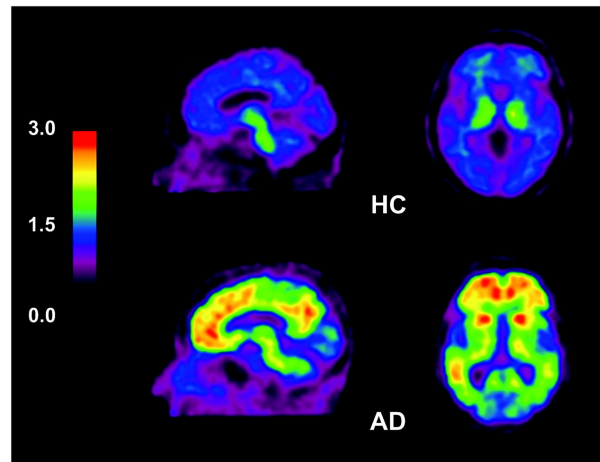


FIGURE 1.4: PIB-PET scan for Normal brain versus AD brain (reproduced from [Ng et al., 2007])

However, PET scans are more expensive and are invasive methods, as they require the injection of a radioactive tracer. However, PET provides complementary information about the disease than MRI do not, and therefore, it remains the second most-widely used modality for the diagnosis of AD. Nowadays, it is possible to pass both exams at the same time, as there are devices that merge the MRI and PET technology. The first device acquired in France is at the Brain&Spine Institute (*Institut du Cerveau et de la Moëlle Épinière*).

1.3.2.3 Lumbar puncture

A lumbar puncture can determine in the cerebrospinal fluid – fluid surrounding the brain and spinal cord –, several biological markers specific for Alzheimer's disease (in particular the τ - and τ -phosphorylated proteins, β -amyloid peptide). There are abnormalities in the concentration of these molecules from the beginning of the disease. The amyloid peptide $A\beta$ -24-1 is lowered, while the concentration of τ -protein is 2 or 3 times higher than normal. This examination constitutes an important contribution to the diagnostic hypothesis, is not expensive but the patient must be hospitalised for a least one day.

1.3.3 Genetic data

Genetic researchers has shown that some genes increase the likelihood of developing AD. It does not mean that the patient who has the specific allele will develop the disease for sure. It also have been shown that genes playing a role in it are not the same depending of the subject age, and in particular the genetic influence will be much stronger if AD develops at 45 years old than at 65 years old.

The hereditary forms represent less than 1% of patients having AD. Usually, for patients who develop AD before 65 years old, the disease is caused by a gene mutation. So far, four genes have been identified: the PSEN1 gene on chromosome 14 (the most frequent

case) [Janssen et al., 2000], the APP (Amyloid Precursor Protein) gene on chromosome 21 (second incriminated in frequency), the PSEN2 gene on chromosome 1 or the gene called SORL1 on chromosome 1. The genetic mutations provoke an increase in amyloid peptide production.

The sporadic forms are more frequent, and concern patients who develop AD after 65 years old. The part of genetics is not predominant in the occurrence of the disease and other environmental factors come into play. AD can be considered as a multifactorial disease. However, there are some genes of predisposition, such as the APOE4 gene located on chromosome 19 and involved in neuronal repair mechanisms.

Being a carrier of the allele ϵ_4 of APOE4 (Apolipoprotein E) gene increases the risk of developing Alzheimer's Disease by 11 times [Mahley et al., 2006]. However, there are many people having this allele who will never declare the disease. In 2013, the largest international study ever conducted on Alzheimer's disease (*International genomics of Alzheimer project*), identified eleven new regions of the genome involved in the occurrence of this neurodegenerative disease and perhaps thirteen others, in validation course [Lambert et al., 2013].

Genetic data are collected using a DNA chip. A DNA chip contains millions of probes to which the genetic variants will hybridise and be detected by fluorescence.

1.4 Features

1.4.1 Neuroimaging data

Different types of features can be extracted from MRI data. The two most common are voxel-based features (a measurement is taken at each voxel) and regional features (there is one measurement for each macroscopic region of the brain). We will work with regional features.

To extract such features, acquired MRI data need to be registered, normalised and corrected. Images from all subjects must be put in the same space, so that voxel-wise correspondence can be realised between subjects. Then an atlas is applied to parcelate the brain into different regions. In our work, we used features extracted with the FreeSurfer software. The regional features represent the average cortical thickness in each region of the cortex and the volume of each subcortical region.

1.4.2 Genetic data

DNA, genes, chromosomes The nucleus of a cell contains all genetic information, under the form of chromosomes. The genetic information is spread in 23 pairs of chromosomes. For each pair, there is a paternal chromosome and a maternal chromosome. Therefore, for the same pair, the two chromosomes will not be identical. The 22 first pairs are the autosomes, whereas the 23rd determined the person's sex: they are the X and Y chromosomes (a woman would have two X chromosomes, while a man would have one X chromosome and one Y chromosome).

Each chromosome contains DNA – *deoxyribonucleic acid* –, that carries the genes (approximately 25,000 genes in the human genome). The DNA carries the genetic code that determines the characteristics of a human, has a double helices complementary structure. The DNA is a sequence of nucleotides: adenine, thymine, guanine and cytosine (also known as A, T, G, C).

A gene is a piece of DNA that encodes for a specific single protein. It is therefore a small portion of a chromosome. Since we have two chromosomes for the same pair, each gene is present twice in the same celled. The two copies of the same genes can be identical or different (since they do not come from the same ancestor), but encode for the same protein.

The genes synthesise proteins; and each protein has a specific role. For instance, actin is a protein that help muscle contraction, while haemoglobin is the protein used to carry oxygen in the blood. There are also genes that encode for the phenotype, such as the colour of the eyes.

A genetic anomaly (mutation or chromosomal anomaly) can disrupt the production of proteins. It will give bad informations that could result in no production, excessive or abnormal production of proteins. At the end, the protein will not play its original role. A genetic abnormality does not always result in a disease, some errors can have no consequences. Examples of genetic mutations are chromosomal inversions (the orientation of the chromosomal segment is inverted), translocations, interstitial deletions. . . A genetic anomaly can happen accidentally during the production of gametes or after fertilisation, or can be inherited.

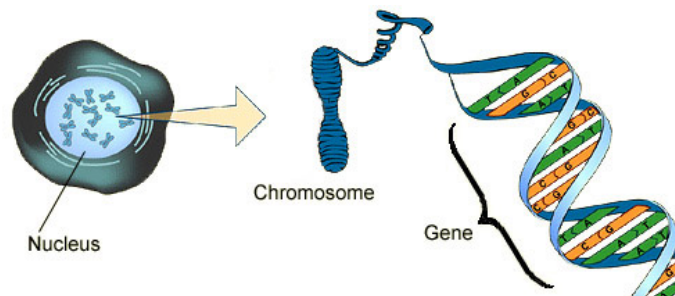


FIGURE 1.5: Nucleus, chromosome and DNA

Alleles The relationship between two different alleles of the same gene is complex and depends of the gene. Some alleles determined integrally the individual, without the contribution of the second allele. These allele are then described as dominant and recessive.

For instance, in the blood group ABO system, the alleles are A, B and O. The possible phenotypes are O, A, B and AB; the O allele is recessive compared to the other two, and the A and B alleles are dominant on O, but A and B are codominant.

Genetic disease Each individual's DNA is half inherited from his father and the other from his mother. Therefore, genetic diseases usually concern not only the affected individual, but also some member of his family (parents, grandparents, brothers, sisters,

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	...
DNA sequence from father	C	T	G	A	A	...
DNA sequence from mother	C	G	G	A	C	...

TABLE 1.1: Sequence of SNPs

	major variant	minor variant
SNP 1	C	A
SNP 2	T	G
SNP 3	G	C
SNP 4	T	A
SNP 5	A	C
⋮	⋮	⋮

TABLE 1.2: Example of table defining major/minor variant in the entire population

uncles, aunts...). A genetic disease is not always inherited, as there are several modes of inheritance: autosomal dominant, autosomal recessive, X-linked.

SNP encoding Single Nucleotide Polymorphisms (or SNP) are the variations of the genome at a specific location. They must concern more than 1% of the population to be considered as a SNP. SNPs represent therefore a tiny portion of the DNA, but account for 90% of human genetic variations. The difference in SNPs can cause disease. For instance, a single pair mutation in the gene APOE (Apolipoprotein E) increases the risk to develop AD.

We give here an illustration of SNP encoding. The table 1.1 represents a sequence of DNA, that is split from its origin. A SNP represents one specific position of the genome.

For each SNP, there is a major variant and a minor variant in the population; the major variant is the variant that the most people have in the entire population. We count for the number of minor variant, using the following mapping:

$$\text{count} : i \mapsto \begin{cases} 0 & \text{if SNP } i \text{ has two major variants} \\ 1 & \text{if SNP } i \text{ has one major variant and one minor variant} \\ 2 & \text{if SNP } i \text{ has two minor variants} \end{cases}$$

If we consider that for the entire population, we have the table 1.2 defining for each SNP the major and minor variant:

We can thus deduce the vector of genetic data:

$$X = \begin{pmatrix} \text{SNP 1} & \text{SNP 2} & \text{SNP 3} & \text{SNP 4} & \text{SNP 5} & \dots \\ 0 & 1 & 0 & 2 & 1 & \dots \end{pmatrix}$$

We can note that the notion of major/minor variant is different from the notion of dominant/recessive variant. The idea is that the average subject would have only major variant, and his vector X of SNPs would be the null vector.

Biological pathways The interactions between proteins (produced by genes) are complex and are usually expressed in terms of biological pathways. A biological pathway is a process through which proteins interact.

1.5 The Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset

1.5.1 Presentation

Presentation The Alzheimer’s Disease Neuroimaging Initiative (ADNI, <http://adni.loni.usc.edu>) collects data for modelling the progression of AD. Data includes cognitive tests, MRI and PET images, CSF, blood biomarkers and genetics. Data are collected, pre-processed, standardized, validated by ADNI. The current database on which researchers of ADNI are working on is the ADNI3 database. The ADNI3 database is built upon the ADNI1, ADNI-GO, ADNI2 database.

ADNI1 dataset In this thesis, we will mainly work with the ADNI1 dataset. The ADNI1 dataset – on the contrary of other ADNI datasets – provides a full SNP genotyping for some subject. The number of SNPs that have been genotyped is around 620,901.

The ADNI1 dataset contains 819 patients with clinical scores, biological biomarkers and genetic variants. Each patient is followed a for a certain time, and several and repeated timepoints are acquired through time. For each timepoint, are acquired clinical scores and biological biomarkers, and the patient’s state is also given: cognitively normal (CN), early mild cognitive impairment (eMCI), late mild cognitive impairment (lMCI), SMC, dementia. Among the MCI subjects (eMCI + lMCI), we prefer to separate them into stable MCI (sMCI) and progressive MCI (pMCI). Stable MCI at T means that the patient remains MCI until *at least* date T , whereas pMCI means that the patient becomes AD before date T .

1.5.2 Descriptive statistics of ADNI1 dataset

Tables 1.3 and 1.4 provide some descriptive statistics for variables measured at study entry of ADNI1 dataset.

Abbreviations used for biological biomarkers are:

- Mid-Temp for the volume of middle temporal gyrus, calculated from medical resonance images;
- Hippo for the volume of hippocampi, calculated from medical resonance images;
- Brain for the volume of the whole brain, calculated from medical resonance images;
- Entorhinal for the volume of the entorhinal region, calculated from medical resonance images;
- FDG for the sum of mean glucose metabolism uptake in regions of angular, temporal, and posterior cingulate, calculated from PET scans;

Variables	AD at baseline	CN at baseline
Total	192	229
Women	91 (47%)	110 (48%)
Number of APOE ϵ 4 alleles		
0	65 (33.9%)	168 (73.4%)
1	91 (47.4%)	56 (24.5%)
2	36 (18.8%)	5 (2.1%)
Age (years)	75.3 (\pm 7.46)	75.7 (\pm 5.02)
Education (years)	14.7 (\pm 3.14)	16.1 (\pm 2.85)
Time in study (years)	1.57 (\pm 0.71)	5.08 (\pm 3.02)
ADAS-Cog13	28.89 (\pm 7.64)	9.50 (\pm 4.19)
RAVLT immediate	23.18 (\pm 7.71)	43.33 (\pm 9.09)
FAQ	15.26 (\pm 7.49)	0.76 (\pm 2.58)
MMSE	23.33 (\pm 2.07)	29.11 (\pm 1.00)
Ventricules [†] (cm ³)	52.8 (\pm 26.8)	37.85 (\pm 19.3)
Hippo [†] (cm ³)	5.56 (\pm 1.13)	7.03 (\pm 0.96)
Brain [†] (cm ³)	947.4 (\pm 107.8)	987.6 (\pm 99.3)
Entorhinal [†] (cm ³)	2.78 (\pm 0.75)	3.69 (\pm 0.66)

TABLE 1.3: Descriptive statistics for variables measured at study entry of ADNI1 participants who are AD or CN at baseline, [†] means that this feature is computed on a subset of ADNI1

Variables	Progressive MCI	Stable MCI	Combined
Total	184	177	361
Women	70 (38%)	59 (33.3%)	129 (36%)
Number of APOE ϵ 4 alleles			
0	64 (34.8%)	97 (54.8%)	161 (44.6%)
1	90 (48.9%)	64 (36.2%)	154 (42.7%)
2	30 (16.3%)	16 (9%)	46 (12.7%)
Age (years)	74.3 (\pm 6.9)	75.4 (\pm 7.6)	74.8 (\pm 7.3)
Education (years)	15.8 (\pm 2.8)	15.46 (\pm 3.17)	15.63 (\pm 3)
Time in study (years)	3.98 (\pm 2.15)	3.5 (\pm 2.69)	3.75 (\pm 2.44)
ADAS-Cog13	20.92 (\pm 5.6)	16.82 (\pm 6.17)	18.91 (\pm 6.23)
RAVLT immediate	27.65 (\pm 6.53)	33.48 (\pm 10.23)	30.51 (\pm 9.02)
FAQ	12.85 (\pm 7.89)	4.24 (\pm 5.2)	8.66 (\pm 7.98)
MMSE	26.69 (\pm 1.73)	27.28 (\pm 1.77)	26.98 (\pm 1.77)
Ventricules [†] (cm ³)	53.4 (\pm 23.9)	47.7 (\pm 26.7)	50.6 (\pm 25.5)
Hippo [†] (cm ³)	5.57 (\pm 1.00)	6.45 (\pm 1.07)	6.00 (\pm 1.12)
Brain [†] (cm ³)	958.7 (\pm 1.14)	993.3 (\pm 100.7)	975.4 (\pm 109.1)
Entorhinal [†] (cm ³)	2.78 (\pm 7.12)	3.4 (\pm 7.3)	3.1 (\pm 0.79)

TABLE 1.4: Descriptive statistics for variables measured at study entry of ADNI1 participants with mild cognitive impairment (MCI), [†] means that this feature is computed on a subset of ADNI1

It can clearly be inferred from these tables that:

- the proportion of subjects having more than one APOE ϵ 4 alleles is higher for the AD patients at baseline;
- the cognitive scores ADAS-Cog13 and FAQ are decreasing when the patient becomes AD; while RAVLT, MMSE are increasing when the patient becomes AD;

- the ventricle volume is much larger for patients having AD than for control patients, whereas the hippocampi is smaller for patients having AD than for control patients.

1.5.3 MCI patients follow-up

In this section, we provide a temporal description of conversion to AD for MCI patients at baseline. For each patient i , we denote T_i^* his conversion time (MCI to AD) and C_i his censored time (time to the end of study). The duration observed for each subject is $T_i = \min(T_i^*, C_i)$.

Observed and censored date Figure 1.6 shows the histogram of the number of visits at each month; visits occur every 6 months; but from 24 months after the baseline date, visits only occur every 12 months. Figure 1.7 gives the histogram of the conversion date T^* and of the censored date C among MCI patients.

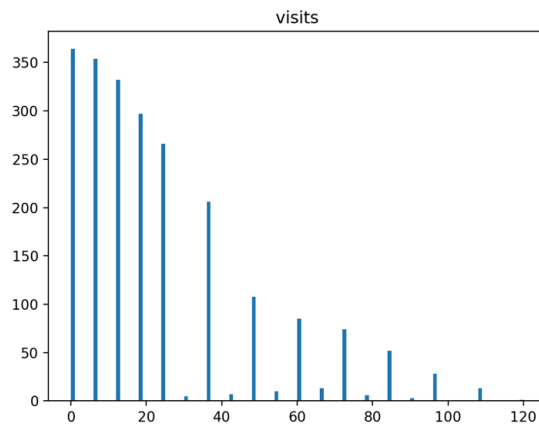


FIGURE 1.6: Histogram of the number of visits at each month among MCI patients at baseline

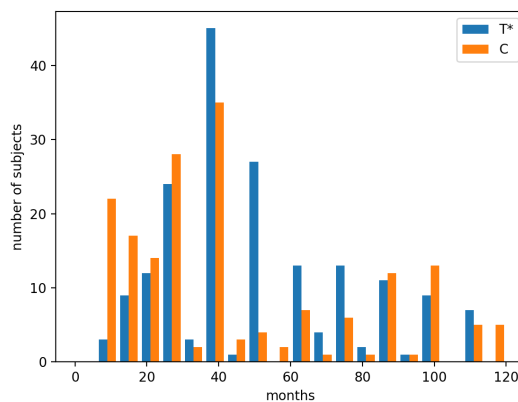


FIGURE 1.7: Histogram of the conversion date T^* and the censored date C

Estimates of the survival and hazard functions Based on figure 1.7, it is possible to compute an estimator of the survival probability $S : t \mapsto \mathbb{P}\{T \geq t\}$ is the conversion time to AD for MCI subjects. Details on how to compute such estimator is given in chapter 3.

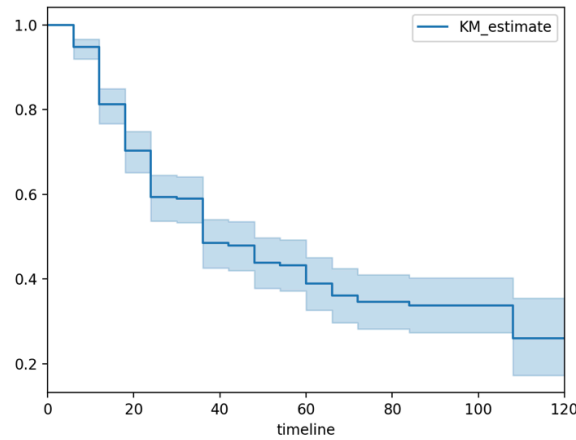


FIGURE 1.8: Kaplan-Meier estimator for the survival function for MCI patients from ADNI1 dataset

It is also possible to provide an estimator of the cumulative hazard function $\hat{H} : t \mapsto \int_0^t h(u)du$ where h is the hazard function, using the Nelson-Aalen estimator. Descriptions on this estimator are given in chapter 3. The hazard function can be deduced from the cumulative hazard function using the formula

$$\hat{h}(t) = \frac{\hat{H}(t + \delta) - \hat{H}(t - \delta)}{2\delta}$$

where δ represents the bandwidth and must be chosen according to the dataset. In our case, given the fact that visits occur every 6 months, we chose $\delta = 6$ months.

The hazard function for the ADNI1 dataset seems to be unimodal. There are some artefacts from 30 months, because visits only occur every 12 months, and a bandwidth of 12 months would be more appropriate to estimate the hazard function. Furthermore, from 100 months, there are very few visits, and most of them represent converted subjects, and therefore, the estimation of the hazard function from 100 months is not reliable.

1.6 Conclusion

Alzheimer's Disease is a form of dementia that creates behavioural and memory troubles. The different modalities involved in the study of Alzheimer's Disease provide different kind of informations. On the one hand, neuropsychological tests provide a clear answer when it comes to assess the diagnosis of AD. Neuroimaging data, such as anatomical MRI or PET scans, can refine the diagnosis realised through neuropsychological tests. On the other hand, genetic data – in particular, having some alleles of some genes such as APOE – can help to identify the risks to develop AD in the future. In the next chapter, we will provide a state of the art of imaging genetics and how to combine both neuroimaging and genetic modalities for the diagnosis and prediction of AD.

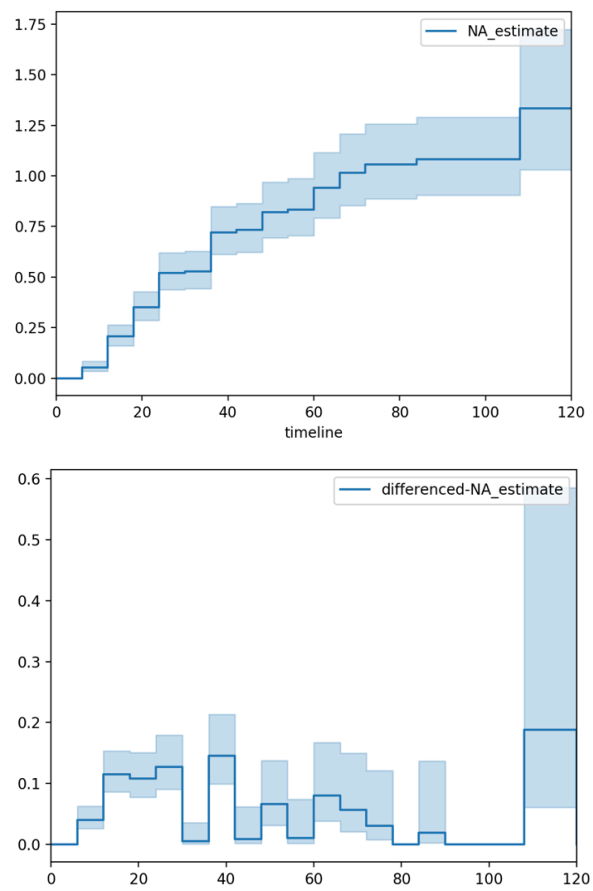


FIGURE 1.9: Nelson-Aalen estimator for the cumulative hazard function (top) and estimated hazard function \hat{h} with bandwidth of 6 months (bottom)

Chapter 2

Statistical and machine learning approaches for Imaging Genetics

2.1 Introduction

In this chapter, we provide a review of the state of the art in Imaging Genetics [Liu and Calhoun, 2014] and its applications to Alzheimer's Disease. Imaging Genetic studies the relation between genetic variations and individual endophenotypes such as neuroimages. The information provided by endophenotypes is closer to the biology of genetic function than the clinical phenotype (disease diagnosis).

This research area focuses on the influence of genes on brain and its pathologies. Using the information from neuroimages and genetics, it try to determine which differences in SNPs can lead to brain pathologies. The main idea behind imaging genetics is that common variants in SNPs lead to common diseases.

The first topic of research in imaging genetics is to study the association between genetic and brain imaging data, using a univariate or multivariate approach. A univariate analysis will associate the genotype to one specific phenotype trait, whereas a multivariate analysis will associate the genotype to several phenotype traits such as a whole neuroimage. Example of multivariate approaches are the partial least squares [Laird and Lange, 2011], the sparse canonical correlation analysis [Du et al., 2014], the sparse regularised linear regression with a ℓ_1 -penalty [Kohannim et al., 2012a], the use of the group lasso penalty [Silver et al., 2012, Silver and Montana, 2012], or a Bayesian modeling that links genetic variants to imaging regions and imaging regions [Chekouo et al., 2016, Batmanghelich et al., 2016]. In these three last examples, the authors consider that the endophenotype can be explained as a sum of effects from genetic variants.

The second topic of research in imaging genetics is to combine these two different modalities (imaging and genetics) for automatic classification or disease evolution of patients. Researchers using such approach consider that the genotype and the neuroimages provide complementary information concerning the subject's disease state, or that knowing the genotype can help to refine the diagnosis using only the imaging modality. In [Wang et al., 2012, Peng et al., 2016] for instance, the authors uses machine learning methods to build predictors for Alzheimer's disease (AD) diagnosis. The main challenges in this topic of research is that the heterogeneous nature of data, and the way both imaging and genetics can be combined efficiently.

Both topic of research are summarised in figure 2.1. In both cases, a challenging issue is the high-dimension of data due to small number of observations.

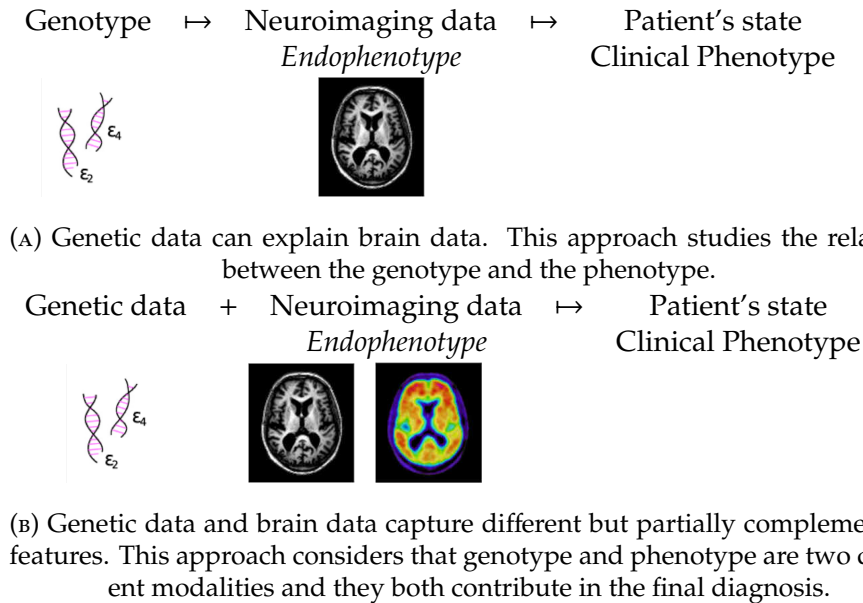


FIGURE 2.1: Two topics of research in imaging genetics

2.2 Univariate and multivariate analyses between genetic and neuroimaging data

In this section, we describe how to explain neuroimages data from genetic data. In [Liu and Calhoun, 2014], the authors have characterise imaging genetics analyses into 4 different types:

- The first type is to consider one gene and one phenotype trait and to perform a univariate analysis.
- The second type is to consider the whole genotype and to perform a univariate analysis with one phenotype trait.
- The third type is to consider one gene and a multiple imaging phenotypes such as brain regions or brain voxels.
- The fourth type is to consider the whole genome and a multiple imaging phenotype and to perform a multivariate analysis between them.

All these type of associations are summarised in figure 2.2. The two main research areas are the univariate association between genotype and one phenotype trait (type 2) and the multivariate analyses, i.e. association between multiple genotypic variables and phenotypic variables (type 4).

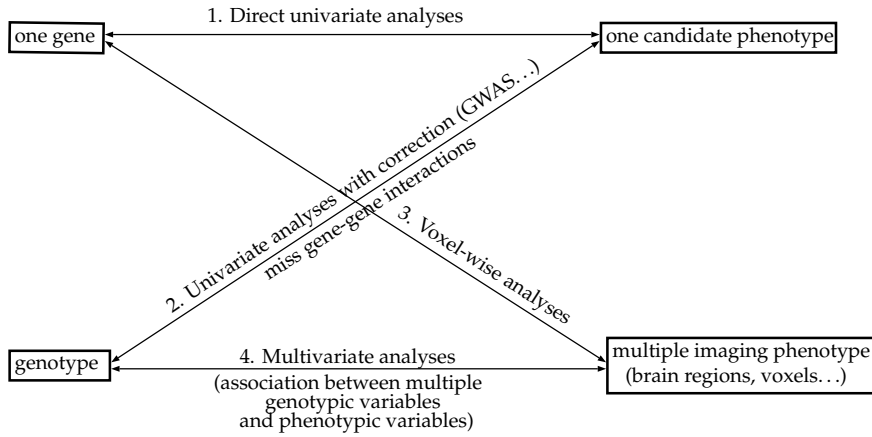


FIGURE 2.2: Four categories of analyses in imaging genetics, reproduced from [Liu and Calhoun, 2014]

2.2.1 Univariate association between genotype and one phenotype trait

Genome Wide Association Study (GWAS) A Genome Wide Association Study (GWAS) is an analysis regarding genetic variations in many individuals to study their correlations with one phenotypic trait. It typically focuses on the association between SNPs and phenotypes, such as diseases.

A comparison of DNA sequences is performed between individuals having several different phenotypes for the same disease. It is intended to identify genetic variants that are most frequently associated with the phenotype. In other words, it consists in categorising patients depending on their disease status, and correlating them with their SNPs.

However, GWAS does not give a causal relationship of the SNPs to the disease but rather informs if they are correlated with this disease. GWAS can have troubles in detecting correlations, as the statistical effect of each SNP (taken independently) is very weak. Furthermore, statistical interdependences between genes and SNPs are ignored in GWAS.

Methodology [Laird and Lange, 2011] Each SNP is independently tested to verify if there is an association with the disease. For each SNP ($m = 1, \dots, M$), we consider the null hypothesis $H_0^{(m)}$, which is there is no association between the SNP m and the disease.

On the one hand, based on data, we can deduce the real distribution, shown on the left. On the other hand, we can compute the hypothetic distribution if the null hypothesis is true.

These two distributions are compared by computing the random variable X :

$$X = \sum_{\text{status}} \sum_{\text{allele}} \left(\frac{\widehat{N}_{\text{status,allele}} - N_{\text{status,allele}}}{N_{\text{status,allele}}} \right)^2 \sim \chi^2(\text{deg} = 1)$$

This variable X follows a χ^2 law of degree 1. The p -value is given by $1 - F^{-1}(X)$. The p -value is the probability of observing something as/more extreme as/than the observed test statistic given that the $H_0^{(m)}$ is true.

Real distribution			Hypothetic distribution if $H_0^{(m)}$ holds			
	Allele A	Allele a		Allele A	Allele a	
CN	$N_{A,CN}$	$N_{a,CN}$	N_{CN}	$\widehat{N}_{A,CN} = \frac{N_A N_{CN}}{N}$	$\widehat{N}_{a,CN} = \frac{N_a N_{CN}}{N}$	N_{CN}
AD	$N_{A,AD}$	$N_{a,AD}$	N_{AD}	$\widehat{N}_{A,AD} = \frac{N_A N_{AD}}{N}$	$\widehat{N}_{a,AD} = \frac{N_a N_{AD}}{N}$	N_{AD}
	N_A	N_a	N	N_A	N_a	N

FIGURE 2.3: Real vs hypothetic distribution (N is the number of subjects, N_A (resp. N_a) the number of subjects with allele A (resp. a), N_{CN} (resp. N_{AD}) the number of subjects who are CN (resp. have AD))

Finally, if the p -value is less than $\frac{\alpha}{M}$, where alpha is the probability to reject the null knowing the null hypothesis $H_0^{(m)}$, we reject the null hypothesis and we consider that there is an association between the SNP and the disease; it is the Bonferroni correction.

We usually prefer to work with $-\log p$ instead of p , as p and $\frac{\alpha}{M}$ are very small. Therefore, we say that there is an association between the SNP and the disease if the $-\log p$ is higher than that line, $-\log\left(\frac{\alpha}{M}\right)$.

Application to Alzheimer's Disease We performed the GWAS on the ADNI1 dataset, and results are shown on figure 2.4. The GWAS identifies highly replicated gene for AD, such as APOE. The x -axis lists the SNPs and the shades indicate different chromosomes. The y -axis reports the negative \log_{10} of the p -value. The horizontal line denotes the statistical significance level (0.05) after Bonferroni correction.

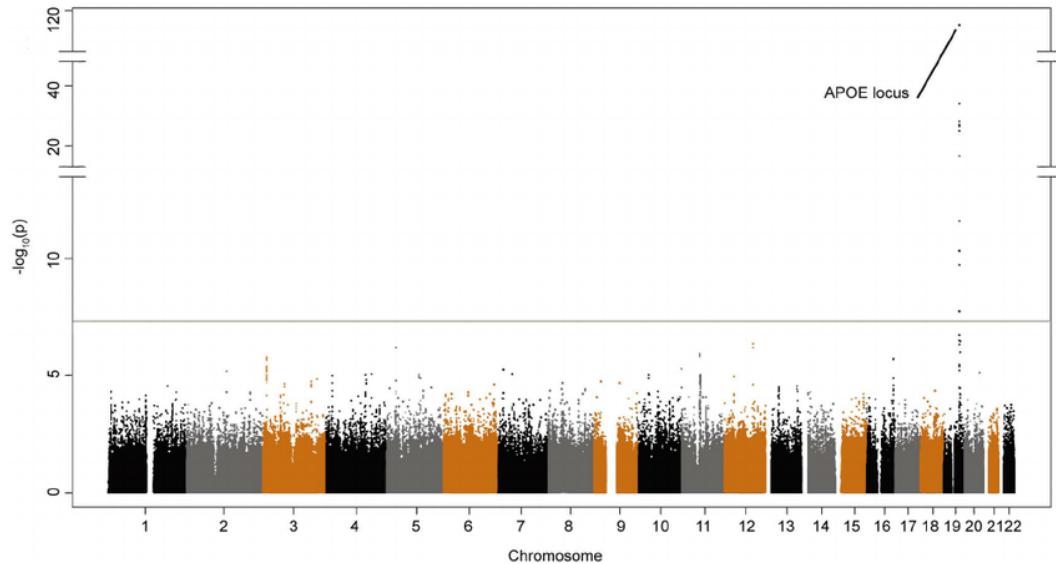


FIGURE 2.4: Genome wide meta-analysis results in AD with 524,993 SNPs. Manhattan plot showing the p -values obtained in the meta-analysis. The end and beginning of a chromosome is denoted by the change of colour pattern of the SNPs (black, grey and brown dots). Genome-wide significance threshold is denoted by a red line. The Y-axis has been truncated (reproduced from [Perez et al., 2014])

2.2.2 Multivariate analyses

Multivariate analyses are more appropriate when we want to explain several biomarkers by a combination of SNPs. There are modelling that only built the association matrix the genetic SNPs, and the neuroimaging features such as the Partial Least Squares (PLS); and other modelling, that compute a linear regression of the biomarkers as a function of genetic data:

$$\text{biomarker} \approx \sum_i \alpha_i \text{SNP}_i$$

These modelling are more frequent in the literature of imaging genetics; we can cite for instance the LASSO regression between SNPs and neuroimaging data (for sparse modelling), Bayesian modelling or random forests on SNPs. The underlying assumption is that these models assume no interactions between SNPs, and that they provide independent effects from each other.

In all cases, multivariate analyses deal with high dimensional data: the number of samples N is much smaller than the number of SNPs M or the number of brain regions q .

2.2.2.1 Partial Least Squares

Among methodological approaches for capturing significant genotype-phenotype interactions, we can cite the partial least squares (PLS) or independent component analysis (ICA). They perform simultaneous regression and dimensionality reduction strategies. Partial least squares (PLS) is an attractive approach because it provides a parsimonious description of multivariate correlation models. Furthermore, it is simple to implement [Lorenzi et al., 2016].

Let K be the number of subjects, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_K) \in \mathbb{R}^{N \times K}$ be the matrix of genetic SNP data and $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_K) \in \mathbb{R}^{M \times K}$ the matrix of neuroimaging data. Both matrices are assumed to be normalised. The goal of PLS is to build two matrices \mathbf{U} and \mathbf{V} that maximises the covariance between \mathbf{XU} and \mathbf{YV} :

$$\max_{\|\mathbf{U}\|=\|\mathbf{V}\|=1} \text{cov}(\mathbf{XU}, \mathbf{YV})$$

Using the singular value decomposition (SVD), the previous formulation is equivalent to find unitary matrices \mathbf{U} and \mathbf{V} such that

$$\mathbf{XY}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{V} \quad \text{with } \mathbf{U} \in \mathcal{U}_M(\mathbb{R}), \mathbf{V} \in \mathcal{U}_K(\mathbb{R})$$

The built mapping $(\mathbf{x}, \mathbf{y}) \mapsto (\mathbf{x}^T \mathbf{U}, \mathbf{y}^T \mathbf{V})$ provides the low dimensional representation on the latent PLS space.

The authors have tested on the ADNI1 and ADNI2 datasets. Figure 2.5 shows the main PLS eigen-component for \mathbf{V} (on the left) for the phenotype feature, and on the right, the main PLS eigen-component for \mathbf{U} are shown. In the first component of \mathbf{V} , ventricles volume is anti-correlated wrt the volume of the other brain areas, whereas on the second component, the hippocampal volume is anti-correlated wrt the other brain structures. In

the matrix \mathbf{U} , it can be seen that chromosome 19 has a largest weight on the first component of \mathbf{U} . Chromosome 19 is the chromosome which contains the most gene APOE.

On the contrary of GWAS, PLS models the joint correlation of SNPs and neuroimaging features and overcomes the classical multiple comparison.

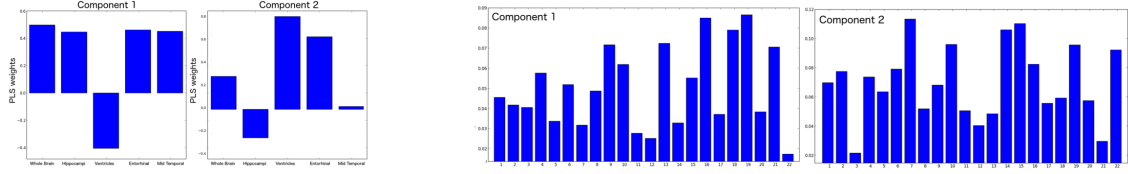


FIGURE 2.5: (left) Main PLS eigen-component of \mathbf{V} for the phenotype features. (right) Chromosome representativeness among the set of most informative SNPs associated to the main PLS eigen-component of \mathbf{U} [Lorenzi et al., 2016].

2.2.2.2 Linear regression between SNPs and neuroimaging data

A simple way to link genetic data (denoted $\mathbf{x} \in \mathbb{R}^m$, where m is the number of SNPs) and neuroimaging variables (denoted $\mathbf{y} \in \mathbb{R}^q$ where q is the number of brain regions) is to perform a linear regression between them.

$$\underbrace{(y_1, \dots, y_q)}_{=\mathbf{y}} = \underbrace{(\tilde{x}_1, \dots, \tilde{x}_m)}_{=\tilde{\mathbf{x}}} \mathbf{w}^* + \mathbf{b}^*$$

As we deal with high dimensional data, it is interesting to penalise \mathbf{w}^* using the LASSO penalty, that provides sparsity in \mathbf{w}^* . Coefficients \mathbf{w}^* and \mathbf{b}^* are find as:

$$(\mathbf{w}^*, \mathbf{b}^*) \in \underset{(\mathbf{w}, \mathbf{b})}{\operatorname{argmin}} \underbrace{\sum_{i=1}^N \sum_{k=1}^q \left(\tilde{y}_k^i - \sum_{j=1}^m w_{k,j} \tilde{x}_j^i - b_k \right)^2}_{=\|\mathbf{Y} - \tilde{\mathbf{X}}\mathbf{w}\|_2^2} + \underbrace{\frac{1}{2} \sum_{k,j} |w_{k,j}|}_{=\|\mathbf{w}\|_1}$$

where $\mathbf{Y} \in \mathbb{R}^{N \times q}$ represents the neuroimaging measures, $\tilde{\mathbf{X}} \in \mathbb{R}^{N \times m}$ the genotype, and N the number of subjects.

In [Kohannim et al., 2012b], they performed this model on each gene (\mathbf{X} represents one single gene), in order to select the most representative SNPs inside a gene. Furthermore, a F -test is realised for subsets of SNP within genes in order to verify that each SNP really has a contributive effect. Selection of most useful SNPs within a gene is realised through partial F -tests were performed for each gene.

2.2.2.3 Generative models

Generative models are widely used in imaging genetics association studies. On the contrary of a discriminative model that would model the conditional probability of the imaging Y given the genetics \mathbf{x} , noted $\mathbb{P}\{Y|\mathbf{X} = \mathbf{x}\}$, a generative model would model the joint probability of $X \times Y$, also noted $\mathbb{P}\{X, Y\}$. These models have some common points, as they

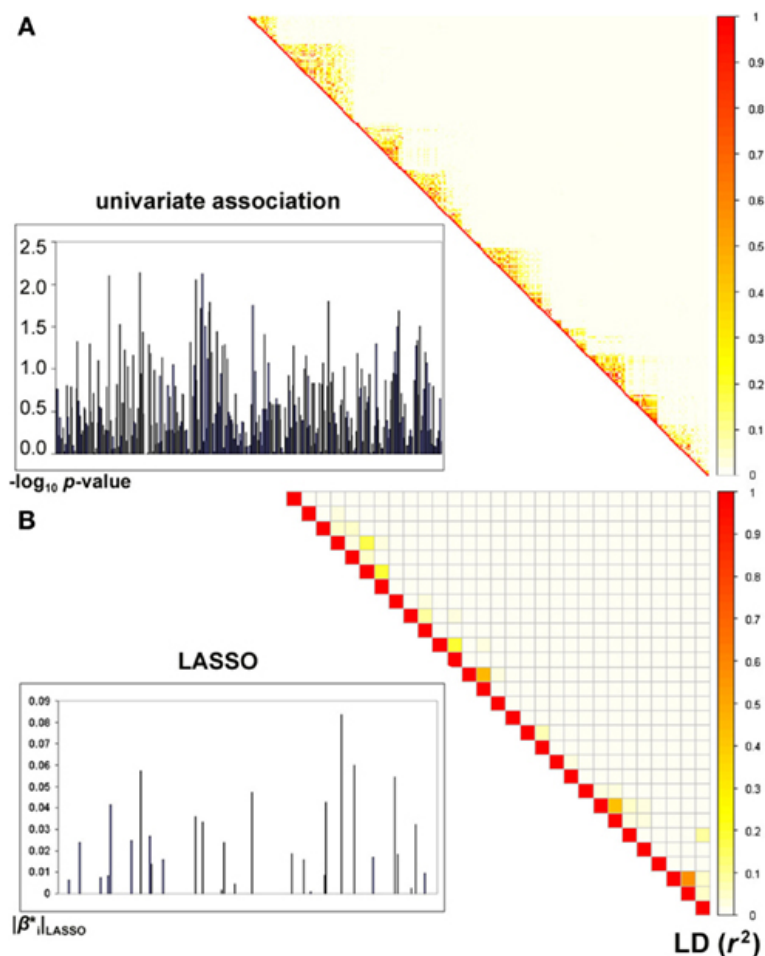


FIGURE 2.6: Comparison between a univariate association test and a LASSO regression with an imaging-derived measure of temporal lobe volume. (reproduced from [Kohannim et al., 2012b])

use linear models to describe Y as a function X , but will not make the same hypothesis and assumptions on the hidden parameters. These models are also complex to design; examples of papers using these models are [Batmanghelich et al., 2016, Chekouo et al., 2016]. In this section, we'll analyse [Batmanghelich et al., 2016].

Bayesian modelling Let \mathbf{x} be the vector of SNP (genetic data) and $\mathcal{D} = \{\mathbf{Y}, \mathbf{z}\}$ be the phenotype (imaging data for \mathbf{Y} , and diagnosis for \mathbf{z}). The goal of a generative model is to use a Bayesian framework in order to link these three variables. Let k be the k^{th} brain region.

For this brain region, \mathbf{a}_k is the vector that will choose the SNPs from the vector \mathbf{x} in order to model a brain region k . Therefore

$$a_{m,k} = \begin{cases} 0 & \text{if SNP } m \text{ is not related to brain region } k \\ 1 & \text{if SNP } m \text{ is related to brain region } k \end{cases}$$

It is assumed that $a_{m,k}$ follows a Bernoulli law of parameter α .

Let b_k be an indicator that indicates if the brain region k is related to the disease z . It is assumed that b_k follows a Bernoulli law of parameter β . If $b_k = 0$, the brain region k will not be studied.

The imaging feature y_k for the brain region k is such that:

$$y_k \sim \begin{cases} \mathcal{N}(0, 1) & \text{if } b_k = 0 \\ \mathcal{N}(\omega_k^\top (\mathbf{x} \cdot \mathbf{a}_k), \sigma_0^2) & \text{if } b_k = 1 \end{cases}$$

where ω_k is the regression coefficient vector, and such that $\omega_k \sim \mathcal{N}(0, \sigma_\omega^2)$. In other words, y_k can be explained as an affine function of genetics. If $b_k = 1$:

$$y_k \approx b_k \underbrace{\left(\sum_m \omega_{k,m} a_{m,k} x_m \right)}_{\text{genetics contribution}} + \underbrace{\varepsilon_k}_{\text{noise}}$$

Finally, the disease status $z \in \{-1, +1\}$ is deduced from the imaging phenotype using a logistic regression:

$$\mathbb{P}\{z = 1 | f, \mathbf{b}, \mathbf{y}\} = \sigma(f(\mathbf{y} \cdot \mathbf{b})) = \sigma\left(f(b_1 y_1, \dots, b_q y_q)\right)$$

where σ is a sigmoid function and f is a gaussian process (latent function).

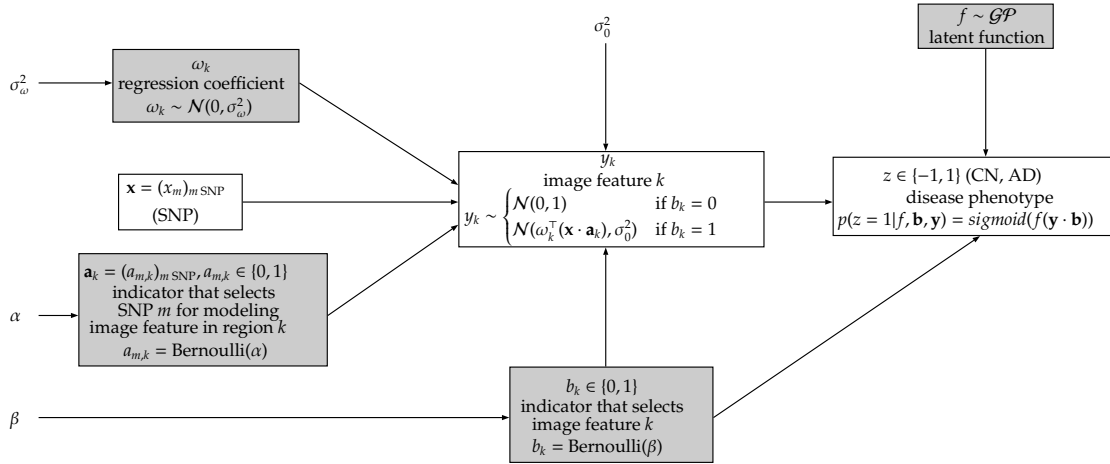


FIGURE 2.7: Relationship between genetic, imaging and clinical measure (synthesised from [Batmanghelich et al., 2016])

Hyperparameters and hidden variables The hyperparameters $\pi = \{\sigma_0^2, \sigma_\omega^2, \alpha, \beta\}$ are fine tuned by the user whereas the hidden variables $\mathcal{Z} = \{f, \mathbf{b}, \mathbf{a}_1, \dots, \mathbf{a}_q, \omega_1, \dots, \omega_q\}$ need to be determined by learning.

For estimating the probability $p(\mathbf{a}_k | \mathbf{x}, \mathbf{y}_k)$ (the association between SNPs and brain region k), the authors used importance sampling and replace the integral by a finite sum. For

$i \in \{1, \dots, L\}$, they set $\pi'(i) = (\log_{10} \alpha(i), \sigma_0^2(i), \sigma_\omega^2(i))$. Then,

$$\begin{aligned} p(a_{m,k} = 1 | \mathbf{x}, \mathbf{y}_k) &= \int p(a_{m,k} = 1 | \mathbf{x}, \mathbf{y}_k, \pi') p(\pi' | \mathbf{x}, \mathbf{y}_k) d\pi' \\ &= \frac{\sum_{i=1}^L p(a_{m,k} = 1 | \mathbf{x}, \mathbf{y}_k, \pi'(i)) \zeta(\pi'(i))}{\sum_{i=1}^L \zeta(\pi'(i))} \end{aligned}$$

where $\zeta(\pi) \propto p(\pi | \mathbf{x}, \mathbf{y}_k) \frac{p(\pi)}{\tilde{p}(\pi)}$ and $\tilde{p}(\cdot)$ is the proposal distribution, in that case, they chose the uniform distribution. For computing $p(a_{m,k} = 1 | \mathbf{x}, \mathbf{y}_k, \pi'(i))$, they use the EM algorithm.

Results In this model, there is no direct association between genetics and the disease, as it focuses on the quantification of genetic associations with imaging variables.

There is a selection of intermediate imaging phenotypes influenced by genetic markers and relevant to the disease. The authors also assume that there is Independence between SNPs and between brain regions. In this model, there is no arbitrary threshold selections. The study of the posterior probability $p(\mathbf{a}_k | \mathbf{x}, \mathbf{y}_k)$ can help to identify genetics that influence brain regions.

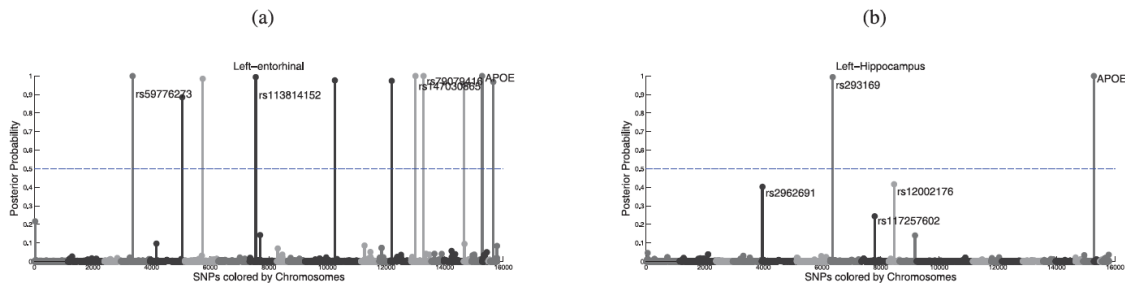


FIGURE 2.8: Posterior relevance of the SNPs $p(\mathbf{a}_k | \mathbf{x}, \mathbf{y}_k)$ with respect to (a) average thickness of the left entorhinal cortex, and (b) volume of the left hippocampus

2.2.2.4 Pathways-based regression modelling

As explained in chapter 2, a gene encodes for a specific protein, and some proteins might interact together in a biological pathway. In other word, a pathway is a series of interactions among proteins in a cell. These interactions conduct to some change in the cell.

In this section, we propose a concise review of the integration of pathways in imaging genetics. Figure 2.9 proposes a simplified representation. For each gene, we can consider only the genotyped SNPs that belong to that gene. A SNP can map to several genes (red squares). A gene encode for a specific protein, some SNPs map to known genes, but there are SNPs that don't map to any gene (white square). The proteins interact within pathways; a same protein can interact in multiple pathways. Therefore, a SNP can map to several genes, and map to several pathways.

The mapping genes to pathway is actually more complicated, as many genes do not map to any known pathway (unfilled circles), while some genes may map to more than one pathway. On the contrary, genes that map to a pathway are in turn mapped to genotyped

SNPs within a specified distance. Concerning SNPs, many SNPs cannot be mapped to a pathway since they do not map to a mapped gene (unfilled squares). Some SNPs (red squares) may map to more than one pathway.

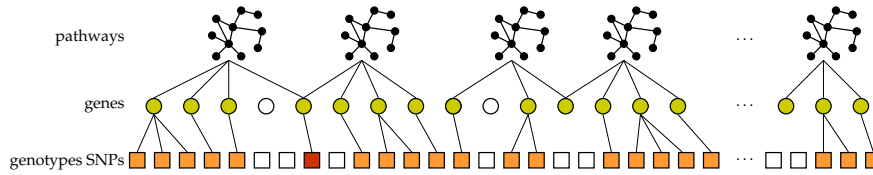


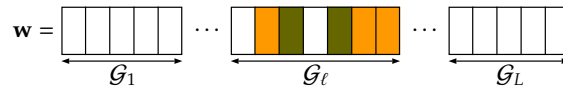
FIGURE 2.9: Pathways, genes and SNPs

Example of pathway database is the KEGG (Kyoto Encyclopedia of Genes and Genomes) database; the pathway they propose for AD can be seen at https://www.genome.jp/kegg-bin/show_pathway?hsa05010. Pathways are experimentally determined, and it is therefore possible to have several pathways for the same disease.

In [Silver and Montana, 2012], the authors propose to group genotyped SNPs by pathways, which is the correct settings for biological process. They perform a regression model that selects all SNPs grouped into pathways, using a group LASSO, where SNPs are grouped into L pathways, denoted $\mathcal{G}_1, \dots, \mathcal{G}_L$. The neuroimaging features \mathbf{y} are then given by:

$$\mathbf{y} = \sum_{\ell=1}^L \langle \mathbf{w}_{\mathcal{G}_\ell}, \mathbf{x}_{\mathcal{G}_\ell} \rangle + b \quad \text{and} \quad (\mathbf{w}^*, b^*) \in \underset{(\mathbf{w}, b)}{\operatorname{argmin}} \left\{ \gamma \|\mathbf{Y} - \mathbf{X}\mathbf{w}\|^2 + \frac{1}{2} \sum_{\ell=1}^L \lambda_\ell \|\mathbf{w}_{\mathcal{G}_\ell}\|_2 \right\}$$

This approach selects only most predictive pathways:



The Group LASSO penalty provides sparsity between groups (in this case between pathways), but no sparsity inside each group.

2.2.2.5 Epistasis effects and Random Forests on Distance Matrices

As shown in the previous section, there are interactions between genes; and therefore, we can consider that interactions between SNPs. But there are also interactions between genes, as the the presence of one gene, can modified the behaviour of another. This is called *epistasis effects*. There is epistasis when one or several genes (dominant or recessive) can prevent the expression of factors located at other genetic sites (locus). Mathematically, this phenomenon can be modelled using the products between SNPs:

$$\text{biomarker} = \sum_i \alpha_i \text{SNP}_i + \sum_{i,j} \beta_{i,j} \text{SNP}_i \times \text{SNP}_j$$

Another possible modelling is to use graphs or trees. In the tree of Figure 2.10, reproduced from [Silver and Montana, 2012], the number of minor variant for SNP α_1 influences

other SNPs. Trees have the advantage to be easily interpretable and doesn't need prior informations on possible interactions.

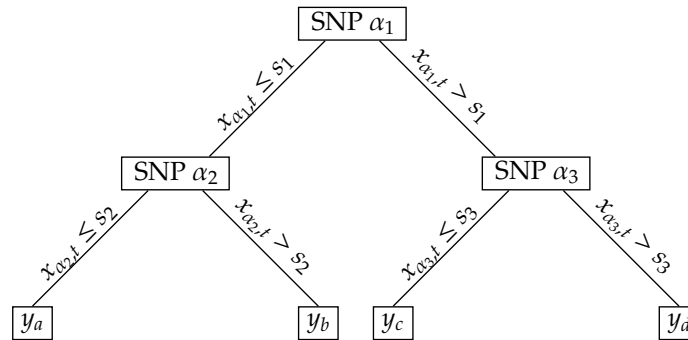


FIGURE 2.10: A tree in a random forest [Silver and Montana, 2012]

We describe how these trees or random forests using these trees can be built [Sim et al., 2013]. Let $S = \{(x^i, y^i, z^i), i = 1 \dots N\}$ the dataset where $x^i = (x_1^i, \dots, x_M^i) \in X$ is the set of SNPs, $y^i \in Y = \mathbb{R}^q$ the voxel-wise brain volumes and $z^i \in Z = \{0, 1, 2\}$ the disease status (corresponding to CN, MCI, AD). In [Sim et al., 2013], the authors took X as the vector of SNPs from chromosome 19 only (7,848) and Y the vector describing the brain longitudinal slope coefficients (148,023).

Usually, we can classify x based on the euclidean distance on \mathcal{Y} . However, the approach chosen was to project the phenotypes y onto a manifold space and to compute the distance matrix, based on the Euclidian distance in the manifold space.

However, these trees will be used for spatial coding, not classification. During a query, for each x , each tree is crossed from the root down to a leaf and the returned label is the unique leaf index, not the (set of) descriptor label(s) y associated with the leaf. An unsupervised transformation of a dataset to a high-dimensional sparse representation. A datapoint is coded according to which leaf of each tree it is sorted into. Using a one-hot encoding of the leaves, it leads to a binary coding with as many ones as there are trees in the forest.

For predicting z from y , a supervised dimensional reduction is realised with a Classification Random Forest. A pairwise proximity matrix between all the subjects is obtained from the endophenotypic vectors y .

They built a forest of 4,000 trees with minimum node size of 5 subjects and a depth equal to 6. Although this model has high complexity, it captures epistasis effects through a random forest.

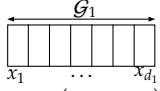
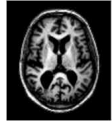
2.3 Combinaison of genetic and neuroimaging data for disease diagnosis

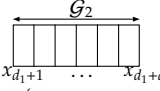
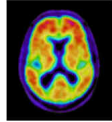
In this section, we describe the state of the art for combining genetics and neuroimaging data in order to predict patient's disease state. Authors that have developed this approach

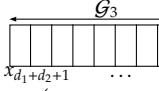
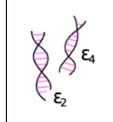
consider that genetics and neuroimaging data provide different type of information concerning the patient's disease state, but that they are complementary and that genetics data can refine the diagnosis given by neuroimaging data.

Notations We denote \mathcal{G}_1 the gray matter volume of region of interest in MRI (dimension d_1), \mathcal{G}_2 the average intensity of region of interest in PET (dimension d_2) and \mathcal{G}_3 the (selected) SNPs (dimension d_3). There are $L = 3$ modalities, and the dimension is $M = d_1 + \dots + d_L$. The input feature \mathbf{x} is the concatenation of the genetic modality, the MRI modality and the PET modality:

$$\mathbf{x} = \underbrace{(x_1, \dots, x_{d_1})}_{=\mathbf{x}_{\mathcal{G}_1}} \underbrace{(x_{d_1+1}, \dots, x_{d_1+d_2})}_{=\mathbf{x}_{\mathcal{G}_2}} \underbrace{(x_{d_1+d_2+1}, \dots, x_{d_1+d_2+d_3})}_{=\mathbf{x}_{\mathcal{G}_3}} \in \mathbb{R}^M$$

\mathcal{G}_1

 $\mathbf{x}_{\mathcal{G}_1} = (x_1, \dots, x_{d_1})$


\mathcal{G}_2

 $\mathbf{x}_{\mathcal{G}_2} = (x_{d_1+1}, \dots, x_{d_1+d_2})$


\mathcal{G}_3

 $\mathbf{x}_{\mathcal{G}_3} = (x_{d_1+d_2+1}, \dots, x_{d_1+d_2+d_3})$


The dataset is denoted $\{\mathbf{x}^i, y^i\}$, for $i \in \{1, \dots, N\}$, where N is the number of subjects. We denote X the feature matrices of $\mathcal{M}_{N,M}(\mathbb{R})$ containing all the \mathbf{x}^i .

Objectives We would like to obtain the subject's diagnosis $y \in \{\text{AD}, \text{CN}\}$ given the input vector \mathbf{x} :

$$y = f(\mathbf{x})$$

There are several ways to define f , as it is a classification problem. For instance, if we use the support vector machine (SVM) framework, f can be defined as:

$$f(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b) = \text{sign} \left(\sum_{j=1}^M w_j \phi_j(\mathbf{x}) + b \right)$$

with $y \in \{-1, +1\}$ and where parameters \mathbf{w}, b are learnt using the samples $\{\mathbf{x}^i, y^i\}$ by solving the following the following loss function:

$$\min_{\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}} \left\{ \frac{1}{N} \sum_{j=1}^N (1 - y^i f(\mathbf{x}^i))_+ + \gamma \|\mathbf{w}\|^2 \right\}$$

where $x_+ = \max(x, 0)$.

Another way to solve this classification problem is to use a logistic regression:

$$f(\mathbf{x}) = \mathbb{P}\{y = 1|\mathbf{x}\} = \frac{1}{1 + e^{-\langle \beta, \mathbf{x} \rangle - \beta_0}}$$

with $y \in \{0, +1\}$. The parameters β, β_0 are determined by minimizing the loss function:

$$J(\beta, \beta_0) = \frac{1}{N} \sum_{i=1}^N \left(-y^i \log f(\mathbf{x}^i) + (1 - y^i) \log(1 - f(\mathbf{x}^i)) \right)$$

2.3.1 Dealing with high dimensional data

The high dimensional hypothesis signifies $M \gg N$. The gram matrix $X^T X$ is an element of $\mathcal{M}_M(\mathbb{R})$, but its rank is lower than $N \ll M$. In this case, this matrix is not invertible and ill-conditioned. It is therefore not possible to solve the previous problem. The standard improvement is to minimise the loss with an additional penalty:

$$\min_{f \in \mathcal{F}} \underbrace{\frac{1}{N} \sum_{i=1}^N \mathcal{L}(f(\mathbf{x}^i), y^i)}_{\text{empirical risk}} + \gamma \underbrace{\Omega(f)}_{\text{penalty}}$$

where \mathcal{L} is the loss function (it depends on the definition of f), Ω is a penalty that will impose regularisation on the function f , and $\gamma > 0$. The penalised problem must be convex w.r.t. the coefficients of the function f , and therefore can be easily solved.

We illustrate different possible penalties, using the logistic regression model.

Ridge regression The ridge penalty is defined by $\Omega(\beta) = \|\beta\|_2^2$. It adds squared magnitude of coefficients to the loss function. When $\gamma = 0$, we get the ordinary minimization problem. When γ becomes large, the penalty will add too much weight to the loss function, and will lead to under-fitting. The ridge penalty usually imposes some kind of regularity in the coefficients of β . The choice of γ is crucial to avoid under-fitting and overfitting issue.

LASSO regression (Least Absolute Shrinkage and Selection Operator) The LASSO penalty is defined by $\Omega(\beta) = \|\beta\|_1$. It adds absolute magnitude of coefficients to the loss function. The difference between the ridge and the LASSO regression, is that the LASSO forces less important features' coefficients to 0, and therefore it removes some features from the model. The larger γ is, the more features will be removed from the model. If γ is too big, no features will be selected in the model. In other words, the LASSO regression provides sparsity between features.

Group LASSO regression Let $\mathcal{G}_1, \dots, \mathcal{G}_L$ be L groups such that $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset$ and $\mathcal{G}_1 \cup \dots \cup \mathcal{G}_L = \{1, \dots, M\}$. The Group LASSO penalty [Ming and Yi, 2006] is defined by:

$$\Omega(\beta) = \sum_{\ell=1}^L \alpha_{\ell} \|\beta_{\mathcal{G}_{\ell}}\|_2$$

where $\beta_{\mathcal{G}_{\ell}}$ is the vector of β where indices \mathcal{G}_{ℓ} are chosen, and $\alpha_{\ell} > 0$ is the weight of the group ℓ (usually, $\alpha_{\ell} = \sqrt{|\mathcal{G}_{\ell}|}$).

The Group LASSO provides sparsity between groups, and not between variables.

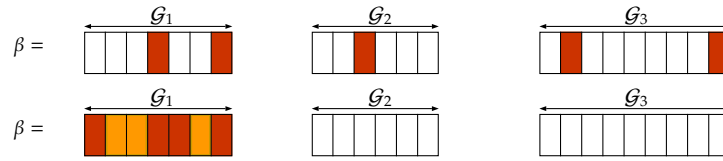


FIGURE 2.11: Illustration of β learnt using the LASSO regression (top), and the Group LASSO regression where $L = 3$ (bottom). The LASSO regression selects only some variables, where the Group LASSO regression selects only some groups.

Overlap Group LASSO regression When dealing with genetics data, groups can be genes or pathways. Genes (or pathways) do overlap, as a SNP can belong to several genes (or pathways). It is not possible to use the original Group LASSO penalty. The solution is to use a design matrix expansion by SNP duplication, although it leads to non orthogonal group (if X is the design matrix, the orthogonal assumption is $X^T X = I$).

Algorithms that solve numerically the Group LASSO must be updated for non orthogonal groups. An efficient way of estimation is to use the block-coordinate descent algorithm for non-orthogonal groups.

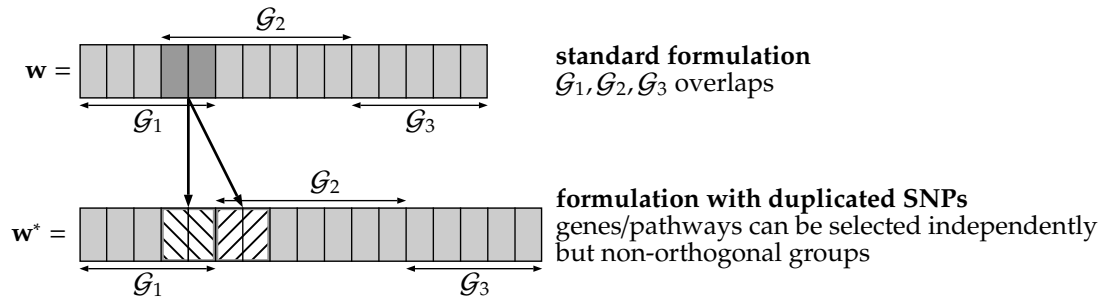


FIGURE 2.12: The problem of overlapping genes/pathways

2.3.2 Multiple Kernel Learning (MKL)

Multiple Kernel Learning On the contrary of the classical SVM, where there is a single kernel K , the Multiple kernel learning uses a set of L kernels K_1, \dots, K_L , and defines the kernel:

$$K(\mathbf{x}, \mathbf{x}^i) = \sum_{\ell=1}^L \theta_{\ell} K_{\ell}(\mathbf{x}, \mathbf{x}^i)$$

It learns the classifier and the combination weights are such that $\theta_{\ell} > 0$ and $\sum_{\ell=1}^L \theta_{\ell} = 1$. It is still a convex optimization problem, and the decision function is given by:

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i=1}^N w_i K(\mathbf{x}, \mathbf{x}_i) + b = \sum_{i=1}^N w_i \sum_{\ell=1}^L \theta_{\ell} K_{\ell}(\mathbf{x}, \mathbf{x}_i) + b \\ &= \sum_{\ell=1}^L \theta_{\ell} \sum_{i=1}^N w_i K_{\ell}(\mathbf{x}, \mathbf{x}_i) + b = \sum_{\ell=1}^L \sqrt{\theta_{\ell}} \langle \mathbf{w}_{\ell}, \psi_{\ell}(\mathbf{x}) \rangle + b \end{aligned}$$

where $\mathbf{w} = (\mathbf{w}_1^T, \dots, \mathbf{w}_L^T)^T$, $\psi_{\ell}(\mathbf{x}) : \mathbf{y} \mapsto K_{\ell}(\mathbf{x}, \mathbf{y})$ and $\mathbf{w}_L = \sqrt{\theta_{\ell}} \sum_{i=1}^N w_i \mathbf{x}_i$.

All the parameters w_i, θ_ℓ, b are learnt together. The multiple kernel learning can efficiently concatenate different modalities.

$$\min_{\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}, \theta \geq 0} \left\{ \frac{1}{N} \sum_{j=1}^N \mathcal{L}(f(\mathbf{x}^j), y^j) + \gamma \frac{1}{2} \sum_{\ell=1}^L \|\mathbf{w}_\ell\|^2 + \gamma' \Omega(\theta) \right\}$$

Equivalently, this problem is similar to, with $w_m \leftarrow \theta_m w_m$:

$$\min_{\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}, \theta \geq 0} \left\{ \frac{1}{N} \sum_{j=1}^N \mathcal{L} \left(\sum_{\ell=1}^L \langle \mathbf{w}_\ell, \psi_\ell(\mathbf{x}^j) \rangle + b, y^j \right) + \gamma \frac{1}{2} \sum_{\ell=1}^L \frac{\|\mathbf{w}_\ell\|^2}{\theta_m} + \gamma' \Omega(\theta) \right\}$$

There are several possible penalties, such as $\Omega(\theta) = \|\theta\|_1$ that will enforce sparse kernel mixtures; or $\Omega(\theta) = \|\theta\|_2$ for non-sparse solution.

ℓ_p -norm MKL In the ℓ_p -MKL, we set $\Omega(\theta) = \|\theta\|_p = \left(\sum_{m=1}^M \theta_m^p \right)^{\frac{1}{p}}$.

Application of ℓ_p -norm MKL to Alzheimer's Disease In [Peng et al., 2016], the authors propose to set one linear kernel per feature and they want to learn the weights for each coefficient using a regulation $\ell_{1,p}$. The advantage of that norm is that it provides sparsity inside groups but also understand that some features can be grouped together. The regularisation is particular; the ℓ_1 norm exploits the data structure while the $\ell_{1,2}$ norm provides sparsity between features.

They start by defining one kernel per modality:

$$K = \sum_{\ell=1}^L \sum_{m \in \mathcal{G}_\ell} \theta_m K_m \quad \text{with} \quad K_m(\mathbf{x}, \mathbf{x}') = x_m x'_m$$

Then, their decision function is given by:

$$f(\mathbf{x}) = \sum_{\ell=1}^L \sum_{m \in \mathcal{G}_\ell} \sqrt{\theta_m} \tilde{\mathbf{w}}_m^\top x_m + b$$

where the parameters $\tilde{\mathbf{w}}, b$ and θ are determined through the following minimization problem:

$$\min_{\theta} \min_{\tilde{\mathbf{w}}, b} \gamma \sum_{i=1}^N \mathcal{L}(f(\mathbf{x}^i), y^i) + \frac{1}{2} \sum_{\ell=1}^L \sum_{m \in \mathcal{G}_\ell} \frac{\|\tilde{\mathbf{w}}_m\|_2^2}{\theta_m} \quad \text{with} \quad \|\theta\|_{1,p} \leq \tau \quad \text{and} \quad 0 \leq \theta$$

The loss used is $\mathcal{L}(t, y) = (1 - yt)_+$. The $\|\cdot\|_{1,p}$ norm is defined by:

$$\|\theta\|_{1,p} = \left(\sum_{\ell=1}^L \gamma_\ell \left(\sum_{m \in \mathcal{G}_\ell} \beta_m |\theta_m| \right)^p \right)^{\frac{1}{p}}$$

The $\ell_{1,p}$ -regularisation provides sparsity inside groups and improves combination of groups. However, such partition is not optimal, as there could be interaction between modalities and within each modality.

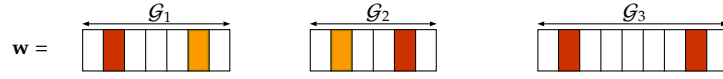


FIGURE 2.13: Parameter \mathbf{w} learnt using the $\ell_{1,p}$ -MKL

Their experiments has shown that the $\ell_{1,p}$ -MKL provides slightly better results for the classification tasks AD vs CN and MCI vs CN. The difference is an improved accuracy of 2%.

2.3.3 Structured sparse regularisation

In the same approach for combining different modalities and extension to multi-class learning – using a multiclass logistic regression –, we can also cite [Wang et al., 2012]. We consider the disease status CN = 1, MCI = 2, and AD = 3. They perform a classification from $\mathbf{x} = (\mathbf{x}_{\mathcal{G}_1}^\top, \dots, \mathbf{x}_{\mathcal{G}_L}^\top)^\top$, to predict $\mathbf{y} = (y_1, y_2, y_3)^\top$ where

$$y_k = \mathbb{P}\{Y = k | \mathbf{x}\} = \frac{e^{\mathbf{w}_{:,k}^\top \mathbf{x}}}{\sum_{k'=1}^3 e^{\mathbf{w}_{:,k'}^\top \mathbf{x}}}$$

They define the matrix $\mathbf{W} \in \mathbb{R}^{M \times 3}$ such that:

$$\mathbf{W} = \begin{pmatrix} \mathbf{w}_{\mathcal{G}_1,1} & \dots & \mathbf{w}_{\mathcal{G}_1,3} \\ \vdots & & \vdots \\ \mathbf{w}_{\mathcal{G}_L,1} & \dots & \mathbf{w}_{\mathcal{G}_L,3} \end{pmatrix} \quad \text{with } \mathbf{w}_{\mathcal{G}_\ell,p} \in \mathbb{R}^{d_\ell}$$

\mathbf{W} is determined by minimizing the quantity

$$\underbrace{\mathcal{L}(\mathbf{W})}_{\text{risk}} + \underbrace{\gamma_1 \|\mathbf{W}\|_{G1} + \gamma_2 \|\mathbf{W}\|_{2,1}}_{\text{regularization}}$$

where

$$\mathcal{L}(\mathbf{W}) = \sum_{i=1}^N \sum_{k=1}^3 \left(y_k^i \log \sum_{c=1}^3 e^{\mathbf{w}_{:,c}^\top \mathbf{x}^i} - y_k^i \mathbf{w}_{:,k}^\top \mathbf{x}^i \right)$$

Regarding to penalties, they propose to use the group ℓ_1 -norm, which provides sparsity between modalities and underlines global relationships between modalities, defined by:

$$\|\mathbf{W}\|_{G1} = \sum_{c=1}^3 \sum_{\ell=1}^L \|\mathbf{w}_{\mathcal{G}_\ell,c}\|_2$$

They also propose to use the $\ell_{2,1}$ -norm which emphasises sparsity between all features and non-sparsity between tasks, defined by:

$$\|W\|_{2,1} = \sum_{m=1}^M \|\mathbf{w}_{m,:}\|_2 \quad \text{where } \mathbf{w}_{m,:} = (\mathbf{w}_{m,1} \quad \mathbf{w}_{m,2} \quad \mathbf{w}_{m,3})$$

The authors of [Wang et al., 2012] have applied their model to data from ADNI, and using the genetic (SNPs), and neuroimaging (MRI and PET scans) modalities. They concluded that their model provided better performances overall, with an average accuracy of 72.6 ± 3.2 %; while the MKL method with ℓ_∞ penalty has an average accuracy of 62.4 ± 3.1 %.

2.4 Conclusion

This chapter provides a review of statistical and machine learning approaches for imaging genetics. One of the research area is to consider that genetics can explain neuroimaging data, but does not relate it with the disease status.

Another research area is to consider that genetics and neuroimaging data capture different but complementary features. The state of the art in this area usually focuses on using linear (or pseudo-linear) models with adapted penalties. Penalties are chosen in order to optimise the weight of each modality in the decision function, but also the optimise the weight of each coefficients within each modality. These penalties are compulsory, as we work with high dimensional data, and help to highlight features that are related to the disease. The question raised from this approach is: Are linear models adapted for the combinaison of modalities from different sources?

It is possible to apply these models to the more interesting classification tasks pMCI (progressive MCI at date T) vs sMCI (stable MCI at date T), with T that could be changed, in order to deduce at the end if the MCI patient has converted to AD, and if yes, when he has converted to AD. In the classification framework, this task will require to run different classifiers for different $T \in \{6, 12, 18, 24, 30, 36, 42, 48, \dots\}$. Furthermore, the path to conversion, using the classification framework, does not ensure that it is monotonous. In the next chapter, we will focus on other approaches for predicting the conversion date from MCI patients.

Chapter 3

Survival analysis: from theory to application for Alzheimer's Disease

3.1 Introduction

The early diagnosis of Alzheimer's disease (AD) is important for providing adequate care and is currently the topic of very active research. In particular, a current challenge is to predict the future occurrence of AD in patients with mild cognitive impairment.

A first way for solving this kind of problem is using classification at fixed time T . If we split the dataset of MCI patients into two subsets of having converted before time T , and having not converted before time T ; and applied the same classification algorithm at different time T , we will not use informations concerning the path of evolution throughout the time. In this sense, we do not ensure the conversion path to AD is monotonous.

Another way is to use regression using the framework of survival analysis. If we denote T be the random variable representing the time to event (the conversion date MCI to AD), we aim – in the framework of survival analysis – at modelling the survival function $S : t \mapsto \mathbb{P}(T \geq t)$.

In this chapter, we propose a review of survival analysis, and some applications using multimodal data to estimate the conversion date to AD from genetics and clinical data.

3.2 Background

The framework of survival analysis allow us to model the time until a pre-specified event from a given starting point, in our case the time of conversion to AD for MCI subjects. The starting point is the entry in the ADNI study, the event is the progression of disease to AD and the endpoint is the time to progression. This section is mainly based on [[van Houwelingen and Putter, 2012](#)].

3.2.1 Assumptions

In the framework of survival analysis,

- (i) the event might not be observed. Researchers have to put an end to the clinical studies at some point in time and we have to deal with (right) censored observations

of the people still MCI at the end of the study, for which only a lower bound for the survival time is known.

- (ii) the event might never happen. For instance, some MCI patients will never convert to AD.
- (iii) the event can only happen once.

3.2.2 Censoring and truncation

For each individual i , we denote T_i^* his (real) survival time (the conversion date MCI to AD) and C_i his censored time (the end of study) and $T_i = \min(T_i^*, C_i)$ be the duration observed in the study. The information available is the observed duration T_i and an indicator $\delta_i = \mathbb{I}\{X_i \leq C_i\}$ with

- $\delta_i = 1$ if the event is observed, and $T_i = T_i^*$ (real dates).
- $\delta_i = 0$ if the event is not observed, and $T_i = C_i$ (censored).

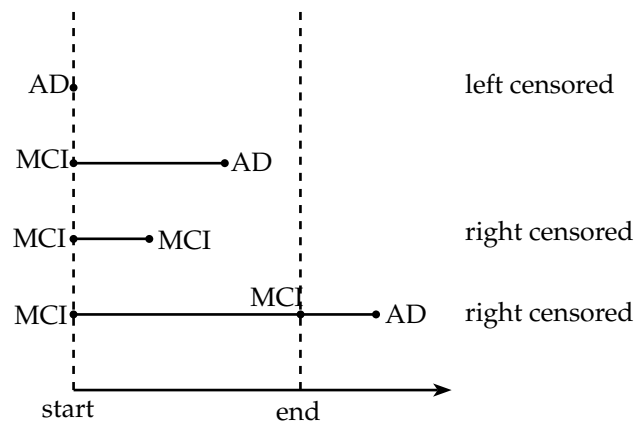


FIGURE 3.1: Different type of censorships

For instance, there is censoring when a subject leaves the study and does not come back or when the study ends whereas some subjects are still MCI. We also assume that T^* and C are independent.

On the contrary of censoring, truncation is deliberate and depends on the design of the study. If there is truncation, some subjects are not observable. Saying that the variable T^* is truncated by a (random) subset $\mathcal{A} \subset \mathbb{R}_+$ means that T^* is only observed if $T^* \in \mathcal{A}$. More specifically, let Z be a independent random variable of T^* .

- There is left truncation if T^* is observable if $T^* > Z$. We observe the couple (T^*, Z) with $T^* > Z$. Only individuals who survive a sufficient time are included in the sample.
- There is right truncation if T^* is observable if $T^* < Z$. We observe the couple (T^*, Z) with $T^* < Z$. Only individuals who have experienced the event by a specified time are included in the sample.

3.2.3 Survival and hazard function

The conversion date T is seen as a continuous random variable with probability density function (p.d.f.) f , and cumulative distribution function (c.d.f.) $F : t \mapsto \mathbb{P}\{T < t\}$. It gives the probability that the event has occurred before time t . However, it is convenient to work with the complement of the cumulative distribution function, the survival function:

$$S(t) = \mathbb{P}\{T \geq t\} = 1 - F(t) = \int_t^{\infty} f(u)du$$

S is a decreasing function, $S(0) = \mathbb{P}\{T \geq 0\} = 1$ and $\lim_{t \rightarrow \infty} S(t) = 0$.

It represents the probability that the conversion has not occurred by duration t . We also define the hazard function, or instantaneous rate of occurrence of the event, defined as

$$h(t) = \lim_{\delta t \rightarrow 0} \frac{\mathbb{P}\{t \leq T < t + \delta t | T \geq t\}}{\delta t}$$

The hazard function is the limit when δt goes to 0 of the ratio between the conditional probability that the event occurs in the interval $[t, t + \delta t]$ given that it did not occur before and δt . The hazard function can be rewritten

$$\begin{aligned} h(t) &= \lim_{\delta t \rightarrow 0} \frac{\mathbb{P}\{t \leq T < t + \delta t\}}{\delta t \times \mathbb{P}\{T \geq t\}} = \lim_{\delta t \rightarrow 0} \frac{\mathbb{P}\{t \leq T < t + \delta t\}}{\delta t} \times \frac{1}{\mathbb{P}\{T \geq t\}} \\ &= \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}(\log S(t)) \end{aligned}$$

Since $S(0) = 1$ (since the event is sure not to have occurred by duration 0), we can solve the above expression to obtain a formula for the probability of surviving to duration t as a function of the hazard at all durations up to t :

$$S(t) = \exp\left(-\int_0^t h(u)du\right)$$

The cumulative hazard H is defined as the sum of the risks you face going from duration 0 to t :

$$H(t) = \int_0^t h(u)du$$

A decreasing hazard function describes a phenomenon where the probability of becoming AD in a fixed time interval in the future decreases over time. On the contrary, an increasing hazard function describes a phenomenon where the probability of becoming AD in a fixed time interval in the future increases over time.

Median survival time The median survival time is calculated as the smallest survival time for which the survivor function is less than or equal to 0.5. The median survival time is considered at the date at which the subject converts to AD.

3.2.4 Kaplan-Meier and Nelson-Aalen estimators

The Kaplan-Meier estimator is a non-parametric statistic used to estimate the survival function from lifetime data. It measures the fraction of patients who have converted for a certain amount of time from the baseline. Let t_1, \dots, t_n be the times when there is conversion to AD or censorship. At time t_i , d_i is the number of conversions that happened, n_i the number of individuals who are still MCI (have not yet converted or have been censored) and $s_i = n_i - d_i$ the total number that haven't failed by time t_i (including censored at t_i). The Kaplan-Meier estimator is given by:

$$\widehat{S}(t) = \prod_{i: t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

From the Kaplan-Meier estimator of the survival function, the Breslow estimator of the cumulative hazard rate function can be obtained with:

$$\tilde{H}(t) = -\log(\widehat{S}(t)) = -\sum_{t_i \leq t} \log\left(1 - \frac{d_i}{n_i}\right)$$

Another non-parametric estimator of the cumulative hazard rate function is the Nelson-Aalen estimator, given by

$$\tilde{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

Both statistics estimate the cumulative number of expected events.

Example We consider the following dataset:

	Observed Survival Time T	Fail Indicator δ
1	12	0
2	24	0
3	18	1
4	6	1
5	24	1
6	30	1
7	24	0
8	24	0
9	12	1
10	12	0

We start by order the event times ($6 < 12 < 18 < 24 < 30$) and create the tabulate counts at each time:

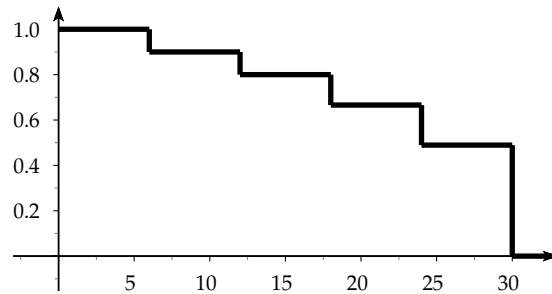
Date t_i	Individuals still MCI n_i	Number of conversion d_i	$s_i = n_i - d_i$
6	10	1	9
12	9	1	8
18	6	1	5
24	5	1	4
30	1	1	0

Then, we can estimate conditional survival probabilities:

$$\mathbb{P}\{T > t_j | T > t_{j-1}\} = 1 - \frac{d_j}{n_j}$$

For instance,

$$\begin{aligned} \mathbb{P}\{T > 0\} &= 1 \\ \mathbb{P}\{T > 6 | T > 0\} &= 1 - \frac{1}{10} = \frac{9}{10} \\ \mathbb{P}\{T > 12 | T > 6\} &= 1 - \frac{1}{9} = \frac{8}{9} \\ \mathbb{P}\{T > 18 | T > 12\} &= 1 - \frac{1}{6} = \frac{5}{6} \\ \mathbb{P}\{T > 24 | T > 18\} &= 1 - \frac{1}{5} = \frac{4}{5} \\ \mathbb{P}\{T > 30 | T > 24\} &= 1 - \frac{1}{1} = 0 \end{aligned}$$



For each $t \in]t_k, t_{k+1}[$, the survival function is given by:

$$S(t) = \mathbb{P}\{T > t\} = \mathbb{P}\{T > t | T > t_k\} \mathbb{P}\{T > t_k | T > t_{k-1}\} \dots \mathbb{P}\{T > t_1\}$$

3.2.5 Adding the covariates for individual predictions

To develop methods for survival data in a population of individuals we need a way of describing the variation among individuals. A popular model that fits in well is to consider the individual specific hazard function $h_i(t)$ and to make the proportional hazards assumption that

$$h_i(t) = c_i h_0(t)$$

The constant c_i , specific to each individual, is called hazard ratio.

3.2.5.1 Cox proportional hazard model

At time t , for an individual whose covariates are $\mathbf{x} = (x_1, \dots, x_n)$, the hazard function is given by:

$$h(t|\mathbf{x}) = h_0(t)e^{\mathbf{x}^\top \boldsymbol{\beta}}$$

In the model, h_0 is the baseline hazard function describing the risks for individuals with $\mathbf{x} = \mathbf{0}$. On the contrary, $e^{\mathbf{x}^\top \boldsymbol{\beta}}$ is a proportionate increase or reduction in risk, related to \mathbf{x} .

This model separates the effect of time from the effect of the covariates. Taking logs, the Cox model is an additive model for the log of the hazard:

$$\log h(t|\mathbf{x}) = \log h_0(t) + \mathbf{x}^\top \boldsymbol{\beta}$$

The effect of the covariates \mathbf{x} is the same at all times. If we integrate the hazard function over time, we obtain $H(t|\mathbf{x}) = H_0(t)e^{\mathbf{x}^\top \boldsymbol{\beta}}$. Then, the survival function is given by:

$$S(t|\mathbf{x}) = (S_0(t))^{e^{\mathbf{x}^\top \boldsymbol{\beta}}}$$

where $S_0(t) = \exp\left(-\int_0^t h_0(u)du\right)$ is a baseline survival function. The effect of the covariate values \mathbf{x} on the survivor function is to raise it to a power given by the relative risk $e^{\mathbf{x}^\top \boldsymbol{\beta}}$.

We also define the prognostic index as $PI(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$.

Example We consider a two-sample problem where the x serves to identify groups one (female) and zero (male). Then the model is

$$h(t|\mathbf{x}) = \begin{cases} h_0(t) & \text{if } x = 0 \\ h_0(t)e^\beta & \text{if } x = 1 \end{cases}$$

$h_0(t)$ represents the risk at time t in group 0, and e^β is the ratio of the risk in group one relative to group zero at any time t . If $e^\beta = 1$, the risk is the same for the two groups. On the contrary, if $e^\beta = 2$ (or $\beta = \ln 2$), then the risk for an individual in group one at any given age is twice the risk of a member of group zero who has the same age.

3.2.5.2 Exponential model

Different forms of proportional hazard models can be obtained by making assumptions about the baseline survival function, or the baseline hazard function. For instance, if the baseline risk is constant over time ($h(t) = \lambda$), then the survival function is $S(t|\mathbf{x}) = \exp(-\lambda t)$. λ can reparameterized in terms of predictor variables.

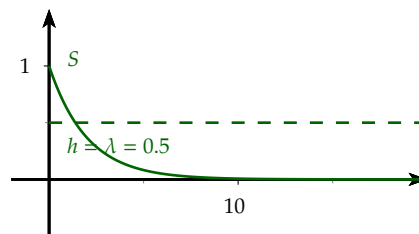


FIGURE 3.2: Survival and hazard functions for exponential model

3.2.5.3 Log-logistic model

The log-logistic survival function is defined by $S(t) = \frac{1}{1 + \lambda t^p}$. The hazard function is given by $h(t) = \frac{\lambda p t^{p-1}}{1 + \lambda t^p}$.

Usually, we set $\alpha = -\log \lambda$ and $p = \frac{1}{\sigma}$, and prefer to estimate α, σ rather than λ, p .

For individual predictions, λ is re-parameterised in terms of predictor variables and regression parameters, whereas the shape parameter p is usually held fixed. When $p > 1$, the hazard is unimodal. On the contrary, when $p < 1$ the hazard decreases.

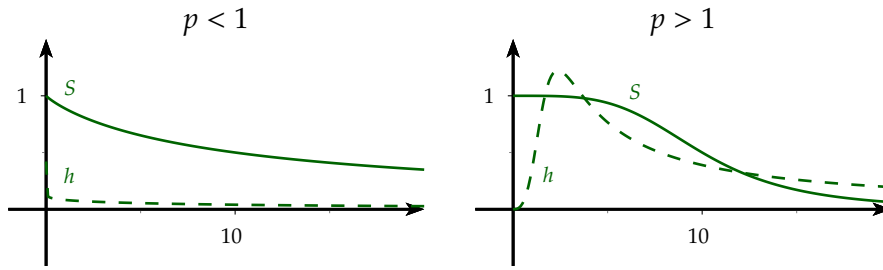


FIGURE 3.3: Survival and hazard functions for log-logistic model

3.2.6 Model fitting

3.2.6.1 Parametric models

When we assume a specific form for the hazard or survival function, the parameters are fit by maximising the appropriate likelihood function. Data are pairs (T_i, δ_i) for $i \in \{1, \dots, n\}$, as described earlier.

- If the subject i became AD at time T_i , then the contribution to the likelihood function is the density at that duration: $L_i = f(T_i) = S(T_i)h(T_i)$.
- If the subject i is still MCI at T_i , all we know under non-informative censoring is that the lifetime exceeds T_i . Therefore, the probability of this event is $L_i = S(T_i)$.

Then, the likelihood function is given by:

$$L = \prod_{i=1}^n f(T_i)^{\delta_i} S(T_i)^{1-\delta_i} = \prod_{i=1}^n h(T_i)^{\delta_i} S(T_i)$$

If we do not take censored data into account, the likelihood would be $L = \prod_{i=1}^n h(T_i)$, and the estimation of is biased.

Example In the case of the exponential model, the hazard is $h(t) = \lambda$.

$$\text{Then, } L = \prod_{i=1}^n \lambda^{\delta_i} e^{-\lambda T_i} \text{ and } \log L = \sum_{i=1}^n \delta_i \ln \lambda - \lambda T_i = D \ln \lambda - \lambda T.$$

where $D = \sum_{i=1}^n \delta_i$ is the total number of deaths, and $T = \sum_{i=1}^n T_i$ the total observation (or exposure) time. Differentiating $\log L$ and solving $-\log L = 0$ gives the maximum likelihood estimator of the hazard:

$$\widehat{\lambda} = \frac{D}{T}$$

3.2.6.2 Semi-parametric models

The hazard remains unspecified, and we only estimate the regression coefficients β (like in the Cox model). Data are of triples $(T_i, \delta_i, \mathbf{x}_i)$ for $i \in \{1, \dots, n\}$, as described earlier.

$$L = \prod_{i=1}^n h(T_i)^{\delta_i} S(T_i) = \prod_{i=1}^n h_0(T_i)^{\delta_i} e^{\delta_i \beta^\top \mathbf{x}_i} (S_0(T_i))^{e^{\beta^\top \mathbf{x}_i}}$$

The log-likelihood is:

$$\begin{aligned} \log L &= \sum_{i=1}^n \delta_i \log h_0(T_i) + \delta_i \beta^\top \mathbf{x}_i + e^{\beta^\top \mathbf{x}_i} \log (S_0(T_i)) \\ &= \sum_{i=1}^n \delta_i \log h_0(T_i) + \delta_i \beta^\top \mathbf{x}_i - e^{\beta^\top \mathbf{x}_i} H_0(T_i) \end{aligned}$$

$$\text{As } H_0(t) = \sum_{t_i \leq t} h_0(T_i),$$

$$\log L = \sum_{i=1}^n \left(-h_0(T_i) \sum_{j \in \mathcal{R}(T_i)} e^{\beta^\top \mathbf{x}_j} + \delta_i \log h_0(T_i) + \delta_i \beta^\top \mathbf{x}_i \right)$$

where $\mathcal{R}(T_i)$ is the risk set of people still alive and in follow-up just prior to T_i .

Let $\widehat{h}_0(T_i|\beta) = \frac{\delta_i}{\sum_{j \in \mathcal{R}(T_i)} e^{\beta^\top \mathbf{x}_j}}$. The resulting log-likelihood is:

$$\ell(\widehat{h}_0(\cdot|\beta), \beta) = \sum_{i=1}^n \delta_i (-1 + \ln \widehat{h}_0(T_i|\beta) + \beta^\top \mathbf{x}_i) = - \sum_{i=1}^n \delta_i + \text{pl}(\beta)$$

where $\text{pl}(\beta)$ is Cox partial log-likelihood:

$$\text{pl}(\beta) = \sum_{i=1}^n \delta_i \cdot \ln \frac{e^{\beta^\top \mathbf{x}_i}}{\sum_{j \in \mathcal{R}(t_i)} e^{\beta^\top \mathbf{x}_j}} = \sum_{i=1}^n \delta_i \left[\beta^\top \mathbf{x}_i - \log \sum_{j \in \mathcal{R}(t_i)} e^{\beta^\top \mathbf{x}_j} \right]$$

The partial likelihood depends only on the ordering of the survival times, not the actual values; so it is invariant to monotone transformation of time. β is estimated by maximising the partial log-likelihood, usually through a gradient descent.

Time is divided into small intervals and we assume that the baseline hazard is constant in each interval. For the Cox PH model, once β estimated, the survival function can be estimated

- by the Kaplan-Meier estimator:

$$\widehat{S}(t|\mathbf{x}) = \prod_{T_i \leq t} (1 - e^{\beta^\top \mathbf{x}} \widehat{h}_0(T_i)) \quad \text{where} \quad \widehat{h}_0(T_i) = \frac{\delta_i}{\sum_{j \in \mathcal{R}(T_i)} e^{\beta^\top \mathbf{x}_j}}$$

- or by the Nelson-Aalen estimator:

$$\widehat{S}(t|\mathbf{x}) = \exp(-\widehat{H}_0(t)e^{\beta^\top \mathbf{x}})$$

3.2.7 Hypotheses behind the Cox proportional hazard model

The Cox proportional hazard model $h(t|\mathbf{x}) = h_0(t)e^{\beta^\top \mathbf{x}}$ can fail to fit data. Some reasons can be:

- the functional form of the covariates: instead of incorporating x_j in the model, it could be interesting to incorporate x_j^2 ou $\log(x_j)$.
- the proportional hazard assumption is not verified: the coefficient β_j is not constant in time.
- the relationship between the predictor $\beta^\top \mathbf{x}$ and the hazard is not log-linear.

We describe how the proportional hazard assumption is verified. It is equivalent to test the null hypothesis:

$$H_0 : \beta_j(t) = \beta_j$$

$\beta_j(t)$ can defined as:

$$\beta_j(t) = \beta_j + \theta_j g_j(t)$$

where g_j is a predictable function of time. Testing H_0 is equivalent to testing:

$$H_0 : \theta_j = 0$$

We order the durations $0 \leq T_{(1)} \leq \dots \leq T_{(n)}$ and $\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(n)}$ the corresponding covariates.

At time $T_{(j)}$, the Schoenfeld residual for any subject $i \in \mathcal{R}(T_{(j)})$ is defined by:

$$\mathbf{x}_{(i)} - \mathbf{a}_{(j)} \quad \text{where} \quad \mathbf{a}_{(j)} = \sum_{\ell \in \mathcal{R}(T_{(j)})} \frac{e^{\beta^\top \mathbf{x}_\ell}}{\sum_{k \in \mathcal{R}(T_{(j)})} e^{\beta^\top \mathbf{x}_k}} \mathbf{x}_\ell$$

The Schoenfeld residual corresponding to $T_{(j)}$ is defined as the sum of the Schoenfeld residuals over all subjects who fail at $T_{(j)}$:

$$r_{\text{Sh}}(j) = \sum_{i \in \mathcal{R}(T_{(j)})} \delta_{ij} (\mathbf{x}_{(i)} - \mathbf{a}_{(j)})$$

where $\delta_{ij} = 1$ if the subject i fails at $T_{(j)}$, 0 otherwise.

The scaled Schoenfeld residual are defined by $\widehat{\mathbf{V}}_{(j)}^{-1} r_{\text{Sh}}(j)$ where $\widehat{\mathbf{V}}_{(j)}$ is the estimator of the covariance matrix of the Schoenfeld residual $r_{\text{Sh}}(j)$. Furthermore, $\mathbb{E}[\widehat{\mathbf{V}}_{(j)}^{-1} r_{\text{Sh}}(j)] \approx \theta_j g_j(T_{(j)})$.

We plot $\widehat{\mathbf{V}}_{(j)}^{-1}r_{\text{Sh}(j)} + \widehat{\beta}$ as a function of time $T_{(j)}$. The hypothesis H_0 is rejected if the graph does not look like an horizontal line.

3.3 Measuring the predictive value of a survival model

In this section, we describe how to assess the performance of a model. Issues to be considered are the model's discriminative ability and the prediction error of the model. The strength of the relation between a predictor and the survival outcome can be realised by AUC-type measures, whereas the prediction error can be measured by the Brier score.

3.3.1 Median survival time

We compare the median survival time of the log-logistic model to the median survival time given by the Kaplan-Meier estimator.

3.3.2 Concordance-index (C-index)

The C-index [Steck et al., 2008] checks if the model orders the conversion dates in the same order as the ground truth. It does not verify how close the estimated conversion date is to the ground truth conversion date.

Two subjects' survival times can be ordered if

1. both of them are uncensored
2. and the uncensored time of one is smaller than the censored survival time of the other.

This definition can be represented by an order graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where the vertices \mathcal{V} represents all the individuals, where each filled vertex indicates an observed/uncensored survival time, while an empty circle denotes a censored observation. Existence of an edge $\mathcal{E}_{i,j}$ implies that $T_i < T_j$. An edge cannot originate from a censored point.

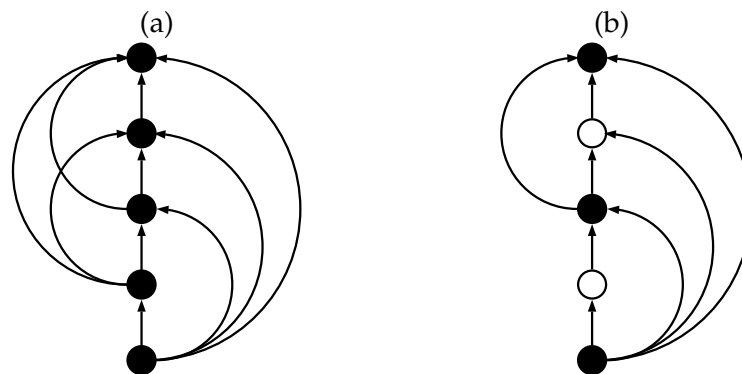


FIGURE 3.4: Order graphs representing the ranking constraints. (a) No censored data and (b) with censored data. The empty circle represents a censored point. The points are arranged in the increasing value of their survival times with the lowest being at the bottom (reproduced from [Steck et al., 2008]).

The C-index is the fraction of all pairs of subjects whose predicted survival times are correctly ordered among all subjects that can actually be ordered.

$$c = \frac{1}{|\mathcal{E}|} \sum_{\mathcal{E}_{ij}} \mathbf{1}_{f(x_i) < f(x_j)} = \frac{1}{|\mathcal{E}|} \sum_{T_i \text{ uncensored}} \sum_{T_j > T_i} \mathbf{1}_{f(x_i) < f(x_j)}$$

where $|\mathcal{E}|$ is the number of edges in the order graph, $f(x_i)$ is the predicted survival time for subject i by the model f .

Therefore, the C-index does not evaluate the distance between the estimated conversion date and the true conversion date. This measure evaluates the accuracy of the ordering of predicted time. As a generalisation of AUC:

- 0.5 is the expected result from random predictions,
- 1.0 is perfect concordance,
- 0.0 is perfect anti-concordance (multiply predictions with -1 to get 1.0)

For standard survival models, the C-index is between 0.6 and 0.7.

3.3.3 Cumulative AUC(t)

Let M_i be the estimated probability value for subject i , and $D_i(t) = 1$ if $T_i \leq t$ et $D_i(t) = 0$ if $T_i > t$. $D_i(t)$ denotes the failure (disease) status at any time t with $D_i(t) = 1$ indicating that subject i has had an event prior to time t .

Sensitivity and specificity become time-dependent functions:

$$\begin{aligned} \text{sensitivity}(m, t) &= \mathbb{P}\{M > m | D(t) = 1\} \\ \text{specificity}(m, t) &= \mathbb{P}\{M \leq m | D(t) = 0\} \end{aligned}$$

Each individual plays the role of a control for times $t < T_i$, but then contributes as a case for later times, $t \geq T_i$. Using these definitions, we can define the corresponding ROC curve for any time t , $\text{ROC}(t)$.

If T denotes the follow-up date, and $f(t)$ the probability density function of the time-to-event outcome related to the model, we define the integrated AUC [Chambless and Diao, 2006] by:

$$\text{iAUC} = \int_0^T \text{AUC}(t) \times f(t) dt$$

As for the AUC, the range for the iAUC is in $[0, 1]$.

3.3.4 Kullback-Leibler divergence and Brier score

If $\widehat{S}(t_0|\mathbf{x})$ denotes the model-based probabilistic prediction for the survival of an individual beyond t_0 given the covariates \mathbf{x} and $y = \mathbb{I}_{\{T > t_0\}}$ be the actual observation (ignoring censoring for the time being), we define the following scores [Graf et al., 1999]:

$$\begin{aligned}\text{KL}(y, \widehat{S}(t_0|\mathbf{x})) &= -y \log \widehat{S}(t_0|\mathbf{x}) - (1 - y) \log(1 - \widehat{S}(t_0|\mathbf{x})) \\ \text{Brier}(y, \widehat{S}(t_0|\mathbf{x})) &= (y - \widehat{S}(t_0|\mathbf{x}))^2\end{aligned}$$

From these measures, a global measure can be built. The time is divided into $L + 1$ intervals $I_1 = [t_0 = 0, t_1[, \dots, I_L = [t_{L-1}, t_L[, I_{L+1} = [t_L, +\infty[$, where t_L is the horizon time.

The global score is then given by:

$$\begin{aligned}\text{Brier}_{\text{global}}(t, \widehat{S}(t_0|\mathbf{x})) &= \sum_{\ell=1}^{L+1} (\mathbb{I}_{\{t \in I_\ell\}} - \widehat{p}_\ell(\mathbf{x}))^2 \\ \text{KL}_{\text{global}}(t, \widehat{p}(\mathbf{x})) &= - \sum_{\ell=1}^{L+1} \mathbb{I}_{\{t \in I_\ell\}} \ln(\widehat{p}_\ell(\mathbf{x}))\end{aligned}$$

where $\widehat{p}_\ell(\mathbf{x}) = \mathbb{P}(T \in I_\ell|\mathbf{x})$.

These scores evaluate how well the estimated survival function fits to the ground truth. Lower are these scores, better the model is.

3.4 Combinaison of multimodal data using the Cox PH framework

In this section, we describe some applications of Cox proportional hazards models using genetics, clinical and neuroimaging data. As for classification, the goal is to provide an answer to the question: Can genetics data help to improve predictions and how to combine them with clinical data?

In [van Houwelingen et al., 2006, van der Laan et al., 2007, Bøvelstad et al., 2009], they applied their models to the breast cancer dataset.

Notations We denote \mathbf{x}_G the vector of SNPs counted by number of minor variants, \mathbf{x}_C the clinical covariates, and T the conversion date from baseline (continuous random variable).

A first model for combining genetics and clinical data In [Bøvelstad et al., 2009], the authors propose the following model:

$$h(t|\mathbf{x}_G, \mathbf{x}_C) = h_0(t) \exp(\mathbf{x}_G^\top \beta + \mathbf{x}_C^\top \gamma)$$

Clinical covariates are supposed to be well-established and are all included in the model. The high-dimensional genomic part is regularised, by adding a penalty on β such as ℓ_1 or ℓ_2 penalty. According to the authors, only the ℓ_2 penalty can improve predictions. In particular, the correlation between genetics and clinical data can lead to dramatic changes in both β and γ compared with the models with only clinical or genomic covariates. For example, adding clinical covariates can completely change the selection of genomic covariates made by a ℓ_1 -regression. The implication is that the prediction model depends

completely on the selection of clinical covariates and cannot easily be transferred from one data set to another.

A second model for combining genetics and clinical data In [van der Laan et al., 2007], the authors propose to start by defining two separate models:

1. the clinical model: $h(t|\mathbf{x}_C) = h_0(t) \exp(\mathbf{x}_C^\top \gamma)$, with $\text{PI}_{\text{clin}}(\mathbf{x}_C) = \mathbf{x}_C^\top \gamma$.
2. the genomic model: $h(t|\mathbf{x}_G) = h_0(t) \exp(\mathbf{x}_G^\top \beta)$, with $\text{PI}_{\text{gen}}(\mathbf{x}_G) = \mathbf{x}_G^\top \beta$.

The super model is defined by

$$h(t|\mathbf{x}_G, \mathbf{x}_C) = h_0(t) \exp(\alpha_1 \text{PI}_{\text{clin}}(\mathbf{x}_C) + \alpha_2 \text{PI}_{\text{gen}}(\mathbf{x}_G))$$

For their breast cancer dataset, this model provides slightly better results than the individual model. The information that is shared by the genomic source and the clinical one and that is responsible for the correlation of the indices is much more relevant than the independent parts. Furthermore, the authors notice that for short term prediction the clinical information is more important, while the genomic information reduces the prediction error in the long term.

Applications to Alzheimer's Disease The authors propose to use the Cox proportional hazard framework with different features has been illustrated in several papers such as [Li et al., 2017, Anderson et al., 2016, Liu et al., 2017].

In [Li et al., 2017], they used the Cox proportional hazard framework with several features. A first model is to use baseline information of some markers (age at baseline, gender, APOE status, education) and cognitive scores (such as ADAS-Cog13, RAVLT (imme), RAVLT (learn), FAQ and MMSE). They also propose to add biological features such as the hippocampus volume, the mid temporal volume, and the FDG.

A second model is to use longitudinal data, by computing 12 MFPC scores from MFPCA (Multivariate Functional Principal Component Analysis), calculated from longitudinal information of cognitive scores before the visit of AD diagnosis or censoring.

They performed an internal validation using repeated a 10-fold cross-validation, and an external validation where ADNI1 is used for parameter estimations, whereas performance measures are assed on ADNI2. They concluded that history of multivariate longitudinal markers can greatly enhance prognostic performance. Furthermore, cognitive and functional variables are the most predictive variables, whereas imaging biomarkers and baseline CSF markers do not improve the predictive power of the model. They also mentioned that their model may produce biased parameter estimates of the predictor effects when failing to account for the competing risk of death in elder population.

In [Liu et al., 2017], the authors also applied the Cox proportional hazard framework using neuroimaging data (structural MRI, FDG-PET features), and clinical variables. They started by applying an independent component analysis on neuroimaging data to extract brain networks in AD and CN groups. Then, they computed independent variates of MCI baseline neuroimaging data. They performed several Cox models using different groups

of features. Although clinical variables provide better accuracy and AUC results at 36 months than neuroimaging data, the combination of both modalities increase the accuracy and AUC. They noticed that patients with reduced gray matter volume in a temporal lobe-related network based on MRI, low glucose metabolism in the posterior DMN based on FDG-PET, positive APOE ϵ_4 status, increased ADAS-Cog and CDR-SB scores were more likely to convert to AD within 36 months.

In [Anderson et al., 2016], the intra-individual cognitive variability (IICV), which estimates the variability between cognitive domains measured at one-time point, is computed as one single feature from four cognitive scores, including RAVLT, total of learning trials, the American National Adult Reading Test and Trail Making Test. A Cox model was used with APOE status, age, education, hippocampus volume loss and IICV. The study showed that the IICV is a significant feature when it comes to predict the conversion date to AD.

3.5 Illness-death models

In [Satizabal et al., 2016], the authors studied the conversion to dementia over three decades of cohorts in the Framingham Heart Study. Their dataset was composed of four cohorts representing four different periods. After applying the Cox PH model on each cohort, they concluded that the incidence of dementia has declined over the course of three decades but the factors contributing to this decline have not been completely identified. The main problem in their methodology, as raised in [Binder and Schumacher, 2016] is that the death event is not independent of the conversion date to dementia. Another problem is that their dataset is not homogeneously built.

In [Leffondré et al., 2013], the authors introduced a multi-state model from modelling an illness-death model. Three states are defined in figure 3.5: disease free (state 0), diseased (state 1), dead (state 2). Transition intensities $\alpha_{k\ell}(t|\mathbf{x})$ from state k to ℓ at age t for covariates \mathbf{x} are defined by:

$$\alpha_{k\ell}(t|\mathbf{x}) = \alpha_{k\ell,0}(t)e^{\beta_{k\ell}^T \mathbf{x}}$$

The exponential term $e^{\beta_{k\ell}^T \mathbf{x}}$ can be interpreted as a hazard ratio (HR) as in the Cox PH model. It is a Markov model where each transition intensity depends on age only.

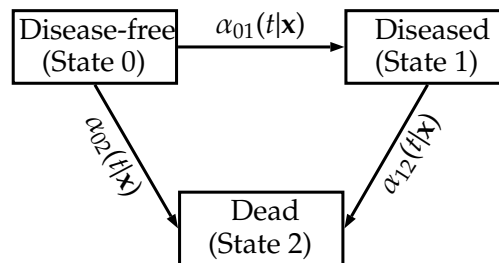


FIGURE 3.5: Illness-death model (reproduced from [Joly et al., 2002], [Leffondré et al., 2013], [Yu et al., 2010])

To understand the bias [Joly et al., 2002], we consider two dates 0 and t_1 and we are interested in modelling $\alpha_{01,0}$ using an exponential model (the intensity functions $\alpha_{k\ell,0}$

are constant). In the classical two-state model, subjects who died between 0 and t_1 are right-censored at 0. An estimator of the survival function at t_1 is:

$$S(t_1) = \frac{\mathbb{P}\{\text{healthy at } t_1\}}{\mathbb{P}\{\text{alive at } t_1\}}$$

Because $S(t_1) = e^{-\alpha_{01,0}t_1}$, an estimator of $\alpha_{01,0}$ is $\alpha_{01,0} = -\frac{1}{t_1} \log S(t_1)$. However, in an illness-death model,

$$\tilde{S}(t_1) = \frac{\mathbb{P}\{\text{healthy at } t_1\}}{\mathbb{P}\{\text{alive at } t_1\}} = \frac{1}{1 + \frac{\alpha_{01,0}}{\alpha_{01,0} + \alpha_{02,0} - \alpha_{12,0}} (e^{(\alpha_{01,0} + \alpha_{02,0} - \alpha_{12,0})t_1} - 1)}$$

Then,

$$\tilde{\alpha}_{01,0} - \alpha_{01,0} = -\frac{1}{t_1} \log \tilde{S}(t_1) + \frac{1}{t_1} \log S(t_1) = -\frac{1}{t_1} \log \left(\frac{\alpha_{01,0} e^{-(\alpha_{12,0} - \alpha_{02,0})t_1} - (\alpha_{12,0} - \alpha_{02,0}) e^{-\alpha_{01,0}t_1}}{\alpha_{01,0} - (\alpha_{12,0} - \alpha_{02,0})} \right)$$

If $\alpha_{12,0} = \alpha_{02,0} = 0$, then both models are identical.

In the general case, the estimation of $\alpha_{k\ell,0}$ [Joly et al., 2002] uses a penalised log-likelihood defined as:

$$\text{pl}(\alpha_{01,0}, \alpha_{12,0}, \alpha_{02,0}) = \ell(\alpha_{01,0}, \alpha_{12,0}, \alpha_{02,0}) - \lambda_{01} \int \alpha''_{01,0}(u)^2 du - \lambda_{12} \int \alpha''_{12,0}(u)^2 du - \lambda_{02} \int \alpha''_{02,0}(u)^2 du$$

where ℓ is the full-likelihood of the illness-death model, and $\lambda_{01}, \lambda_{12}, \lambda_{02}$ are positive smoothing parameters.

Compared with other survival models such as the Cox proportional hazard model and the Weibull model [Leffondré et al., 2013], the illness-death model provides better estimates of the effects on disease of the effects on disease of exposures that were associated with death. This model is able to account for the probability of developing the disease between the last visit and death, on the contrary of classical survival model.

3.6 Conclusion

This chapter provides a review of survival analysis and its applications for disease conversion date prediction. Although survival analysis make strong assumptions concerning the construction of the dataset (censoring and truncation) and the disease (all patients are assumed to convert), survival analysis also provides a good regression framework for directly estimating the conversion date to disease. We also note that most survival models have underlying assumptions that are important to verify when applied to specific dataset; and that death is a competitive risk when dealing with elder population.

Part II

Contributions

Chapter 4

Multilevel Modeling with Structured Penalties for Classification from Imaging Genetics Data

A preliminary version of this chapter has been published in the proceedings of the MICCAI 2017 Workshop on Imaging Genetics.

Pascal Lu, Olivier Colliot. Multilevel Modeling with Structured Penalties for Classification from Imaging Genetics data. 3rd MICCAI Workshop on Imaging Genetics (MICGen 2017), Sep 2017, Quebec City, Canada. pp.230-240

4.1 Introduction

In this chapter, we propose a framework for automatic classification of patients from multimodal genetic and brain imaging data by optimally combining them. Additive models with unadapted penalties (such as the classical group lasso penalty or ℓ_1 -multiple kernel learning) treat all modalities in the same manner and can result in undesirable elimination of specific modalities when their contributions are unbalanced.

To overcome this limitation, we introduce a multilevel model that combines imaging and genetics and that considers joint effects between these two modalities for diagnosis prediction. Furthermore, we propose a framework allowing to combine several penalties taking into account the structure of the different types of data, such as a group lasso penalty over the genetic modality and a ℓ_2 -penalty on imaging modalities. Finally, we propose a fast optimization algorithm, based on a proximal gradient method.

The model has been evaluated on genetic (single nucleotide polymorphisms - SNP) and imaging (anatomical MRI measures) data from the ADNI database, and compared to additive models [Wang et al., 2012, Aiolli and Donini, 2015]. It exhibits good performances in AD diagnosis; and at the same time, reveals relationships between genes, brain regions and the disease status.

4.2 State of the art

The research area of imaging genetics studies the association between genetic and brain imaging data [Liu and Calhoun, 2014]. A large number of papers studied the relationship between genetic and neuroimaging data by considering that a phenotype can be explained by a sum of effects from genetic variants. These multivariate approaches use partial least squares [Lorenzi et al., 2016], sparse canonical correlation analysis [Du et al., 2014], sparse regularized linear regression with a ℓ_1 -penalty [Kohannim et al., 2012a], group lasso penalty [Silver et al., 2012, Silver and Montana, 2012], or Bayesian model that links genetic variants to imaging regions and imaging regions to the disease status [Batmanghelich et al., 2016].

But another interesting problem is about combining genetic and neuroimaging data for automatic classification of patients. In particular, machine learning methods have been used to build predictors for heterogeneous data, coming from different modalities for brain disease diagnosis, such as Alzheimer's disease (AD) diagnosis. However, challenging issues are high-dimensional data, small number of observations, the heterogeneous nature of data, and the weight for each modality.

A framework that is commonly used to combine heterogeneous data is multiple kernel learning (MKL) [Gönen and Alpaydn, 2011]. In MKL, each modality is represented by a kernel (usually a linear kernel). The decision function and weights for the kernel are simultaneously learnt. Moreover, the group lasso [Ming and Yi, 2006, Meier et al., 2008] is a way to integrate structure inside data. However, the standard ℓ_1 -MKL and group lasso may eliminate modalities that have a weak contribution. In particular, for AD, imaging data already provides good results for its diagnosis. To overcome this problem, different papers have proposed to use a $\ell_{1,p}$ -penalty [Kloft et al., 2011] to combine optimally different modalities [Wang et al., 2012, Peng et al., 2016].

These approaches do not consider potential effects between genetic and imaging data for diagnosis prediction, as they only capture brain regions and SNPs separately taken. Moreover, they put on the same level genetic and imaging data, although these data do not provide the same type of information: given only APOE genotyping, subjects can be classified according to their risk to develop AD in the future; on the contrary, imaging data provides a photography of the subject's state at the present time.

Thereby, we propose a new framework that makes hierarchical the parameters and considers interactions between genetic and imaging data for AD diagnosis. We started with the idea that learning AD diagnosis from imaging data already provides good results. Then, we considered that the decision function parameters learnt from imaging data could be modulated, depending on each subject's genetic data. In other words, genes would express themselves through these parameters. Considering a linear regression that links these parameters and the genetic data, it leads to a multilevel model between imaging and genetics. Our method also proposes potential relations between genetic and imaging variables, if both of them are simultaneously related to AD. This approach is different from the modeling proposed by [Batmanghelich et al., 2016], where imaging variables are predicted from genetic variables, and diagnosis is predicted from imaging variables.

Furthermore, current approaches [Wang et al., 2012, Peng et al., 2016, Aiolli and Donini, 2015]

do not exploit data structure inside each modality, as it is logical to group SNPs by genes, to expect sparsity between genes (all genes are not linked to AD) and to enforce a smooth regularization over brain regions for imaging modality. Thus, we have imposed specific penalties for each modality by using a ℓ_2 -penalty on the imaging modality, and a group lasso penalty over the genetic modality. It models the mapping of variants into genes, providing a better understanding of the role of genes in AD.

To learn all the decision function parameters, a fast optimization algorithm, based on a proximal gradient method, has been developed. Finally, we have evaluated our model on 1,107 genetic (SNP) and 114 imaging (anatomical MRI measures) variables from the ADNI database and compared it to additive models [Wang et al., 2012, Aioli and Donini, 2015].

4.3 Model set-up

4.3.1 Multilevel Logistic Regression with Structured Penalties

Let $\{(\mathbf{x}_G^k, \mathbf{x}_I^k, y^k), k = 1, \dots, N\}$ be a set of labeled data, with $\mathbf{x}_G^k \in \mathbb{R}^{|\mathcal{G}|}$ (genetic data), and $\mathbf{x}_I^k \in \mathbb{R}^{|\mathcal{I}|}$ (imaging data) and $y^k \in \{0, 1\}$ (diagnosis). Genetic, imaging and genetic-imaging cross products training data are assumed centered and normalized.

We propose the following Multilevel Logistic Regression model:

$$p(y = 1 | \mathbf{x}_G, \mathbf{x}_I) = \sigma(\boldsymbol{\alpha}(\mathbf{x}_G)^\top \mathbf{x}_I + \alpha_0(\mathbf{x}_G)) \quad \text{with } \sigma : x \mapsto \frac{1}{1 + e^{-x}}$$

where $\alpha_0(\mathbf{x}_G)$ is the intercept and $\boldsymbol{\alpha}(\mathbf{x}_G) \in \mathbb{R}^{|\mathcal{I}|}$ is the parameter vector. On the contrary of the classical logistic regression model, we propose a multilevel model, for which the parameter vector $\boldsymbol{\alpha}(\mathbf{x}_G)$ and the intercept $\alpha_0(\mathbf{x}_G)$ depend on genetic data \mathbf{x}_G .

This is to be compared to an additive model, where the diagnosis is directly deduced from genetic and imaging data put at the same level. We assume that $\boldsymbol{\alpha}$ and α_0 are affine functions of genetic data \mathbf{x}_G :

$$\boldsymbol{\alpha}(\mathbf{x}_G) = \mathbf{W}\mathbf{x}_G + \boldsymbol{\beta}_I \quad \text{and} \quad \alpha_0(\mathbf{x}_G) = \boldsymbol{\beta}_G^\top \mathbf{x}_G + \beta_0$$

where $\mathbf{W} \in \mathcal{M}_{|\mathcal{I}|, |\mathcal{G}|}(\mathbb{R})$, $\boldsymbol{\beta}_I \in \mathbb{R}^{|\mathcal{I}|}$, $\boldsymbol{\beta}_G \in \mathbb{R}^{|\mathcal{G}|}$ and $\beta_0 \in \mathbb{R}$. Therefore, the probability becomes $p(y = 1 | \mathbf{x}_G, \mathbf{x}_I) = \sigma(\mathbf{x}_G^\top \mathbf{W}^\top \mathbf{x}_I + \boldsymbol{\beta}_I^\top \mathbf{x}_I + \boldsymbol{\beta}_G^\top \mathbf{x}_G + \beta_0)$. Figure 4.1 summarizes the relations between parameters.

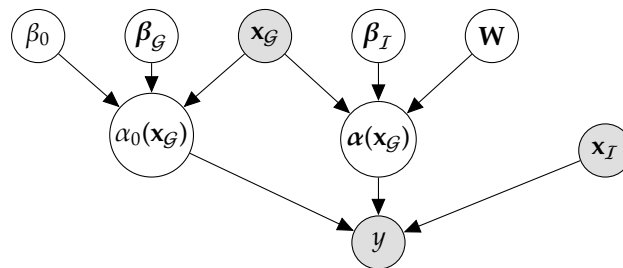


FIGURE 4.1: The disease status y is predicted from imaging data \mathbf{x}_I and the parameters $\beta_0(\mathbf{x}_G), \boldsymbol{\beta}(\mathbf{x}_G)$ (which are computed from genetic data \mathbf{x}_G)

The parameters $\mathbf{W}, \beta_I, \beta_G, \beta_0$ are obtained by minimizing the objective:

$$S(\mathbf{W}, \beta_I, \beta_G, \beta_0) = R_N(\mathbf{W}, \beta_I, \beta_G, \beta_0) + \Omega(\mathbf{W}, \beta_I, \beta_G)$$

$$\text{with } R_N(\mathbf{W}, \beta_I, \beta_G, \beta_0) = \frac{1}{N} \sum_{k=1}^N \left\{ -y^k \left((\mathbf{x}_G^k)^\top \mathbf{W}^\top \mathbf{x}_I^k + \beta_I^\top \mathbf{x}_I^k + \beta_G^\top \mathbf{x}_G^k + \beta_0 \right) \right. \\ \left. + \log \left(1 + e^{(\mathbf{x}_G^k)^\top \mathbf{W}^\top \mathbf{x}_I^k + \beta_I^\top \mathbf{x}_I^k + \beta_G^\top \mathbf{x}_G^k + \beta_0} \right) \right\}$$

$$\text{and } \Omega(\mathbf{W}, \beta_I, \beta_G) = \lambda_W \Omega_W(\mathbf{W}) + \lambda_I \Omega_I(\beta_I) + \lambda_G \Omega_G(\beta_G)$$

$\Omega_W, \Omega_I, \Omega_G$ are respectively the penalties for $\mathbf{W}, \beta_I, \beta_G$, whereas $\lambda_W > 0, \lambda_I > 0, \lambda_G > 0$ are respectively the regularization parameters for $\Omega_W, \Omega_I, \Omega_G$.

Genetic data are a sequence of single-polymorphism nucleotides (SNP) counted by minor allele. A SNP can belong (or not) to one gene ℓ (or more) and therefore participate in the production of proteins that interact inside pathways. We decided to group SNPs by genes, and designed a penalty to enforce sparsity between genes and regularity inside genes. Given that some SNPs may belong to multiple genes, the group lasso with overlap penalty [Jacob et al., 2009] is more suitable, with genes as groups. To deal with this penalty, an overlap expansion is performed. Given $\mathbf{x} \in \mathbb{R}^{|\mathcal{G}|}$ a subject's feature vector, a new feature vector is created $\tilde{\mathbf{x}} = (\mathbf{x}_{\mathcal{G}_1}^\top, \dots, \mathbf{x}_{\mathcal{G}_L}^\top)^\top \in \mathbb{R}^{\sum_{\ell=1}^L |\mathcal{G}_\ell|}$, defined by the concatenation of copies of the genetic data restricted by group \mathcal{G}_ℓ . Similarly, the same expansion is performed on β_G, \mathbf{W} to obtain $\tilde{\beta}_G \in \mathbb{R}^{\sum_{\ell=1}^L |\mathcal{G}_\ell|}$ and $\tilde{\mathbf{W}} \in \mathbb{R}^{|\mathcal{I}| \times (\sum_{\ell=1}^L |\mathcal{G}_\ell|)}$. This group lasso with overlap penalty is used for the matrix \mathbf{W} and for β_G .

For imaging variables, the ridge penalty is considered: $\Omega_I(\beta_I) = \|\beta_I\|_2^2$. In particular, brain diseases usually have a diffuse anatomical pattern of alteration throughout the brain and therefore, regularity is usually required for the imaging parameter. Finally, Ω is defined by:

$$\Omega(\tilde{\mathbf{W}}, \tilde{\beta}_G, \beta_I) = \lambda_W \sum_{i=1}^{|\mathcal{I}|} \sum_{\ell=1}^L \theta_{\mathcal{G}_\ell} \|\tilde{\mathbf{W}}_{i, \mathcal{G}_\ell}\|_2 + \lambda_I \|\beta_I\|_2 + \lambda_G \sum_{\ell=1}^L \theta_{\mathcal{G}_\ell} \|\tilde{\beta}_{\mathcal{G}_\ell}\|_2$$

4.3.2 Minimization of $S(\mathbf{W}, \beta_I, \beta_G, \beta_0)$

From now on, and for simplicity reasons, $\tilde{\mathbf{W}}, \tilde{\beta}$ and $\tilde{\mathbf{x}}$ are respectively denoted as \mathbf{W}, β and \mathbf{x} . Let Φ be the function that reshapes a matrix of $\mathcal{M}_{|\mathcal{I}|, |\mathcal{G}|}(\mathbb{R})$ to a vector of $\mathbb{R}^{|\mathcal{I}| \times |\mathcal{G}|}$ (i.e. $\mathbf{W}_{i, g} = \Phi(\mathbf{W})_{i|\mathcal{G}|+g}$):

$$\Phi : \mathbf{W} \mapsto ((\mathbf{W}_{1,1}, \dots, \mathbf{W}_{1,|\mathcal{G}|}), \dots, (\mathbf{W}_{|\mathcal{I}|,1}, \dots, \mathbf{W}_{|\mathcal{I}|,|\mathcal{G}|}))$$

We will estimate $\Phi(\mathbf{W})$ and then reshape it to obtain \mathbf{W} . The algorithm developed is based on a proximal gradient method [Hastie et al., 2015, Beck and Teboulle, 2009].

Algorithm 1: Training the multilevel logistic regression

1 **Input:** $\{(\mathbf{x}_I^k, \mathbf{x}_G^k, y^k), k = 1, \dots, N\}$, $\delta = 0.8$, $\varepsilon_0 = 1$, $\eta = 10^{-5}$;

2 **Initialization:** $\mathbf{W} = \mathbf{0}$, $\beta_I = \mathbf{0}$, $\beta_G = \mathbf{0}$, $\beta_0 = 0$, $\varepsilon = \varepsilon_0$ and converged = False ;

3 **while** not(converged) **do**

4 Compute $R_N = R_N(\mathbf{W}, \beta_I, \beta_G, \beta_0)$;

5 Compute $\nabla R_N = \frac{1}{N} \sum_{k=1}^N \begin{pmatrix} \Phi((\mathbf{x}_I^k)^\top \mathbf{x}_G^k) \\ \mathbf{x}_I^k \\ \mathbf{x}_G^k \\ 1 \end{pmatrix} \left[\sigma((\mathbf{x}_G^k)^\top \mathbf{W}^\top \mathbf{x}_I^k + \beta_I^\top \mathbf{x}_I^k + \beta_G^\top \mathbf{x}_G^k + \beta_0) - y^k \right]$

6 Compute $\omega = \beta - \varepsilon \nabla_{(\mathbf{W}, \beta_I, \beta_G)} R_N$;

7 Update $\widehat{\mathbf{W}}_{G_\ell, i} = \max\left(0, 1 - \frac{\varepsilon \lambda_{G_\ell} \theta_{G_\ell}}{\|\omega_{G_\ell + i|G_\ell}\|_2}\right) \omega_{G_\ell + i|G_\ell}$ for $(i, \ell) \in \llbracket 1, |I| \rrbracket \times \llbracket 1, L \rrbracket$;

8 Update $\widehat{\beta}_I = \frac{\omega_{I+|G|I}}{1+2\varepsilon\lambda_I}$ for $i \in \llbracket 1, |I| \rrbracket$ (imaging modality) ;

9 Update $\widehat{\beta}_{G_\ell} = \max\left(0, 1 - \frac{\varepsilon \lambda_{G_\ell} \theta_{G_\ell}}{\|\omega_{G_\ell + (|G|+1)|I}\|_2}\right) \omega_{G_\ell + (|G|+1)|I}$ for $\ell \in \llbracket 1, L \rrbracket$;

10 Update $\widehat{\beta}_0 = \beta_0 - \varepsilon \frac{\partial R_N}{\partial \beta_0}$ and $\widehat{G} = \frac{1}{\varepsilon} \left[\begin{pmatrix} \Phi(\mathbf{W}) \\ \beta_I \\ \beta_G \\ \beta_0 \end{pmatrix} - \begin{pmatrix} \Phi(\widehat{\mathbf{W}}) \\ \widehat{\beta}_I \\ \widehat{\beta}_G \\ \widehat{\beta}_0 \end{pmatrix} \right]$;

11 **if** $R_N(\widehat{\mathbf{W}}, \widehat{\beta}_I, \widehat{\beta}_G, \widehat{\beta}_0) > R_N - \varepsilon \nabla R_N^\top \widehat{G} + \frac{\varepsilon}{2} \|\widehat{G}\|_2^2$ **then**

12 | $\varepsilon = \delta \varepsilon$;

13 **else**

14 | converged = $\left| S(\widehat{\mathbf{W}}, \widehat{\beta}_I, \widehat{\beta}_G, \widehat{\beta}_0) - S(\mathbf{W}, \beta_I, \beta_G, \beta_0) \right| < \eta \left| S(\mathbf{W}, \beta_I, \beta_G, \beta_0) \right|$;

15 | $\mathbf{W} = \widehat{\mathbf{W}}$, $\beta_I = \widehat{\beta}_I$, $\beta_G = \widehat{\beta}_G$, $\beta_0 = \widehat{\beta}_0$, $\varepsilon = \varepsilon_0$;

16 **end**

17 **end**

18 **return** $(\mathbf{W}, \beta_I, \beta_G, \beta_0)$

The parameters $\mathbf{w}^{(t+1)} = (\Phi(\mathbf{W}^{(t+1)}), \beta_I^{(t+1)}, \beta_G^{(t+1)}, \beta_0^{(t+1)})$ are updated with:

$$\begin{aligned} \mathbf{w}^{(t+1)} &= \underset{\mathbf{w}}{\operatorname{argmin}} R_N(\mathbf{w}) + [\mathbf{w} - \mathbf{w}^{(t)}]^\top \nabla R_N(\mathbf{w}^{(t)}) + \frac{1}{2\varepsilon} \|\mathbf{w} - \mathbf{w}^{(t)}\|_2^2 + \Omega(\mathbf{w}) \\ &= \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\omega^{(t)} - \mathbf{w}^{(t)}\|_2^2 + \varepsilon \Omega(\mathbf{w}) \right\} \text{ with } \omega^{(t)} = \mathbf{w}^{(t)} - \varepsilon \nabla R_N(\mathbf{w}^{(t)}) \end{aligned}$$

The idea is to update $\mathbf{w}^{(t+1)}$ from $\mathbf{w}^{(t)}$ with a Newton-type algorithm without the constraint Ω given a stepsize ε , and then to project the result onto the compact set defined by Ω . Regarding the stepsize ε , a backtracking line search [Beck and Teboulle, 2009] is performed. Let $\widehat{G}(\mathbf{w}^{(t)}, \varepsilon) = \frac{1}{\varepsilon} [\mathbf{w}^{(t)} - \mathbf{w}^{(t+1)}]$ be the step in the proximal gradient update. A line search is performed over ε until the inequality is reached:

$$R_N(\mathbf{w}^{(t+1)}) \leq R_N(\mathbf{w}^{(t)}) - \varepsilon \nabla R_N(\mathbf{w}^{(t)})^\top \widehat{G}(\mathbf{w}^{(t)}, \varepsilon) + \frac{\varepsilon}{2} \|\widehat{G}(\mathbf{w}^{(t)}, \varepsilon)\|_2^2$$

The minimization algorithm stops when $|S(\mathbf{w}^{(t+1)}) - S(\mathbf{w}^{(t)})| \leq \eta |S(\mathbf{w}^{(t)})|$, where $\eta = 10^{-5}$. The whole algorithm is summarized below:

4.3.3 Probabilistic formulation

We provide a probabilistic formulation for the model, although it has not been implemented in practise. The conditional probability is given by

$$p(y = 1 | \mathbf{x}_{\mathcal{G}}, \mathbf{x}_I) = \sigma(\mathbf{x}_{\mathcal{G}}^T \mathbf{W}^T \mathbf{x}_I + \boldsymbol{\beta}_I^T \mathbf{x}_I + \boldsymbol{\beta}_{\mathcal{G}}^T \mathbf{x}_{\mathcal{G}} + \beta_0)$$

- For each region $i \in \mathcal{I}$ and gene \mathcal{G}_ℓ , $\mathbf{W}_{i, \mathcal{G}_\ell} \sim \text{M-Laplace}(0, \lambda_W)$ (M-Laplace stands for "Multi-Laplacian prior"). In other words:

$$p(\mathbf{W}; \lambda_W, \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}}) \propto \prod_{i=1}^{|\mathcal{I}|} \prod_{\ell=1}^L e^{-\lambda_W \theta_{\mathcal{G}_\ell} \|\mathbf{W}_{i, \mathcal{G}_\ell}\|_2}$$

- For each region $i \in \mathcal{I}$, $\boldsymbol{\beta}_i \sim \mathcal{N}(0, \frac{1}{2\lambda_I})$, i.e. $p(\boldsymbol{\beta}_I; \lambda_I) \propto e^{-\lambda_I \|\boldsymbol{\beta}_I\|_2^2}$
- For each gene \mathcal{G}_ℓ , $\boldsymbol{\beta}_{\mathcal{G}_\ell} \sim \text{M-Laplace}(0, \lambda_{\mathcal{G}})$, i.e.

$$p(\boldsymbol{\beta}_{\mathcal{G}}; \lambda_{\mathcal{G}}, \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}}) \propto \prod_{\ell=1}^L e^{-\lambda_{\mathcal{G}} \theta_{\mathcal{G}_\ell} \|\boldsymbol{\beta}_{\mathcal{G}_\ell}\|_2}$$

Let $Y = (y^1, \dots, y^N)$, $X_I = (\mathbf{x}_I^1, \dots, \mathbf{x}_I^N)$ and $X_{\mathcal{G}} = (\mathbf{x}_{\mathcal{G}}^1, \dots, \mathbf{x}_{\mathcal{G}}^N)$.

The generative model is given by:

$$\begin{aligned} & p(\mathbf{W}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_{\mathcal{G}}, \beta_0, Y, X_I, X_{\mathcal{G}}; \lambda_W, \lambda_I, \lambda_{\mathcal{G}}, \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}}) \\ & \stackrel{\text{Bayes}}{=} p(Y, X_I, X_{\mathcal{G}} | \mathbf{W}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_{\mathcal{G}}) p(\mathbf{W}; \lambda_W, \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}}) p(\boldsymbol{\beta}_I; \lambda_I) p(\boldsymbol{\beta}_{\mathcal{G}}; \lambda_{\mathcal{G}}, \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}}) p(\beta_0) \\ & \stackrel{\text{obs iid}}{=} \left(\prod_{k=1}^N p(y = y^k, \mathbf{x}_I^k, \mathbf{x}_{\mathcal{G}}^k | \mathbf{W}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_{\mathcal{G}}) \right) \\ & \quad p(\mathbf{W}; \lambda_W, \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}}) p(\boldsymbol{\beta}_I; \lambda_I) p(\boldsymbol{\beta}_{\mathcal{G}}; \lambda_{\mathcal{G}}, \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}}) p(\beta_0) \\ & \propto \prod_{k=1}^N \sigma\left(\left(\mathbf{x}_{\mathcal{G}}^k\right)^T \mathbf{W}^T \mathbf{x}_I^k + \boldsymbol{\beta}_I^T \mathbf{x}_I^k + \boldsymbol{\beta}_{\mathcal{G}}^T \mathbf{x}_{\mathcal{G}}^k + \beta_0\right)^{y^k} \\ & \quad \prod_{k=1}^N \left[1 - \sigma\left(\left(\mathbf{x}_{\mathcal{G}}^k\right)^T \mathbf{W}^T \mathbf{x}_I^k + \boldsymbol{\beta}_I^T \mathbf{x}_I^k + \boldsymbol{\beta}_{\mathcal{G}}^T \mathbf{x}_{\mathcal{G}}^k + \beta_0\right)\right]^{1-y^k} \\ & \quad \left(\prod_{i=1}^{|\mathcal{I}|} \prod_{\ell=1}^L e^{-\lambda_W \theta_{\mathcal{G}_\ell} \|\mathbf{W}_{i, \mathcal{G}_\ell}\|_2} \right) \times e^{-\lambda_I \|\boldsymbol{\beta}_I\|_2^2} \times \left(\prod_{\ell=1}^L e^{-\lambda_{\mathcal{G}} \theta_{\mathcal{G}_\ell} \|\boldsymbol{\beta}_{\mathcal{G}_\ell}\|_2} \right) \end{aligned}$$

The *maximum a posteriori* estimation is given by:

$$\begin{aligned} (\widehat{\mathbf{W}}, \widehat{\boldsymbol{\beta}}_I, \widehat{\boldsymbol{\beta}}_{\mathcal{G}}, \widehat{\beta}_0) & \in \underset{\mathbf{W}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_{\mathcal{G}}, \beta_0}{\text{argmax}} p(\mathbf{W}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_{\mathcal{G}}, \beta_0 | Y, X_I, X_{\mathcal{G}}; \lambda_W, \lambda_I, \lambda_{\mathcal{G}}, \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}}) \\ & \in \underset{\mathbf{W}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_{\mathcal{G}}, \beta_0}{\text{argmax}} p(\mathbf{W}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_{\mathcal{G}}, \beta_0, Y, X_I, X_{\mathcal{G}}; \lambda_W, \lambda_I, \lambda_{\mathcal{G}}, \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}}) \end{aligned}$$

It is equivalent to minimize the function S defined by:

$$\begin{aligned} S(\mathbf{W}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_G, \beta_0) &= -\log p(Y, \mathbf{W}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_G, \beta_0, X_I, X_G; \lambda_W, \lambda_I, \lambda_G, \mathcal{G}, \boldsymbol{\theta}_G) \\ &= R_N(\mathbf{W}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_G, \beta_0) + \Omega(\mathbf{W}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_G) \end{aligned}$$

4.4 Experimental results

4.4.1 Dataset

We worked with a subset of ADNI1, in which all subjects have been genotyped. This subset contains 707 subjects, with 156 Alzheimer’s Disease patients (denoted AD), 196 MCI patients at baseline who progressed to AD (denoted pMCI, as progressive MCI), 150 MCI patients who remain stable (denoted sMCI, as stable MCI) and 201 healthy control subjects (denoted CN).

In ADNI1 GWAS dataset, 620,901 SNPs have been genotyped, but we selected 1,107 SNPs based on the 44 first top genes related to AD (from AlzGene, <http://www.alzgene.org>) and on the Illumina annotation using the Genome build 36.2. Group weighting for genes is based on gene size: for group \mathcal{G}_ℓ , the weight $\theta_{\mathcal{G}_\ell} = \sqrt{|\mathcal{G}_\ell|}$ ensures that the penalty term is of the order of the number of parameters of the group.

The parameter λ_G influences the number of groups that are selected by the model. In particular, the group \mathcal{G}_ℓ enters in the model during the first iteration if $\left\| \nabla_{\beta_{\mathcal{G}_\ell}} R_N(\mathbf{0}) \right\|_2 > \lambda_G \theta_{\mathcal{G}_\ell}$. This inequality gives an upper bound for λ_G . The same remark can be done for λ_W . Regarding MRI modality, we used the segmentation of FreeSurfer which gives the volume of subcortical regions (44 features) and the average cortical region thickness (70 features). Therefore, there are $1,107 \times 114 = 126,198$ parameters to infer for \mathbf{W} , 114 parameters for $\boldsymbol{\beta}_I$ and 1,107 parameters for $\boldsymbol{\beta}_G$.

4.4.2 Results

We ran our multilevel model and compared it to the logistic regression applied to one single modality with simple penalties (lasso, group lasso, ridge), to additive models ([Wang et al., 2012], [Aiolli and Donini, 2015] EasyMKL with a linear kernel for each modality, and the model $p(y = 1 | \mathbf{x}_G, \mathbf{x}_I) = \sigma(\boldsymbol{\beta}_I^\top \mathbf{x}_I + \boldsymbol{\beta}_G^\top \mathbf{x}_G + \beta_0)$ with our algorithm under the constraint $\boldsymbol{\beta}_G \neq \mathbf{0}$), and to the multiplicative model with \mathbf{W} only, where $p(y = 1 | \mathbf{x}_G, \mathbf{x}_I) = \sigma(\mathbf{x}_G^\top \mathbf{W}^\top \mathbf{x}_I + \beta_0)$. We considered two classification tasks: “AD versus CN” and “pMCI versus CN”. Four measures are used: the sensitivity (SEN), the specificity (SPE), the precision (PRE) and the balanced accuracy between the sensitivity and the specificity (BACC). A 10-fold cross validation is performed. The parameters $\lambda_W, \lambda_I, \lambda_G$ are optimised between $[10^{-3}, 1]$. Classification results for these tasks are shown on table 7.1. It typically takes between 5 and 8 minutes to learn the parameters.

We compared our approach to [Wang et al., 2012, Aiolli and Donini, 2015], for which the codes are available. The features that are selected by [Wang et al., 2012, Aiolli and Donini, 2015] are similar to ours for each modality taken separately. For instance, for [Wang et al., 2012] and the task “AD versus CN”, SNPs that have the most important weights are in genes

TABLE 4.1: Classification results for different modalities and methods

		AD VERSUS CN (%)			
MODALITY	METHOD & PENALTY	SEN	SPE	PRE	BACC
SNPs only	logistic regression (lasso ℓ_1)	69.4	77.5	71.1	73.4
SNPs grouped by genes	logistic regression (group lasso)	69.4	77.5	71.1	73.4
MRI (cortical)	logistic regression (ridge ℓ_2)	84.4	89.5	87.1	86.9
MRI (subcortical)	logistic regression (ridge ℓ_2)	80.0	86.0	83.2	83.0
SNP + MRI (all)	[Aiolli and Donini, 2015] MKL	89.4	85.0	83.0	87.2
SNP + MRI (all)	[Wang et al., 2012]	89.4	88.0	85.7	88.7
SNP + MRI (all)	additive model (β_I, β_G only)	88.8	89.5	87.6	89.1
SNP + MRI (all)	multiplicative model (\mathbf{W} only)	89.4	87.0	85.0	88.2
SNP + MRI (all)	multilevel model (all)	90.6	87.0	85.5	88.8

		pMCI VERSUS CN (%)			
MODALITY	METHOD & PENALTY	SEN	SPE	PRE	BACC
SNPs only	logistic regression (lasso ℓ_1)	72.0	77.0	75.9	74.5
SNPs grouped by genes	logistic regression (group lasso)	72.0	77.0	75.9	74.5
MRI (cortical)	logistic regression (ridge ℓ_2)	74.0	76.0	76.4	75.0
MRI (subcortical)	logistic regression (ridge ℓ_2)	73.0	76.5	76.6	74.7
SNP + MRI (all)	[Aiolli and Donini, 2015] MKL	77.0	73.5	75.1	75.3
SNP + MRI (all)	[Wang et al., 2012]	79.5	81.5	82.4	80.5
SNP + MRI (all)	additive model (β_I, β_G only)	80.5	81.0	82.0	80.8
SNP + MRI (all)	multiplicative model (\mathbf{W} only)	81.0	81.5	82.9	81.3
SNP + MRI (all)	multilevel model (all)	82.5	83.0	84.1	82.8

APOE (rs429358), BZW1 (rs3815501) and MGMT (rs7071424). However, the genetic parameter vector learnt from [Wang et al., 2012] or [Aiolli and Donini, 2015] is not sparse, in contrary of ours. Furthermore, for [Aiolli and Donini, 2015], the weight for the imaging kernel is nine times much larger than the weight for the genetic kernel. These experiments show that the additive model with adapted penalties for each modality provides better performances than [Aiolli and Donini, 2015], but our additive, multiplicative and multilevel models provide similar performances.

4.4.3 Parameters

Regarding MRI features, the most important features (in weight) are the left/right hippocampus, the left/right Amygdala, the left/right entorhinal and the left middle temporal cortices. Regarding genetic features, the most important features in weight are SNPs that belong to gene APOE (rs429358) for both tasks "AD versus CN" and "pMCI versus CN".

Regarding the matrix \mathbf{W} , the couples (brain region, gene) learnt through the task "pMCI versus CN" are shown on Fig. 4.2. It can be seen that \mathbf{W} has a sparse structure. Among the couples (brain region, gene) that have non null coefficients for the both tasks "AD versus CN" and "pMCI versus CN", there are (Left Hippocampus, MGMT), (Right Entorhinal, APOE) or (Left Middle Temporal, APOE). Only couples related to AD are selected by the model.

Figure 4.3 provides another possible interpretation of vector β_I and columns of \mathbf{W} . By projecting the coefficients onto the brain, it is possible to visualise brain regions that are strongly related to AD (first row), and brain regions, combined with SNP “rs429358” of APOE, related to AD (second row).

Another possible visualisation of matrix \mathbf{W} is to display a Manhattan plot for some particular brain regions. In figures 4.4 and 4.5, Manhattan plot for the rows of \mathbf{W} corresponding to the left entorhinal, the left hippocampus, the left middle temporal and the right entorhinal are shown. These plots highly SNPs that are strongly correlated with these brain regions and AD.

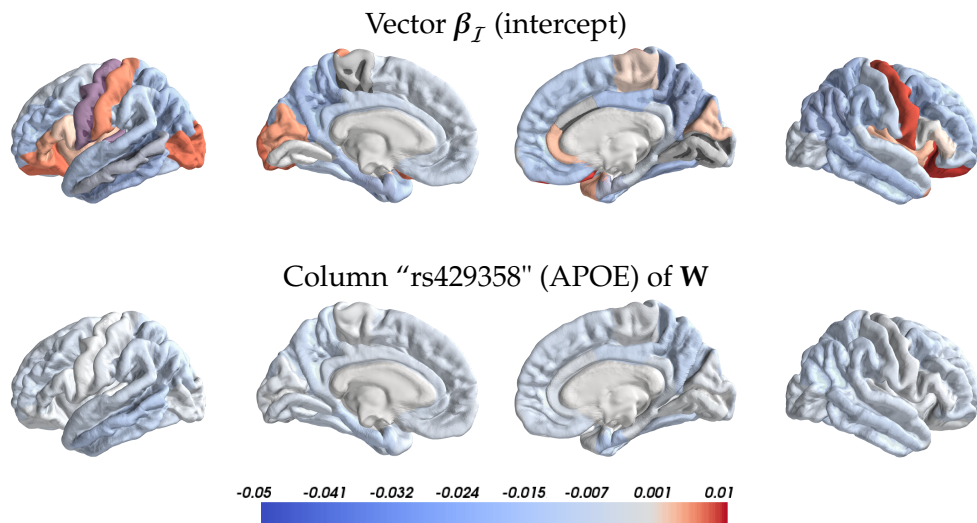


FIGURE 4.3: Intercept β_I and slope \mathbf{W} in the function $\alpha(x_G) = \mathbf{W}x_G + \beta_I$

4.5 Conclusion

In this chapter, we developed a novel approach to integrate genetic and brain imaging data for prediction of disease status. Our multilevel model takes into account potential interactions between genes and brain regions, but also the structure of the different types of data through the use of specific penalties within each modality. When applied to genetic and MRI data from the ADNI database, the model was able to highlight brain regions and genes that have been previously associated with AD, thereby demonstrating the potential of our approach for imaging genetics studies in brain diseases.

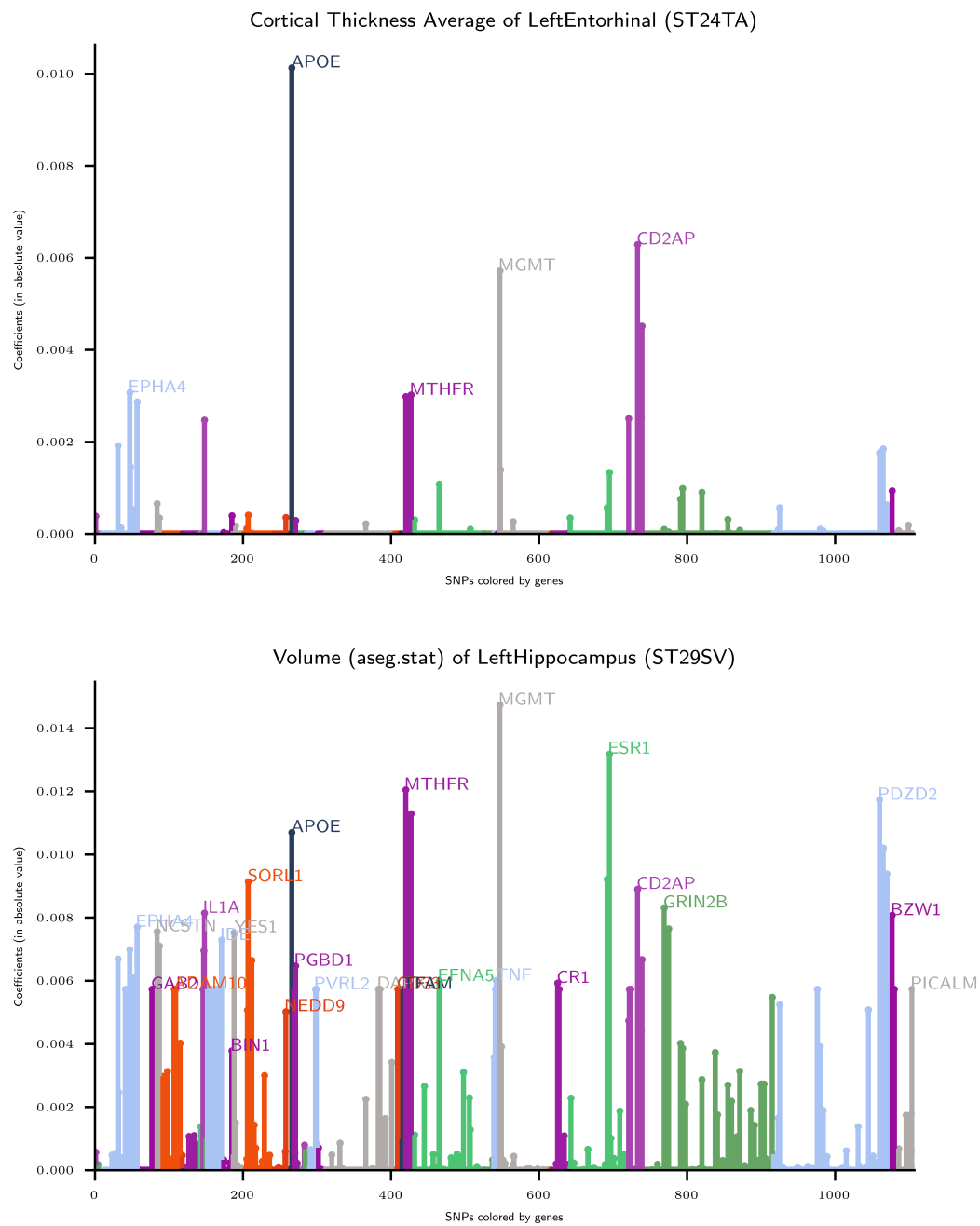


FIGURE 4.4: Rows of W

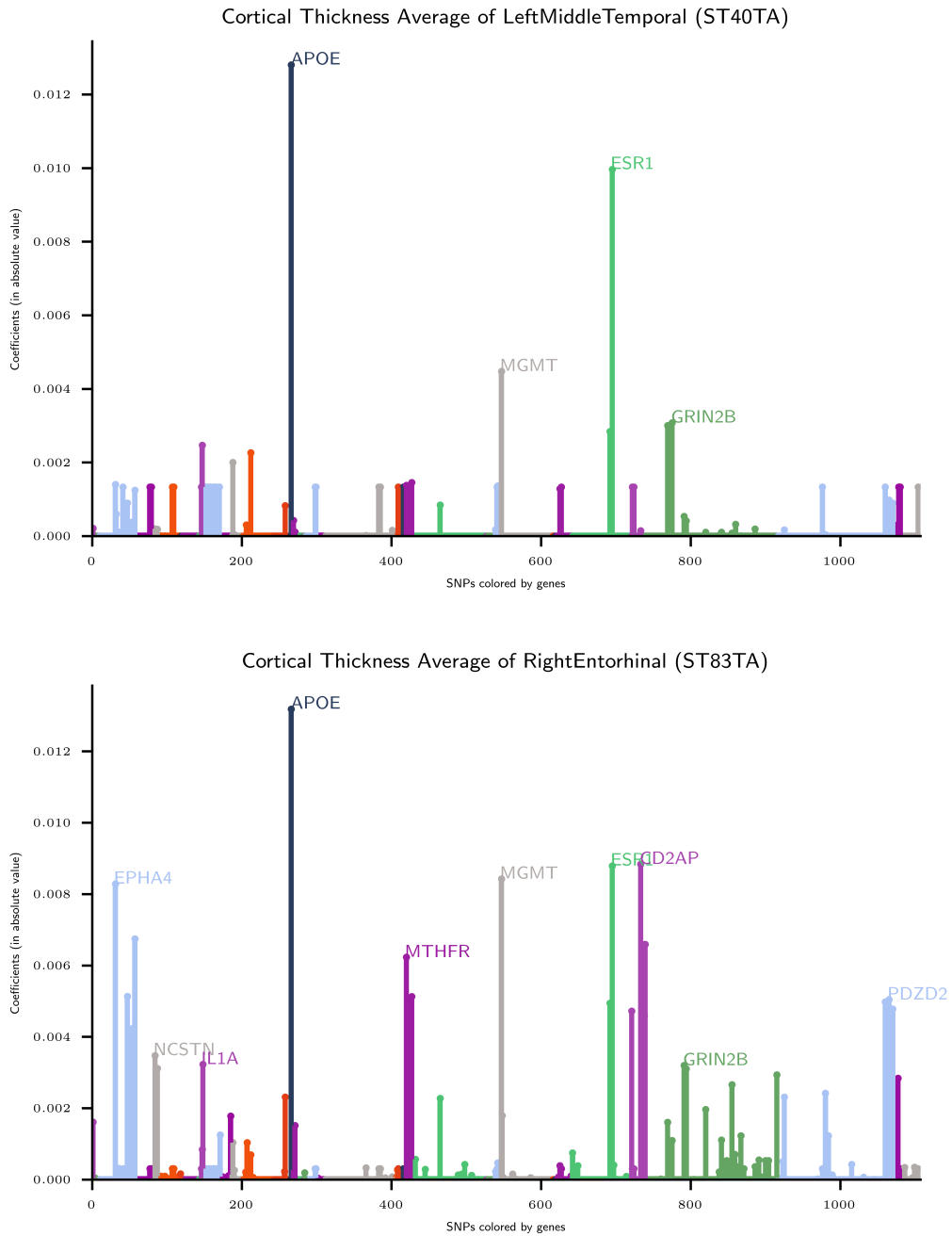


FIGURE 4.5: Rows of W

Chapter 5

Contributions to the TADPOLE 2017-2022 challenge

5.1 Introduction

The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge aims at comparing methods and algorithms at predicting the future evolution of patients at risk of Alzheimer's Disease. Details concerning the project are available on their website:

<https://tadpole.grand-challenge.org>

Launch in 2017, the TADPOLE requires participants to submit their monthly forecasts for a period of 5 years from January 2018 to December 2022.

We start by giving some details on the organisation of the TADPOLE challenge, and the metrics that they use to assess the performance of each submission. Then, we describe our methodology and contributions to the TADPOLE challenge; and provide some results, after the first phase of evaluation in January 2019.

5.2 Project details

This section is mainly based on [[Marinescu et al., 2018](#)].

5.2.1 Overview

TADPOLE Challenge aims at comparing methods and algorithms at predicting the future evolution of patients at risk of Alzheimer's Disease. Participants can train their models on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study or add any other datasets. Once the model trained, they have to make monthly forecasts for a period of 5 years from January 2018 to December 2022, for all ADNI3 rollover participants. Predictions that are asked by TADPOLE are:

- the probabilistic clinical diagnosis, which can be cognitively normal (CN), mild cognitive impairment (MCI) or Alzheimer's disease (AD).
- the cognitive score ADAS-Cog13 (Alzheimer's Disease Assessment Scale Cognitive Subdomain),

- the total volume of the ventricles divided by intra-cranial volume.

These forecasts will be compared with the corresponding future measurements in ADNI3, that will be obtained after the TADPOLE submission deadline. The timeline is described in table 5.1.

Submission deadline	15 th November 2017
Test set complete	November 2018
Evaluation results on website	January 2019
Review first phase	March 2019

TABLE 5.1: TADPOLE timeline

Figure 5.1 shows an example of submission file.

RID	Month	Date	p_{CN}	p_{MCI}	p_{AD}	ADAS	ADAS 50% CI lower	ADAS 50% CI upper	Ventri.	Ventri. 50% CI lower	Ventri. 50% CI upper
55	1	2018-01									
55	2	2018-02									
55	3	2018-03									
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
55	60	2022-12									
56	1	2018-01									
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
56	60	2022-12									
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

FIGURE 5.1: Submission file format for TADPOLE (where Ventri. stands for ventricle volume and CI for confidence interval)

5.2.2 Datasets

Different kind of datasets are built by the TADPOLE Challenge: a training dataset, a prediction dataset and a test dataset. The organisers provide a training dataset and a prediction dataset. The training dataset contains measurements with associated outcomes and can be used to train algorithms. The prediction dataset contains only longitudinal or baseline measurements, without the associated outcomes. The algorithm uses it as input to predict the outcome (patient status) that are in the test dataset. During the TADPOLE challenge, this test dataset does not exist; and will be built after the submission deadline. More precisely, the organisers named their dataset D1, D2, D3 and D4 where:

- D1 is the training dataset containing all longitudinal data from the entire ADNI history (ADNI1, ADNI GO and ADNI2). Each individual has been assessed at least two separate visits across the study. Each dataset contains measurements (clinical, biological biomarkers) for every individual.
- D2 is the longitudinal prediction dataset. It contains all past time-points for individuals for whom it is required to make forecasts.

- D3 is the cross-sectional prediction dataset. It contains only one time-point for individuals for whom it is required to make forecasts. Although forecasts will be worst for D3 than for D2, D3 is the information that is typically available when a cohort is created for a clinical trial.
- D4 is the test set, on which the forecast will be compared. It will be built on ADNI3 rollovers after the challenge submission deadline.

It is also possible to work on other training datasets than D1 in order to build the model. All these datasets are summarised in Figures and .

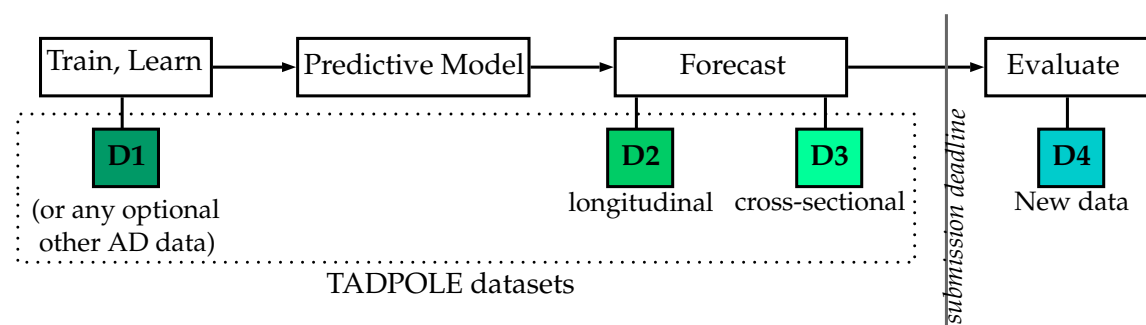


FIGURE 5.2: TADPOLE Challenge design. Participants are required to train a predictive model on a training dataset (D1 and/or others) and make forecasts for different datasets (D2, D3) by the submission deadline. Evaluation will be performed on a test dataset (D4) that is acquired after the submission deadline (reproduced from [Marinescu et al., 2018]).

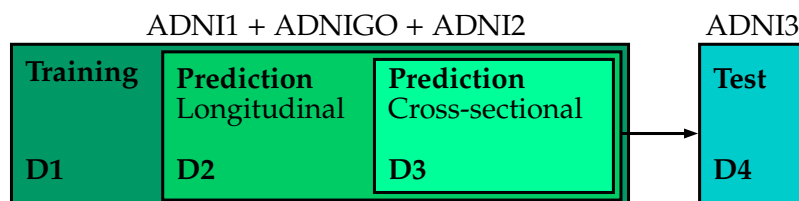


FIGURE 5.3: Venn diagram of the ADNI datasets for training (D1), longitudinal prediction (D2), cross-sectional prediction (D3) and the test set (D4). D3 is a subset of D2, which in turn is a subset of D1. Other non-ADNI data can also be used for training (reproduced from [Marinescu et al., 2018]).

Descriptive statistics of D1, D2, D3 datasets Table 5.2 provide some descriptive statistics of D1, D2, D3 and D4 datasets, grouped by diagnosis at the first available visit (CN, MCI or AD).

5.2.3 Metrics

5.2.3.1 Multiclass Area Under the ROC Curve

The Multiclass Area Under the ROC Curve (mAUC) is an extension of the Area under the ROC (Receiver Operating Characteristic) curve for problems with more than two classes.

Dataset		D1	D2	D3
Number of subjects		1667	896	896
CN	Number (%)	508 (30.5%)	369 (41.2%)	299 (33.4%)
	Visits per subject	8.3 (4.5)	8.5 (4.9)	1.0 (0.0)
	Age	74.3 (5.8)	73.6 (5.7)	72.3 (6.2)
	Gender (% male)	48.6%	47.2%	43.5%
	MMSE	29.1 (1.1)	29.0 (1.2)	28.9 (1.4)
	Converters	18 (3.5%)	9 (2.4%)	
MCI	Number (%)	841 (50.4%)	458 (51.1%)	269 (30.0%)
	Visits per subject	8.2 (3.7)	9.1 (3.6)	1.0 (0.0)
	Age	73.0 (7.5)	71.6 (7.2)	71.9 (7.1)
	Gender (% male)	59.3%	56.3%	58.0%
	MMSE	27.6 (1.8)	28.0 (1.7)	27.6 (2.2)
	Converters	117 (13.9%)	37 (8.1%)	
AD	Number (%)	318 (19.1%)	69 (7.7%)	136 (15.2%)
	Visits per subject	4.9 (1.6)	5.2 (2.6)	1.0 (0.0)
	Age	74.8 (7.7)	75.1 (8.4)	72.8 (7.1)
	Gender (% male)	55.3%	68.1%	55.9%
	MMSE	23.3 (2.0)	23.1 (2.0)	20.5 (5.9)
	Converters			

TABLE 5.2: Descriptive statistics of D1, D2, D3 datasets (reproduced from <https://tadpole.grand-challenge.org/Results/>)

For classification of a class i against another class j , the AUC $\widehat{A}(c_i|c_j)$ is defined by:

$$\widehat{A}(i|j) = \frac{S_i n_i (n_i + 1) / 2}{n_i n_j}$$

where n_i (resp. n_j) is the number of points belonging to class i (resp. j); while S_i is the sum of the ranks of the class i test points after ranking all the class i and j data points in increasing likelihood of belonging to class i .

The average AUC for classes i and j is defined as:

$$\widehat{A}(i, j) = \frac{1}{2} (\widehat{A}(i|j) + \widehat{A}(j|i))$$

The overall mAUC is obtained by averaging $\widehat{A}(i, j)$ over all pairs of classes:

$$\text{mAUC} = \frac{2}{L(L-1)} \sum_{i=2}^L \sum_{j=1}^{i-1} \widehat{A}(i, j) = \frac{1}{3} (\widehat{A}(\text{MCI}, \text{CN}) + \widehat{A}(\text{AD}, \text{CN}) + \widehat{A}(\text{AD}, \text{MCI}))$$

5.2.3.2 Balanced Classification Accuracy

The Balanced Classification Accuracy is a generalisation of the classification accuracy measure that accounts for the imbalance in the numbers of data-points belonging to each class. On the contrary of the mAUC, the balanced accuracy is not probabilistic. A hard classification to the class (CN, MCI, or AD) with the highest likelihood needs to be assigned

to each datapoint. The balanced accuracy for class c_i is given by:

$$BCA_i = \frac{1}{2} \left[\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right]$$

where TP, FP, TN, FN are respectively the number of true positives, false positives, true negatives and false negatives for classification as class c_i .

The overall BCA is defined as the mean of all the balanced accuracies for every class.

$$BCA_i = \frac{1}{3} (BCA_{CN} + BCA_{MCI} + BCA_{AD})$$

5.3 Contributions

Our contribution to the TADPOLE project will be limited to the prediction of the clinical diagnosis of Alzheimer's Disease. TADPOLE requires to forecast a probability for each date and subject. For that purpose, we will provide a cross-sectional approach.

5.3.1 Disease progression paths

For prediction of diagnosis, there are six possible paths of disease progression, shown on figure 5.4. Although there are other paths of disease progression – such as AD → MCI –, we considered that there are not medically relevant and represent only some isolated cases.

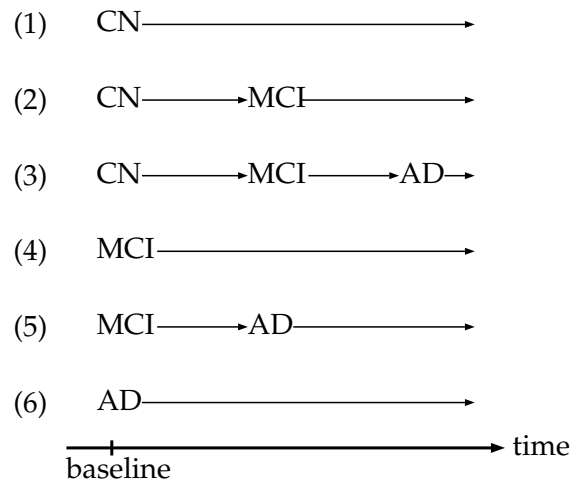


FIGURE 5.4: Possible paths

Based on D1 (in particular on table 5.2), it is possible to quantify the proportion of each path, as show on figure 5.5. Among subjects who are cognitively normal at baseline, only 3.5% will convert to MCI. Among patients who have mild cognitively normal at baseline, approximatively 13.9% will convert to AD.

Based on figure 5.5, we will make the assumption Cognitively Normal subjects and AD patients will not change state and remain constant throughout time. Therefore, only the disease path "MCI to AD" need to be modelled. If we denote $p_{MCI}(t)$ the probability to

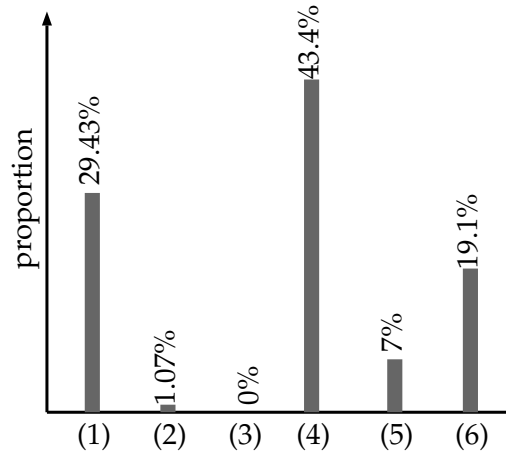


FIGURE 5.5: Proportion of each possible path based on D1 dataset

remain MCI at time t , the probability to be AD is $p_{AD}(t) = 1 - p_{MCI}(t)$ and the probability to be CN is then $p_{CN}(t) = 0$.

5.3.2 Estimation of the conversion date to AD for MCI patients

Survival analysis The framework of survival analysis allows us to estimate directly the conversion date T to AD by estimating the survival function:

$$S : t \mapsto \mathbb{P}\{T \geq t\}$$

We made the underlying hypothesis that all MCI patients will convert to AD.

Aalen model Let $\mathbf{x} = (x_1, \dots, x_r)$ be the vector of covariates. In the Aalen model [Aalen, 1980], the hazard function is given by:

$$h(t|\mathbf{x}) = a_0(t) + a_1(t)x_1 + \dots + a_r(t)x_r$$

a_0 is a baseline function; while the regression functions a_1, \dots, a_r measure the influence of the respective covariates.

The survival function can be deduced:

$$S(t|\mathbf{x}) = \mathbb{P}\{T \geq t|\mathbf{x}\} = \exp\left(-\int_0^t h(u)du\right) = \exp(-A_0(t) - A_1(t)x_1 - \dots - A_n(t)x_r)$$

where $A_j(t) = \int_0^t a_j(u)du$ is the cumulative regression function.

Estimation of regression functions in the Aalen model For each individual $i \in \{1, \dots, n\}$, we denote h_i the hazard rate of individual i . We note $h(t) = (h_1(t), \dots, h_n(t))^T$ the column vector of hazard rates.

The additive model is given by $h(t) = Y(t)\alpha(t)$ where $\alpha(t) = (a_0(t), a_1(t), \dots, a_r(t))^\top$ is the regression information and $Y(t) \in \mathcal{M}_{n,r+1}(\mathbb{R})$ the design matrix where i^{th} row of $Y(t)$ is:

$$Y_i(t) = \begin{cases} (1, x_1^i, \dots, x_r^i) = (1, \mathbf{x}^i) & \text{if the } i^{\text{th}} \text{ individual is a member of the risk set at time } t \\ (0, \dots, 0) & \text{otherwise} \end{cases}$$

It is easier to estimate the cumulative regression functions defined by $A_j(t) = \int_0^t a_j(u)du$ than $a_j(t)$. Let $A(t) = (A_0(t), \dots, A_r(t))^\top$. $A(t)$ is estimated by an approach similar to the ordinary linear model:

$$\widehat{A}(t) = \sum_{T_{(k)} \leq t} X(T_{(k)})I_k$$

where $T_{(1)} < T_{(2)} < \dots$ are the ordered event times and I_k is a column vector consisting of zeros except for a one in the place corresponding to the subject who experiences an event at time $T_{(k)}$. The matrix $X(t)$ is defined as the ordinary least squares inverse of $Y(t)$:

$$X(t) = (Y(t)^\top Y(t))^{-1} Y(t)^\top$$

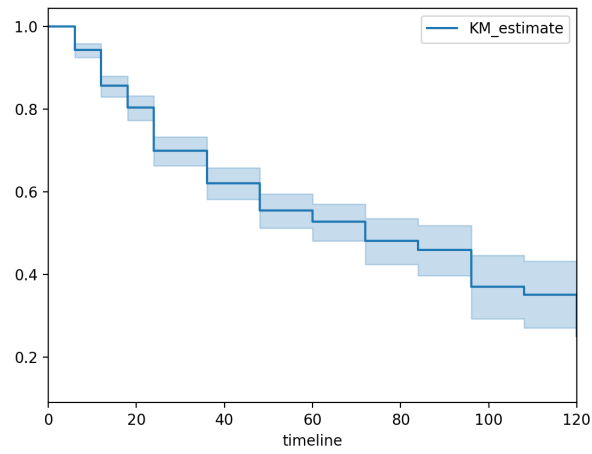
Features We decide to work with cross-sectional data, and use only one time-point for each subject.

We chose several non-time dependent features such as genetics (APOE4, as some alleles of this gene increase the factor of development of AD) and the level of education (PTEDUCAT). We also took some time-dependent features such as:

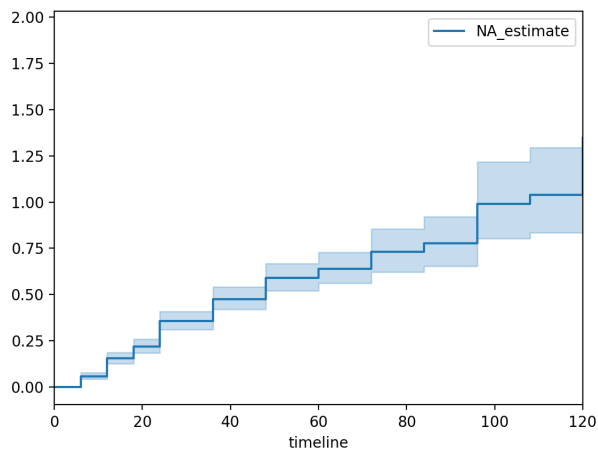
- cognitive and clinical scores: CDRSB, ADAS11, ADAS13, MMSE, RAVLT (immediate), FAQ.
- volumetric of the following anatomical structured, calculated using medical resonance images: Hippocampus, Ventricles, Whole Brain, Entorhinal, Fusiform, Mid Temporal, ICV.
- the sum of mean glucose metabolism uptake in regions of angular, temporal, and posterior cingulate, calculated from PET scans: FDG.
- the concentration of amyloid peptide $A\beta$ -24-1, τ and τ -phosphorylated proteins in the cerebrospinal fluid (CSF) features: ABETA_UPENNBIOMK9_04_19_17, TAU_UPENNBIOMK9_04_19_17, PTAU_UPENNBIOMK9_04_19_17.
- the subject's age: AGE.

These time-dependant features provide a good "picture" of the subject's state at the time they were acquired. At the end, our model will use 20 relevant features, which are all known to be associated to AD diagnosis and/or with progression to AD.

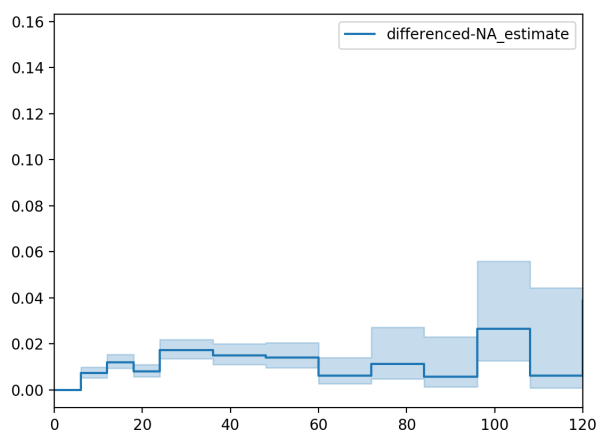
We do not apply any pre- or post-processing. Missing values are imputed from the average on the population. There is no confounder correction.



(A) Kaplan-Meier estimator of $\widehat{S} : t \mapsto \mathbb{P}\{T \geq t\}$



(B) Nelson-Aalen estimator of the cumulative hazard function \widehat{H}



(C) hazard function \widehat{h} , computed from the cumulative hazard function with a bandwidth of 6 months

FIGURE 5.6: Estimators for the survival, cumulative hazard, and hazard functions

hazard function and survival function Average hazard function and survival function estimated using respectively the Kaplan-Meier and Nelson-Aalen estimators are shown in Figures 5.6a and 5.6c. The median conversion time is 72 months.

Cumulative regression functions We work only with cross-sectional data. To train our model, we chose the first timepoint (baseline) for each subject of D1, and we tested on the other timepoints.

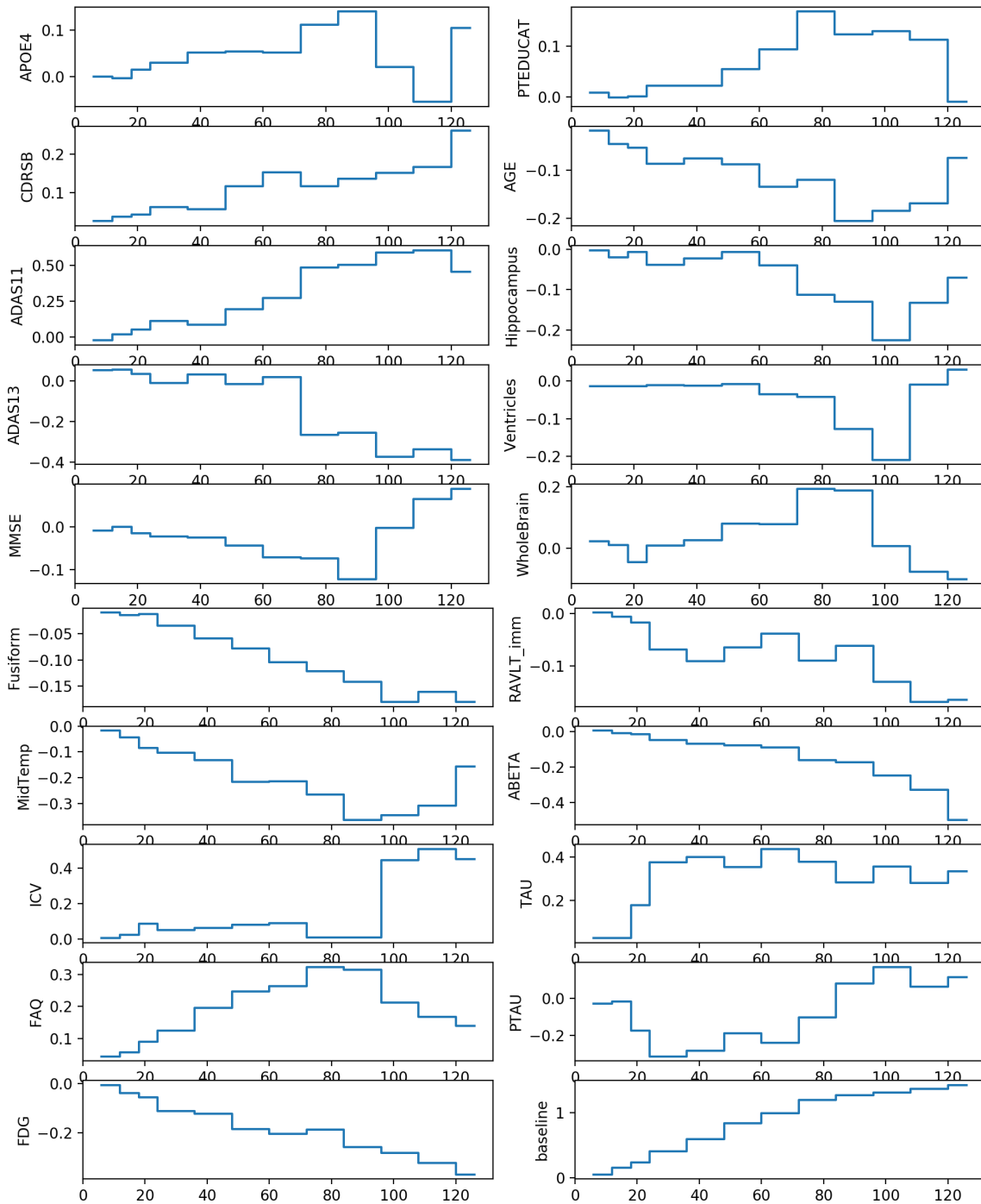


FIGURE 5.7: Cumulative regression functions of $\hat{A}(t)$

To learn our model, we chose the last timepoint for each subject of D2. The cumulative regression functions are determined using the implementation of Lifelines (<https://lifelines.readthedocs.io/en/latest/>) on the dataset D2. There are 755 observations, including 455 censored.

The components of $\widehat{A}(t)$ are to be plotted on figure 5.7 and give information about effects of covariates. Both signs and slopes and $A_j(t)$ bring informations.

Processing time On a Macbook Pro 2016, it typically takes approximatively 20 seconds for training, and less than 0.02 seconds for forecasting the predictions for one subject.

5.4 Results and evaluation on D4

Dataset D4 Table 5.3 provide some descriptive statistics of D1, D2, D3 and D4 datasets, grouped by diagnosis at the first available visit (CN, MCI or AD). D4 is built from ADNI3 data acquired until January 2019.

Status at baseline	CN	MCI	AD
Number (%)	94 (42.9%)	90 (41.1%)	29 (13.2%)
Visits per subject	1.0 (0.2)	1.1 (0.3)	1.1 (0.3)
Age	78.4 (7.0)	79.4 (7.0)	82.2 (7.6)
Gender (% male)	47.9%	64.4%	51.7%
MMSE	29.1 (1.1)	28.1 (2.1)	19.4 (7.2)
Converters		9 (10.0%)	9 (31.0%)

TABLE 5.3: Descriptive statistics of D4 dataset (219 subjects, reproduced from <https://tadpole.grand-challenge.org/Results/>)

Figure 5.8 shows that conversion paths (2) and (3) are non-existent, and therefore modelling only the conversion path (5) is relevant. However, on the contrary of D1, D4 contains other paths of conversions, in particular with AD as starting point.

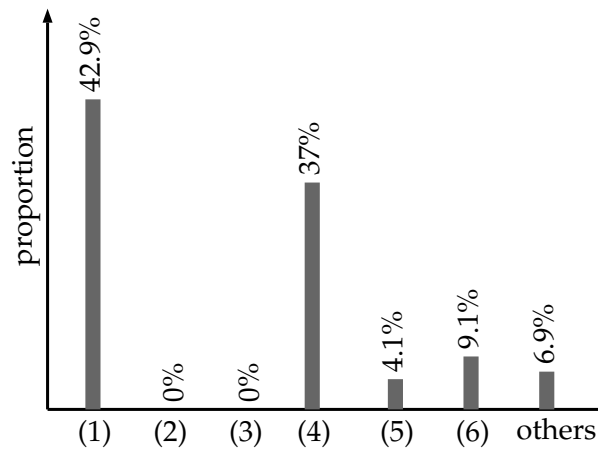


FIGURE 5.8: Proportion of each possible paths for D4 dataset

Participating teams and rankings All results are available at:

<https://tadpole.grand-challenge.org/Results/>

There were a total of 33 participating teams, who submitted a total of 58 forecasts from D2, 34 forecasts from D3, and 6 forecasts from custom prediction sets. Methods used by participating teams are described at:

<https://docs.google.com/document/d/1bh7QoI4Uyz2Q9GMnghmlsJvS1SirI7ttWugmRUGE8ic/edit>

The methodology we used was ranked 15th in terms of mAUC (0.876), but came first in terms of balanced accuracy (0.85); details results are shown on table 5.4 and 5.5. Although our model is a cross-sectional model, we preferred to work with the longitudinal dataset D2 (instead of D3); and used the last timepoint for prediction of dataset D2.

RANK	TEAM NAME	MAUC	BCA	Feature selection	Number of features	Missing data imputation	Diagnosis prediction
1	Frog	0.931	0.849	automatic	70+420*	none	gradient boosting
2	Threedays	0.921	0.823	Manual	16	none	RF
3	EMC-EB	0.907	0.805	automatic	200-338	nearest-neighbour	SVM classifier
4	GlassFrog-Average	0.902	0.825	semi-automatic	all	forward-fill	multi-state model
5	GlassFrog-LCMEM-HDR	0.902	0.825	Manual	7	linear model	multi-state model
6	GlassFrog-SM	0.902	0.825	semi-automatic	all	forward-fill/linear	multi-state model
7	Apocalypse ★	0.902	0.827	Manual	16	population average	SVM
8	EMC1-Std	0.898	0.811	automatic	250	nearest neighbour	DPM + 2D spline + SVM
9	CBIL	0.897	0.803	Manual	21	linear interpolation	LSTM
10	CN2L-RandomForest	0.896	0.792	automatic	all	forward-filling	RNN
11	EMC1-Custom	0.892	0.798	automatic	250	nearest neighbour	DPM + 2D spline + SVM
12	BGU-LSTM	0.883	0.779	automatic	67	none	feed-forward NN
13	DIKU-GeneralisedLog-Custom	0.878	0.790	semi-automatic	18	none	Bayesian classifier/LDA + DPM
14	DIKU-GeneralisedLog-Std	0.877	0.790	semi-automatic	18	none	Bayesian classifier/LDA + DPM
15	ARAMIS-Pascal ★	0.876	0.850	Manual	20	population average	Aalen model
16	VikingAI-Sigmoid	0.875	0.760	Manual	10	none	DPM + ordered logit model
17	Tohka-Ciszek-RandomForestLin	0.875	0.796	Manual	32	nearest neighbour	-
18	IBM-OZ-Res	0.868	0.766	Manual	oct-15	filled with zero	stochastic gradient boosting
19	BORREGOTECHTY	0.866	0.808	automatic	100 + 400*	nearest-neighbour	regression ensemble
20	VikingAI-Logistic	0.865	0.754	Manual	10	none	DPM + ordered logit model
21	lmaUCL-Std	0.859	0.781	Manual	5	regression	multi-task learning
22	lmaUCL-Covariates	0.852	0.760	Manual	5	regression	multi-task learning
23	Chen-MCW-Stratify	0.848	0.783	Manual	9	none	linear regression
24	AlgosForGood	0.847	0.810	Manual	16+5*	forward-filling	Aalen model
25	lmaUCL-halfD1	0.845	0.753	Manual	5	regression	multi-task learning
26	Sunshine-Conservative ★	0.845	0.816	semi-automatic	6	population average	SVM
27	CN2L-Average	0.843	0.792	automatic	all	forward-filling	RNN/RF
28	BGU-RF	0.838	0.673	automatic	67+1340*	none	semi-temporal RF
29	BenchmarkSVM	0.836	0.764	Manual	6	mean of previous values	SVM
30	Chen-MCW-Std	0.836	0.778	Manual	9	none	linear regression
31	FortuneTellerFish-Control	0.834	0.692	Manual	19	nearest neighbour	multiclass ECOC SVM
32	BGU-RFFIX	0.831	0.673	automatic	67+1340*	none	semi-temporal RF
33	Sunshine-Std ★	0.825	0.771	semi-automatic	6	population average	SVM
34	CyberBrains	0.823	0.747	Manual	5	population average	linear regression
44	BenchmarkLastVisit	0.774	0.792	None	3	none	constant model

TABLE 5.4: Ranking based on mAUC

RANK	TEAM NAME	MAUC	BCA	Feature selection	Number of features	Missing data imputation	Diagnosis prediction
1	ARAMIS-Pascal ★	0.876	0.850	Manual	20	population average	Aalen model
2	Frog	0.931	0.849	automatic	70+420*	none	gradient boosting
3	Apocalypse ★	0.902	0.827	Manual	16	population average	SVM
4	GlassFrog-Average	0.902	0.825	semi-automatic	all	forward-fill	multi-state model
5	GlassFrog-LCMEM-HDR	0.902	0.825	Manual	7	linear model	multi-state model
6	GlassFrog-SM	0.902	0.825	semi-automatic	all	forward-fill/linear	multi-state model
7	Threedays	0.921	0.823	Manual	16	none	RF
8	Sunshine-Conservative ★	0.845	0.816	semi-automatic	6	population average	SVM
9	EMC1-Std	0.898	0.811	automatic	250	nearest neighbour	DPM + 2D spline + SVM
10	AlgosForGood	0.847	0.810	Manual	16+5*	forward-filling	Aalen model
11	BORREGOTECMTY	0.866	0.808	automatic	100 + 400*	nearest-neighbour	regression ensemble
12	EMC-EB	0.907	0.805	automatic	200-338	nearest-neighbour	SVM classifier
13	CBIL	0.897	0.803	Manual	21	linear interpolation	LSTM
14	EMC1-Custom	0.892	0.798	automatic	250	nearest neighbour	DPM + 2D spline + SVM
15	Tohka-Ciszek-RandomForestLin	0.875	0.796	Manual	32	nearest neighbour	-
16	CN2L-RandomForest	0.896	0.792	automatic	all	forward-filling	RNN
17	CN2L-Average	0.843	0.792	automatic	all	forward-filling	RNN/RF
18	BenchmarkLastVisit	0.774	0.792	None	3	none	constant model
19	Orange	0.774	0.792	Manual	17	none	clinician's decision tree
20	DIKU-GeneralisedLog-Custom	0.878	0.790	semi-automatic	18	none	Bayesian classifier/LDA + DPM
21	DIKU-GeneralisedLog-Std	0.877	0.790	semi-automatic	18	none	Bayesian classifier/LDA + DPM
22	Chen-MCW-Stratify	0.848	0.783	Manual	9	none	linear regression
23	lmaUCL-Std	0.859	0.781	Manual	5	regression	multi-task learning
24	BGU-LSTM	0.883	0.779	automatic	67	none	feed-forward NN
25	Chen-MCW-Std	0.836	0.778	Manual	9	none	linear regression
26	Sunshine-Std ★	0.825	0.771	semi-automatic	6	population average	SVM
27	IBM-OZ-Res	0.868	0.766	Manual	oct-15	filled with zero	stochastic gradient boosting
28	BenchmarkSVM	0.836	0.764	Manual	6	mean of previous values	SVM
29	VikingAI-Sigmoid	0.875	0.760	Manual	10	none	DPM + ordered logit model
30	lmaUCL-Covariates	0.852	0.760	Manual	5	regression	multi-task learning
31	VikingAI-Logistic	0.865	0.754	Manual	10	none	DPM + ordered logit model
32	lmaUCL-halfD1	0.845	0.753	Manual	5	regression	multi-task learning
33	BenchmarkMixedEffectsAPOE	0.822	0.749	None	4	none	Gaussian model
34	CyberBrains	0.823	0.747	Manual	5	population average	linear regression
35	SBIA	0.776	0.721	Manual	30-70	dropped visits with missing data	SVM + density estimator

TABLE 5.5: Ranking based on BCA

5.5 Conclusion

Drawing conclusions based on the ranking provided by TADPOLE would be too premature, as the dataset D4 is still in construction until 2022; however it can be noticed that making the assumption that CN subjects remains constant throughout time is valid. Furthermore, using few features but relevant features provides results that are as good as using all features; leading to models that are easy to compute and understand.

Among the best methods for predicting this disease status, there are on the one hand survival models such as the Aalen model or the multi-state model; and on the other hand, machine learning approaches such as gradient boosting, SVM and random forest.

Chapter 6

Multimodal survival models for predicting the conversion to Alzheimer's Disease

6.1 Introduction

In this chapter, we focus on the application of survival models for the prediction of the conversion date to AD for MCI patients, using their genetics and clinical data at baseline. By using survival models, we make the underlying assumption that MCI patients convert to AD, but the conversion date to AD is not always observed. Furthermore, in this problem, death can be a competitive risk with AD; however this information remains unknown in the ADNI1 dataset.

We start by comparing some baseline models for the prediction of AD (classification at fixed time T , Cox proportional hazard model, Aalen model, log-logistic model) using basic multimodal genetics (APOE, gender) and clinical scores (such as MMSE, ADAS13, RAVLT).

Then, we propose a parametric survival model based on multimodal data to estimate the conversion date to AD from genetics and clinical data. We chose the log-logistic model which provides a parametric framework where the parameters depends on both clinical and genetic data. The hazard function is unimodal, which seems well-suited to our model. In our proposed formulation, genetic data \mathbf{x}_G only influences the speed $v(\mathbf{x}_G)$ at which the conversion would happen, whereas clinical data \mathbf{x}_C influences the initial state of the subject $p(\mathbf{x}_C)$. If S denotes the survival function and $t_{1/2}$ the median survival time, we set $t_{1/2} = p(\mathbf{x}_C)/v(\mathbf{x}_G)$, and $S'(t_{1/2}) = -v(\mathbf{x}_G)$. By determining $v(\mathbf{x}_G)$ and $p(\mathbf{x}_C)$, we are able to determine the associated survival function S .

Dataset and features In this chapter, we will work with MCI patients at baseline from the ADNI1 dataset, whose descriptive statistics are given in chapter . We will select only seven representative features which can be grouped into two categories: non-time varying covariates (APOE status, gender (man = 0, woman = 1), education) and time-varying covariates measured at baseline (MMSE, ADAS13, RAVLT, age).

6.2 Comparison between classification and survival analysis

In this section, features are centred and normalised.

6.2.1 Models

Let $\mathbf{x} = (x_1, \dots, x_r)$ be the vector of covariates for a subject.

Cox proportional hazard model The hazard and survival function for the Cox model are respectively given by:

$$h(t) = h_0(t)e^{\beta^\top \mathbf{x}} = h_0(t)e^{\beta_1 x_1 + \dots + \beta_r x_r} \quad \text{and} \quad S(t) = (S_0(t))^{e^{\beta^\top \mathbf{x}}}$$

Log-logistic model The survival function is defined by:

$$S(t|\mathbf{x}) = \frac{1}{1 + (\lambda(\mathbf{x})t)^p}$$

and the hazard function by:

$$h(t|\mathbf{x}) = \frac{p\lambda(\mathbf{x})(\lambda(\mathbf{x})t)^{p-1}}{1 + (\lambda(\mathbf{x})t)^p}$$

We can set for instance $\lambda(\mathbf{x}) = e^{\alpha_0 + \alpha_1 x_1 + \dots + \alpha_r x_r}$ and $p = e^\beta$

Aalen additive model The hazard function [Aalen, 1980] is given by:

$$h(t|\mathbf{x}) = a_0(t) + a_1(t)x_1 + \dots + a_r(t)x_r$$

where a_0 is a baseline function; while the regression functions a_1, \dots, a_r measure the influence of the respective covariates. The survival function can be deduced:

$$S(t|\mathbf{x}) = \mathbb{P}\{T \geq t|\mathbf{x}\} = \exp\left(-\int_0^t h(u)du\right) = \exp(-A_0(t) - A_1(t)x_1 - \dots - A_r(t)x_r)$$

where $A_j(t) = \int_0^t a_j(u)du$ is the cumulative regression function.

Classification at fixed time T For $T \in \{6, 12, 18, 24, 36, 48, 60, 72, 84\}$, we ran the classical logistic regression defined by:

$$\mathbb{P}\{y = 1|T, \mathbf{x}\} = \frac{1}{1 + e^{-\beta(T)^\top \mathbf{x} - \beta_0(T)}}$$

where $\beta(T)$ is a time-dependent vector, and $\beta_0(T)$ is a time-varying intercept. At month T , the patient i , whose observed date is T_i and if conversion has occurred is represented by δ_i is included in the model if:

- it did not convert before month T ($\delta_i = 0$ and $T_i \geq T$); its corresponding label is $y_i = 0$.
- it converted before month T ($\delta_i = 1$ and $T_i \leq T$); its corresponding label is $y_i = 1$.
- it converted after month T ($\delta_i = 1$ and $T_i > T$); its corresponding label is $y_i = 0$.

At time T , the patient i whose conversion is not observed $\delta_i = 0$ and followed until $T_i < T$ is ignored.

Table 6.1 gives the number of positive labels and total number of labels for different T . The training set varies throughout time. Although it is not a survival model, [Yu et al., 2011] proposes to transform the logistic regression at fixed time T into a survival model by imposing some constraints on β and β_0 .

Months T	6	12	18	24	36	48	60	72	84	96
Number of positive labels	19	65	99	131	158	166	174	180	181	181
Total number of samples	361	339	322	308	278	241	235	227	219	207

TABLE 6.1: Number of positive labels and total number of labels throughout time

6.2.2 Effect of covariates

We remind that when the patient becomes AD, RAVLT and MMSE are decreasing whereas ADAS-Cog13 is increasing. We also remind that the population with one or two alleles of APOE $\epsilon 4$ is more important for AD than for CN.

Figure 6.1 shows the coefficients β determined for the Cox PH model. The sign for each coefficient is coherent with the direction of variation of each coefficient when the disease progresses.

When we deal with the Cox PH model, it is important to test if each variable in the model verifies the proportional hazard assumption. Results shown on table 6.2 show that gender, APOE4 and education do not verify this assumption. As they do not respect this assumption, the functional form is incorrect, and that there may be non-linear missing terms. For instance, we can in the model education^2 such that the proportional hazard assumption is respected. For APOE4 and gender, as they are discrete variables, it is possible to use stratification; in other words, split the dataset into subgroups depending on APOE4 and gender status, and run a Cox PH model on each subgroup.

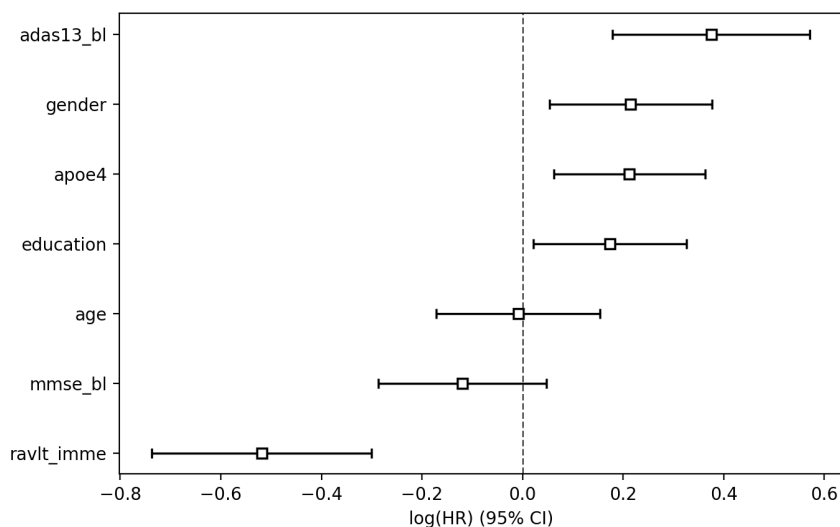


FIGURE 6.1: Coefficients learnt for the Cox PH model, with confidence intervals

		test_statistic	p	$-\log_2(p)$
adas13_bl	km	2.93	0.09	3.53
	rank	2.51	0.11	3.14
age	km	2.06	0.15	2.72
	rank	3.30	0.07	3.85
apoe4	km	8.82	<0.005	8.39
	rank	9.99	<0.005	9.31
education	km	6.48	0.01	6.52
	rank	5.80	0.02	5.96
gender	km	5.85	0.02	6.00
	rank	5.64	0.02	5.83
mmse_bl	km	0.15	0.70	0.52
	rank	0.17	0.68	0.56
ravlt_imme	km	1.60	0.21	2.28
	rank	1.10	0.29	1.77

TABLE 6.2: Testing the proportional hazard hypothesis

Figure 6.2 provides the coefficients for the log-logistic model. We can notice the sign for all coefficients is exactly the opposite of the coefficients in the Cox model; which remains still coherent. We also provide in figure 6.3 the survival function and cumulative hazard function estimated by the parametric log-logistic model. Although the curve seems close to the non-parametric estimators, they diverge for later dates of conversion.

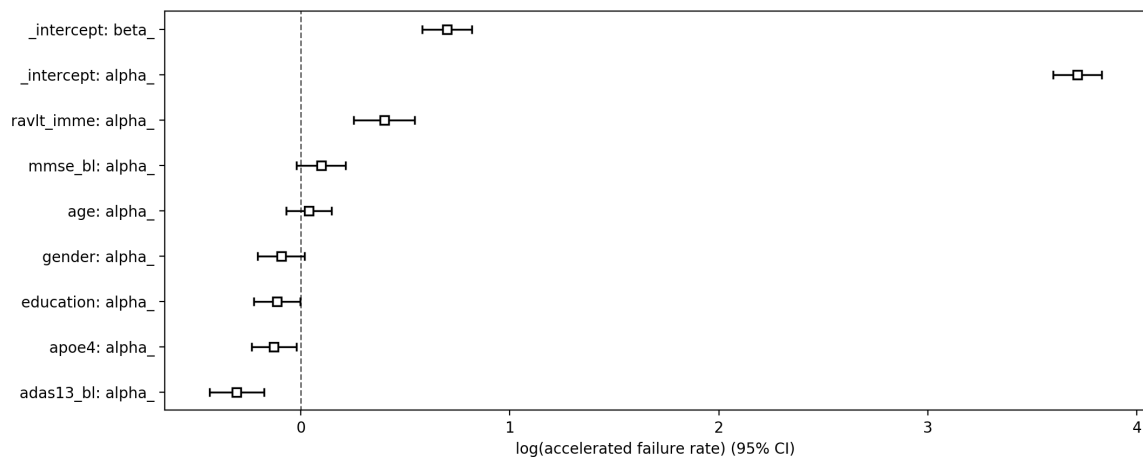


FIGURE 6.2: Coefficients of the log-logistic model

On figure 6.4, are shown the cumulative regression functions in the Aalen additive model as a function of time and the evolution of the parameter vector β and the intercept β_0 . Although both models are different, we can notice that there are similarities between the variations and the signs for each parameter corresponding to the same covariate.

6.2.3 Assessment of the predictive value

To assess the predictive value for each model, and to be able to compare different families of methods, we will use the Area Under the ROC Curve (AUC) and the balanced accuracy

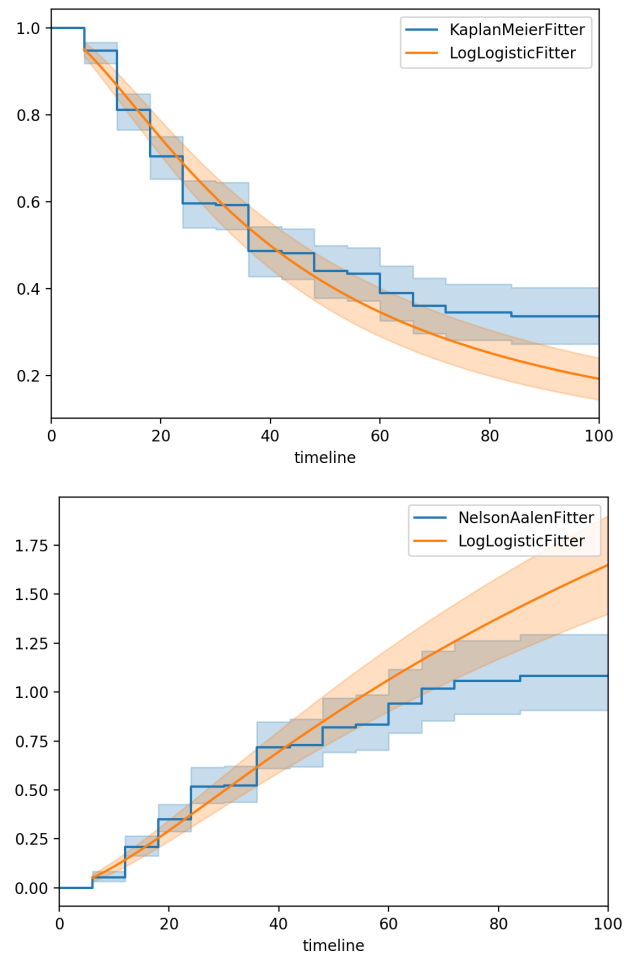


FIGURE 6.3: Comparison between the survival function estimated using the Kaplan-Meier estimator and the log-logistic regression; and between the cumulative hazard estimated using the Nelson-Aalen estimator and the log-logistic regression

(BAC) as in chapter 5. We will use also the concordance index (C-index) defined in [Steck et al., 2008], that verifies if the predicted order between dates coincides with the ground truth. We performed a stratified shuffle split with 10 splits and a test size of 0.3. All the results are shown on figure 6.5.

6.2.4 Conclusion

Both survival models and classification models provide similar performances when it comes to predict the conversion date to Alzheimer's Disease: they have similar AUC (in terms of classification), C-index, although survival models are slightly better on the balanced accuracy metric. The coefficients learnt in each model are coherent, and do respect the direction of variation to the disease. Although the Aalen additive model and logistic regression at fixed time T are two different models, it can be noticed that the variations and signs of each parameter are identical.

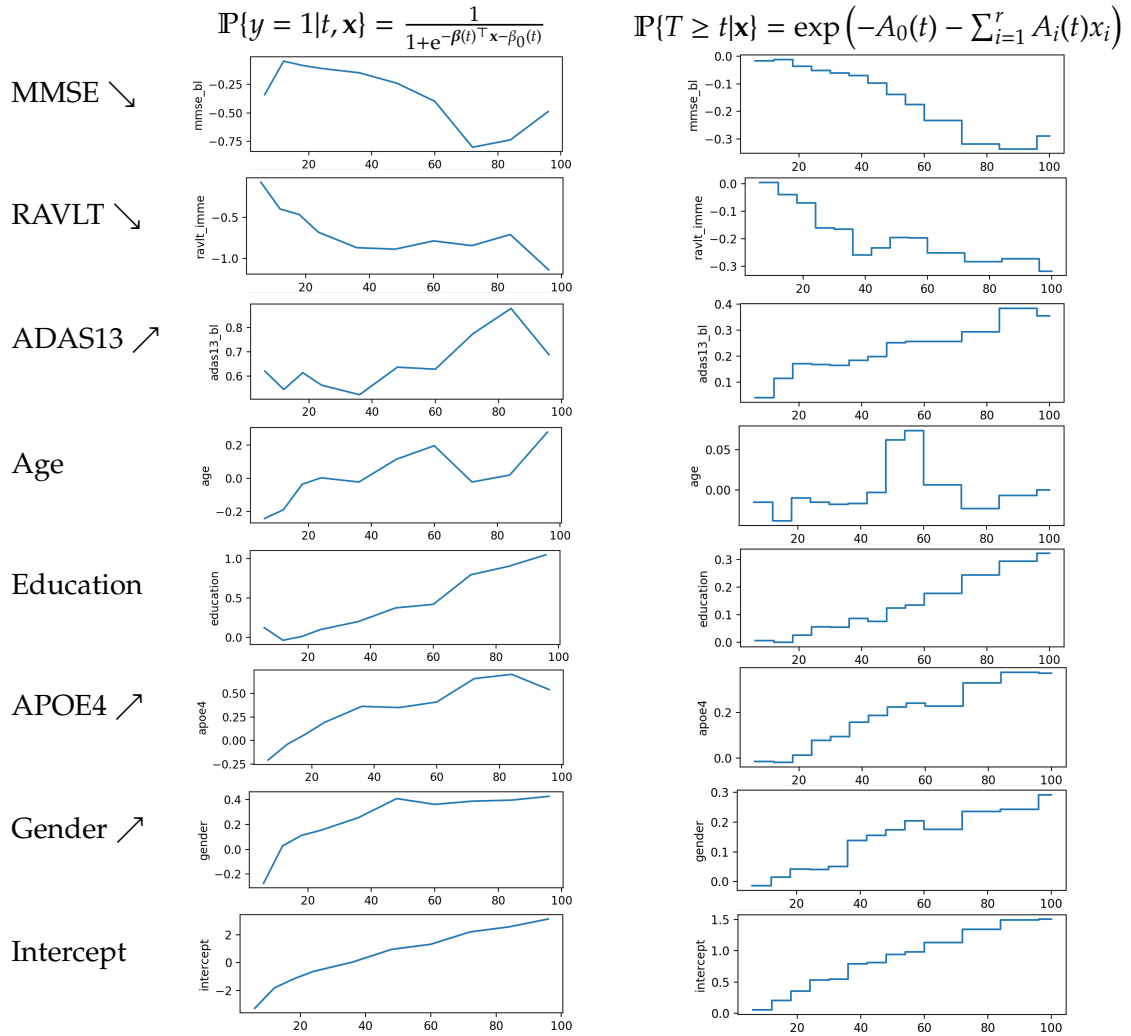


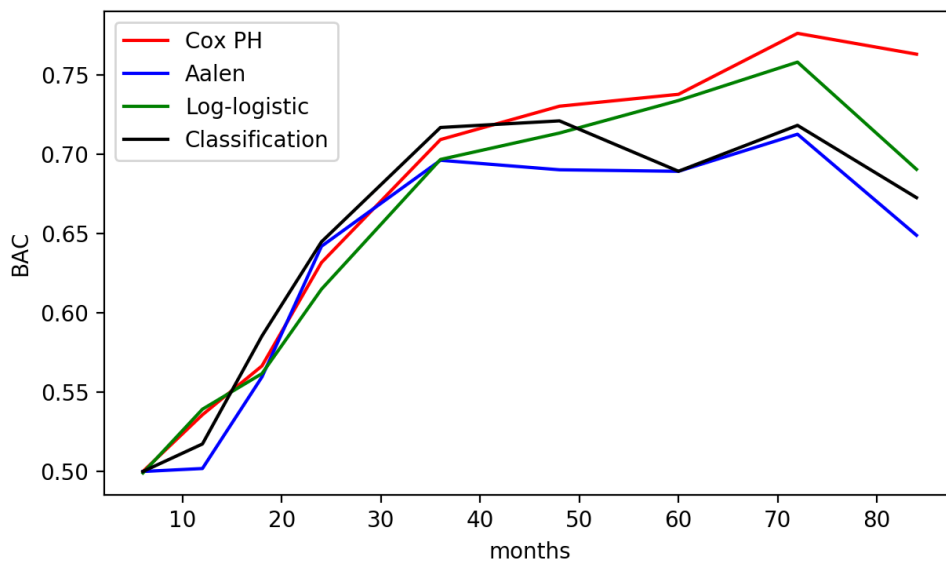
FIGURE 6.4: Coefficients of the parameter vector β of the logistic regression at fixed time t (left); Cumulative regression functions in the Aalen additive model (right)

6.3 A multilevel log-logistic model

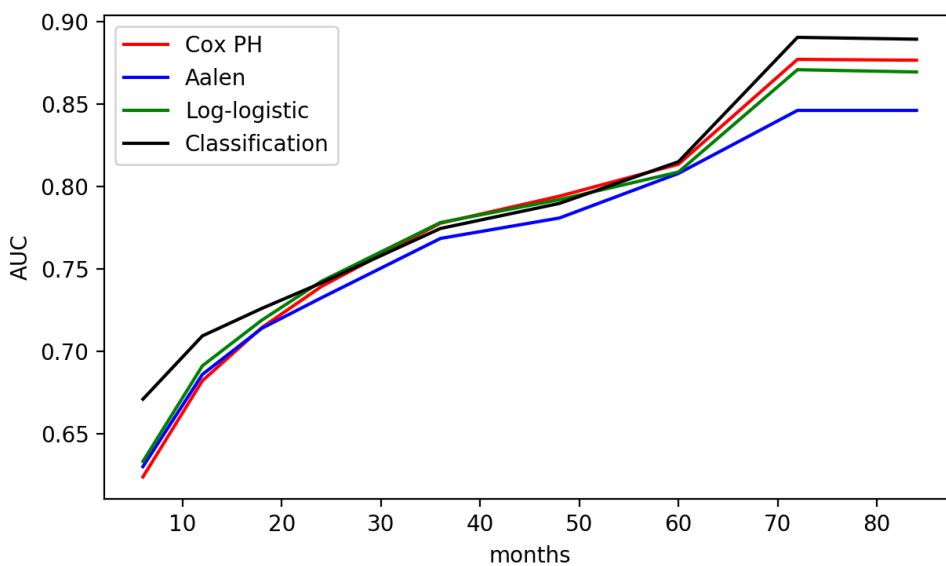
6.3.1 Motivation

Clinical biomarkers and genetic data do not provide the same type of information regarding a subject. On the one hand, clinical biomarkers provide different measurements reflecting the current subject's state. On the other hand, genetic data provide information regarding the characteristic of subject and can help to identify if the subject can develop AD in the future. Therefore, putting them on the same level, for instance like in the Cox model, could not be optimal. On the contrary, genetic data only influences the speed at which the conversion would happen, whereas clinical data influences the initial state of the subject.

Survival function depending on APOE status We split the original MCI datasets into three sub-groups depending on the number of allele ε_4 in gene APOE, and we ran the Kaplan-Meier estimators on these different sub-groups, shown on figure 6.6. The median survival time is 66 months for APOE 0, 36 months for APOE 1, and 24 months for APOE 2.



(A) Balanced Accuracy for different models as function of time



(B) AUC for different models as function of time

	C-index
Cox proportional hazard model	0.720 ± 0.031
Aalen additive model	0.702 ± 0.028
Log-logistic model	0.723 ± 0.031
Logistic regression at fixed time	0.721 ± 0.033

(C) Concordance index for different models

FIGURE 6.5: Assessment of the predictive value for the different models: Balanced Accuracy (top), AUC (middle), C-index (bottom)

The different survival functions start at the same point $(1, 0)$, but the the speed of conversion is clearly different depending on APOE status.

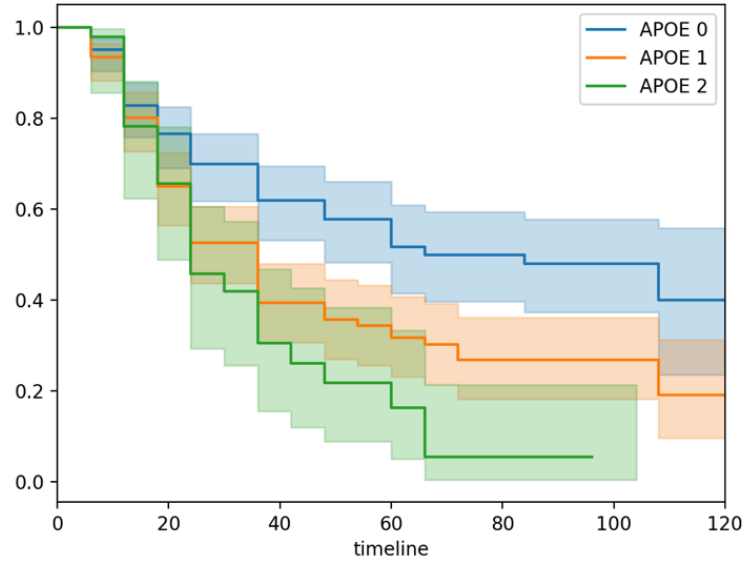


FIGURE 6.6: Estimator of Kaplan-Meier of the survival function $S(t) = \mathbb{P}\{T > t | \text{APOE}\}$

Difference with the original log-logistic model We chose the log-logistic model which provides a parametric framework where the parameters depends on both clinical and genetic data. Furthermore, we expect the hazard function to be unimodal, in other words $p > 1$.

$$S(t|\mathbf{x}) = \frac{1}{1 + \lambda t^p} \quad \text{with } \lambda > 0 \text{ and } p > 0$$

The standard log-logistic model parametrise the parameters as a function of covariates, but would put them on the same level. On the contrary, the multilevel log-logistic model specifies a speed of conversion (depending on genetic covariates \mathbf{x}_G) and the initial state of the subject (depending on clinical covariates \mathbf{x}_C).

6.3.2 Parametrisation

Let \mathbf{x} be the covariates. We expect the survival function taking the following form:

$$S(t|\mathbf{x}) = \frac{1}{1 + \lambda(\mathbf{x})t^{p(\mathbf{x})}} \quad \text{with } \lambda(\mathbf{x}) > 0 \text{ and } p(\mathbf{x}) > 0$$

For instance, we can choose $\lambda(\mathbf{x}) = e^{\beta^\top \mathbf{x}}$ and $p(\mathbf{x}) = e^{\gamma^\top \mathbf{x}}$

However, based on the motivation, we expect that the genetic data modifies the speed at which the subject becomes AD, and that the clinical data translates the curve, indicating therefore the initial position. We define the following variables $v(\mathbf{x}_G) = \beta_G^\top \mathbf{x}_G$ (speed) and $p(\mathbf{x}_I) = \beta_I^\top \mathbf{x}_I$ (position).

We chose to set the median time $t_{1/2}$ (the time when $S(t) = \frac{1}{2}$) such that:

$$t_{1/2} = \frac{v(\mathbf{x}_G)}{p(\mathbf{x}_I)} = \frac{\beta_G^\top \mathbf{x}_G}{\beta_I^\top \mathbf{x}_I}$$

and at the median time $t_{1/2}$, the slope is

$$S'(t_{1/2}) = -v(\mathbf{x}_G) = -\beta_G^\top \mathbf{x}_G$$

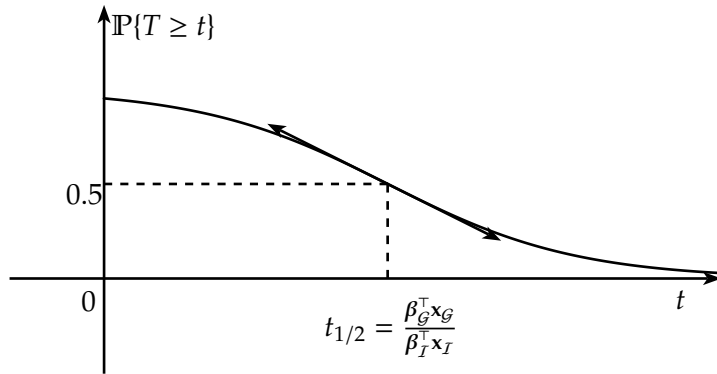


FIGURE 6.7: Log-logistic model

The conditions $\beta_G^\top \mathbf{x}_G > 0$ (all subjects convert) and $\beta_I^\top \mathbf{x}_I > 0$ (conversion date is positive) must always be verified.

We start by defining the median time from the model

$$S(t) = \mathbb{P}\{T \geq t\} = \frac{1}{1 + \lambda t^p} \Leftrightarrow t = \left(\frac{1}{S(t)} - 1 \right)^{\frac{1}{p}} \frac{1}{\lambda^{\frac{1}{p}}}$$

Then, $S(t_{1/2}) = 0.5$ gives the following median survival time:

$$t_{1/2} = \frac{1}{\lambda^{\frac{1}{p}}}$$

Furthermore, the density f of T is given by:

$$f(t) = -S'(t) = \frac{\lambda p t^{p-1}}{(1 + \lambda t^p)^2}$$

When $S(t_{1/2}) = 0.5$, then:

$$\begin{cases} t_{1/2} &= \frac{1}{\lambda^{\frac{1}{p}}} \\ S'(t_{1/2}) &= -p \frac{\lambda^{\frac{1}{p}}}{4} = -\frac{p}{4t_{1/2}} \end{cases}$$

Therefore:

$$\begin{cases} p &= -4t_{1/2} S'(t_{1/2}) \\ \lambda &= \frac{1}{t_{1/2}^p} = t_{1/2}^{4t_{1/2} S'(t_{1/2})} \end{cases}$$

Since we assume that the conversion date is $t_{1/2} = \frac{\beta_I x_I}{\beta_G x_G}$ and $S'(\tau) = -\beta_G x_G$. Then:

$$\begin{cases} p = 4\beta_I x_I \\ \lambda = \left(\frac{\beta_G x_G}{\beta_I x_I}\right)^{4\beta_I x_I} = \exp\left(4\beta_G x_G \ln\left(\frac{\beta_G x_G}{\beta_I x_I}\right)\right) \end{cases}$$

Finally, the survival function is given by:

$$S(t|x_G, x_I) = \mathbb{P}\{T \geq t\} = \frac{1}{1 + \exp\left(4\beta_I x_I \left(\ln t + \ln\left(\frac{\beta_G x_G}{\beta_I x_I}\right)\right)\right)}$$

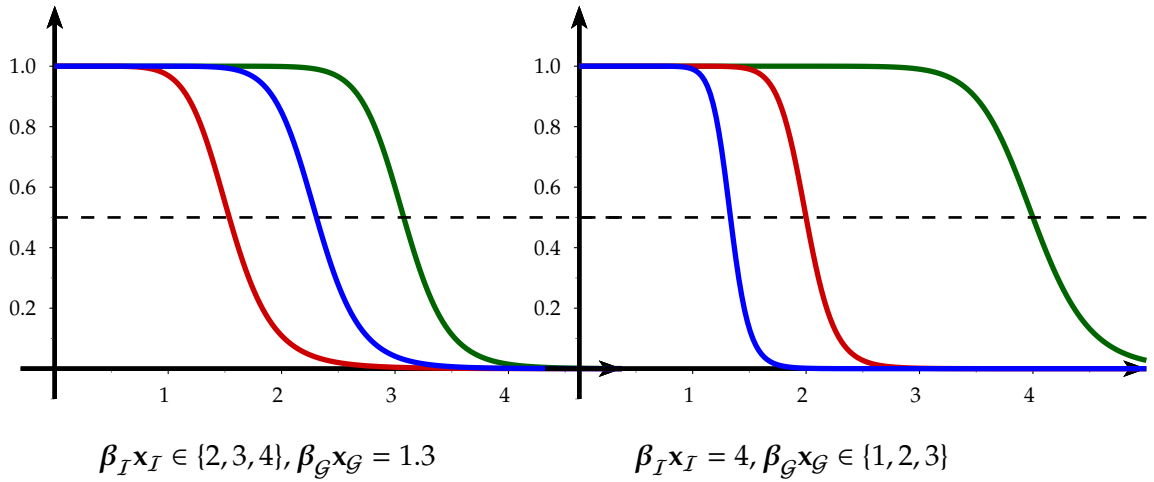


FIGURE 6.8: Effect of $\beta_I x_I$ and $\beta_G x_G$ on $S(t)$. It can be seen that $\beta_I x_I$ only translates the curve of $S(t)$ (for fixed $\beta_G x_G$), whereas $\beta_G x_G$ changes both the shape and translation of the curve of $S(t)$

6.3.3 Model fitting

If n denotes the number of subjects, the likelihood is given by:

$$L(\beta_I, \beta_G) = \prod_{i=1}^n f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

where $\delta_i \in \{0, 1\}$ indicates if the conversion date is observed for a subject i . We set $u = \ln t$. The likelihood becomes:

$$L(\beta_I, \beta_G) = \prod_{i=1}^n f(u_i)^{\delta_i} S(u_i)^{1-\delta_i} = \prod_{i=1}^n \frac{\left(4\beta_I x_I^i \exp\left(4\beta_I x_I^i \left(u_i - \ln\left(\frac{\beta_I x_I^i}{\beta_G x_G^i}\right)\right)\right)\right)^{\delta_i}}{\left(1 + \exp\left(4\beta_I x_I^i \left(u_i - \ln\left(\frac{\beta_I x_I^i}{\beta_G x_G^i}\right)\right)\right)\right)^{\delta_i+1}}$$

The log-likelihood is given by:

$$-\log L(\beta_I, \beta_G) = \sum_{i=1}^n -\delta_i \ln 4 - \delta_i \ln(\beta_I x_I^i) - 4\delta_i \beta_I x_I^i \left(u_i - \ln \left(\frac{\beta_I x_I^i}{\beta_G x_G^i} \right) \right) \\ + (\delta_i + 1) \ln \left(1 + \exp \left(4\beta_I x_I^i \left(u_i - \ln \left(\frac{\beta_I x_I^i}{\beta_G x_G^i} \right) \right) \right) \right)$$

We are searching for a minimum of $-\log L$ on

$$\mathcal{D} = \{\beta_I, \beta_G | \forall i \in \{1, \dots, n\}, \beta_I x_I^i > 0, \beta_G x_G^i > 0\}$$

In two dimensions (if we ran the model with only one covariate for instance), the gray zone in figure 6.9 represents \mathcal{D} .

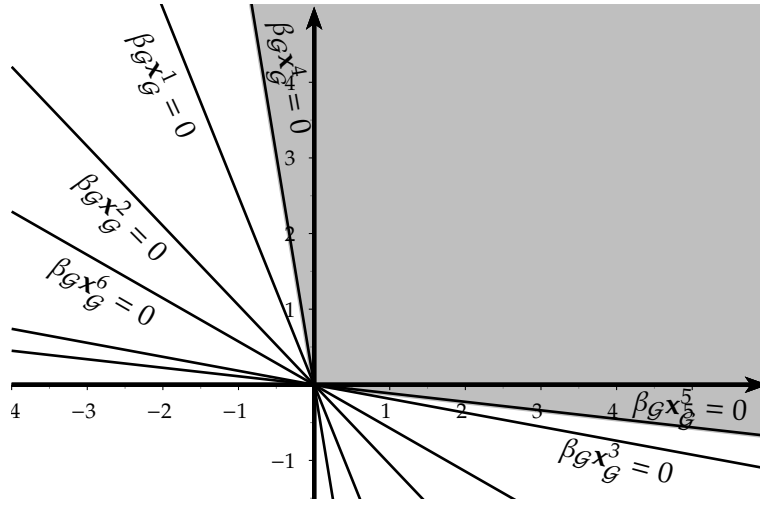


FIGURE 6.9: Example of domain \mathcal{D} in two-dimensions

We used a modified gradient descent to respect the constraints.

Algorithm 2: Training the multilevel log-logistic model

```

1 Initialization:  $\delta = 0.005, \beta_I = \delta \mathbf{1}, \beta_G = \delta \mathbf{1}$ ;
2 while not(converged) do
3    $\varepsilon = \varepsilon_0$ ;
4    $\begin{pmatrix} \beta_I^{new} \\ \beta_G^{new} \end{pmatrix} \leftarrow \begin{pmatrix} \beta_I \\ \beta_G \end{pmatrix} - \varepsilon \nabla(-\log L(\beta_G, \beta_I));$ 
5   while  $(\beta_I^{new}, \beta_G^{new}) \notin \mathcal{D}$  do
6      $\varepsilon \leftarrow \frac{\varepsilon}{2}$ ;
7      $\begin{pmatrix} \beta_I^{new} \\ \beta_G^{new} \end{pmatrix} \leftarrow \begin{pmatrix} \beta_I \\ \beta_G \end{pmatrix} - \varepsilon \nabla(-\log L(\beta_G, \beta_I));$ 
8   end
9 end
```

6.3.4 Experiments and results

Simulated data We started by testing our model on simulated data. They were created using the following procedure:

- simulated clinical data \mathbf{x}_C are sampled through a normal law $\mathcal{N}(\mathbf{m}_C, \Sigma_C)$ where \mathbf{m}_C is the mean for the clinical covariates of ADNI1, and Σ_C the covariance matrix for the clinical data of ADNI1.
- simulated genetic data \mathbf{x}_G are sampled through a discrete law where the probabilities for an allele is the frequency of this allele in the dataset ADNI1.
- the simulated conversion date T is chosen as the median date estimated by the multilevel log-logistic model.

The model has a C-index of 1, an iAUC of 1, and the median survival time was the same as expected (36 months). Because data are simulated through the model, we can expect that the learnt model returns the correct results.

We selected the following covariates:

Effects of covariates We ran our model on the ADNI1 dataset. On the contrary of previous models, data must not be normalised and centred to learn the multilevel log-logistic model.

For the multilevel log-logistic model, when $\beta_I^\top \mathbf{x}_I$ increases, then the conversion date $t_{1/2}$ increases. On the contrary, when $\beta_G^\top \mathbf{x}_G$ increases, then $t_{1/2}$ decreases.

For the Cox model, when $\beta_I^\top \mathbf{x}_I$ or $\beta_G^\top \mathbf{x}_G$ increases, then the conversion date $t_{1/2}$ decreases.

Table 6.3 shows the coefficients learnt by a Cox model and by the multilevel log-logistic model. The genetic features gender and APOE seem to contribute in both models in the estimation of the conversion date, whereas the age does not seem to be a relevant feature. We can also notice that as ADAS 13, increases when the subject progress to AD, the corresponding coefficient is positive. On the contrary, the MMSE and the RAVLT immediate decrease when the subject progress to AD, and the coefficients for both models are negative.

Coefficient	Our model	Normalised Cox PH
MMSE (at baseline) ↘	0.0044	-0.12
ADAS13 (at baseline) ↗	-0.0097	0.36
Age ↗	0.00084	-0.012
Education	-0.0025	0.16
RAVLT (at baseline) ↘	0.0098	-0.51
Intercept (clinics)	0.22	
APOE4 ↗	0.0037	0.20
Gender ↗	0.00082	0.19
Intercept (genetics)	0.0083	

TABLE 6.3: Coefficients for the Cox model and log-logistic model (↗ means that the higher this covariates is, the risk to have AD increases; whereas ↘ means that the higher this covariates is, the risk to have AD decreases)

Assessment of the predictive value

The estimated survival function $t \mapsto \widehat{S}(t)$ and hazard function $t \mapsto \widehat{h}(t)$ learnt by both models are shown in figure 6.10 (in orange for the logistic model, in blue for the non-parametric estimator). We can see that the learnt curve fits well on earlier conversion dates, but does not fit well for later conversion date. In particular, the estimated hazard for later conversion date is much higher than the Nelson-Aalen estimated hazard.

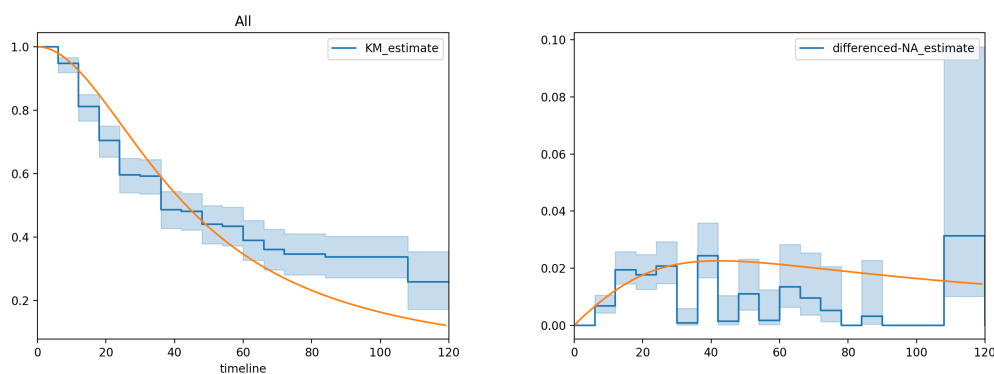


FIGURE 6.10: Estimated survival function \widehat{S} for the Kaplan Meier estimate and for the multilevel log-logistic model; Estimated hazard function \widehat{h} for the Nelson-Aalen estimated hazard and for the multilevel log-logistic model

To assess the predictive value of our model, we perform a stratified shuffle split with 10 splits and a test size of 0.2.

As the conversion date T is in the set $\{6, 12, 18, 24, 36, 42, 54, 60, 72\}$, we set the intervals $I_0 = [0, 6[, I_1 = [6, 12[, I_2 = [12, 18[, I_3 = [18, 24[, I_4 = [24, 36[, I_5 = [36, 42[, \dots$

Table 6.4 shows the results for the Cox model and the multilevel log-logistic model. The AUC for each date is given in table 6.5.

	Cox PH	Multilevel log-logistic
Median survival time (36 months)	36	43.4
C-index	0.72	0.65
iAUC	0.84	0.81

TABLE 6.4: Average results on ADNI1

Months	6	12	18	24	36	42	60
Cox PH Model	0.67	0.71	0.79	0.83	0.83	0.83	0.86
Multilevel log-logistic Model	0.63	0.69	0.77	0.79	0.79	0.79	0.82

TABLE 6.5: AUC for each date

All the metrics (C-index, iAUC, $AUC(t)$) show that the Cox model predictive value is stronger than the multilevel log-logistic model. We also notice that the median survival time of the multilevel log-logistic model is positively biased.

Evaluation on group based on APOE status We can also assess the model on groups of subject depending of their APOE status. The median survival times are shown on Figure 6.11, and we plotted the average trajectories for these groups.

	Kaplan-Meier	Multilevel log-logistic
Median survival time (months)	36	43.4
Median survival time for APOE 0 (months)	66	55.8
Median survival time for APOE 1 (months)	36	39.3
Median survival time for APOE 2 (months)	24	30.4

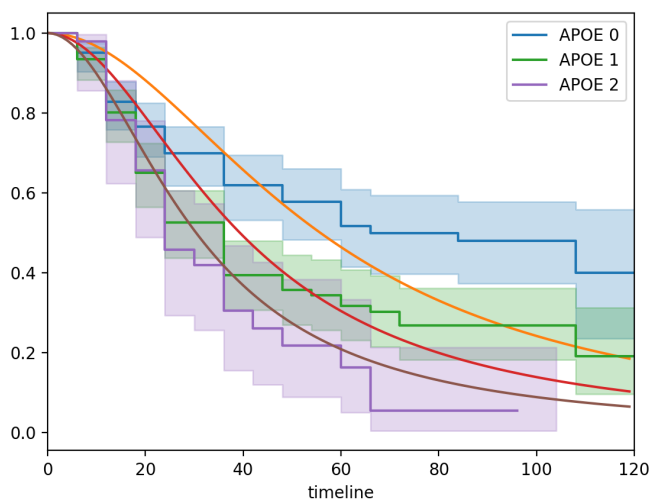


FIGURE 6.11: Median survival time, estimated survival function \widehat{S} for the Kaplan Meier estimate and for the multilevel log-logistic model

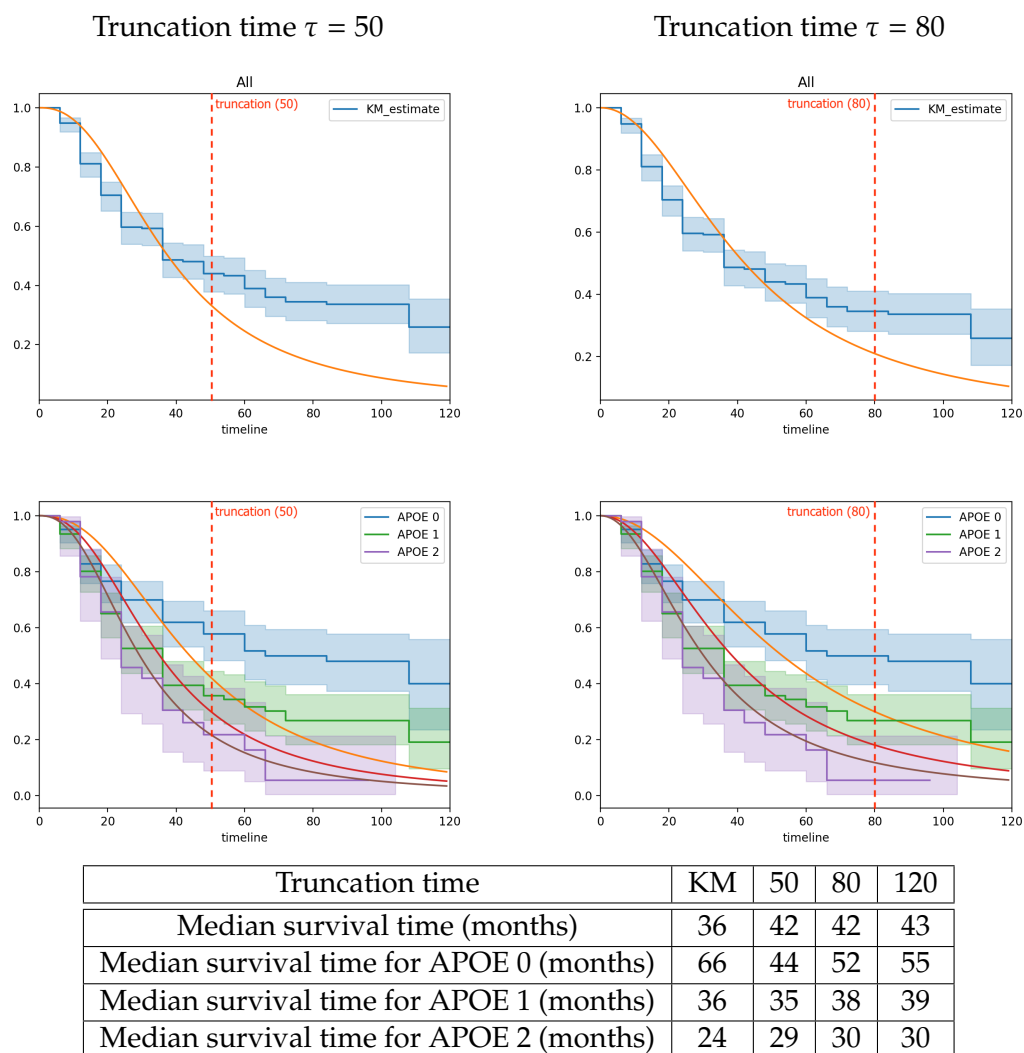
On the Kaplan-Meier estimates, we notice that the order between the three curves is respected. Furthermore, the distance between the curves for APOE 0 and APOE 1 is larger than the distance between the curves for APOE 2 and APOE 1.

As for global curve, the subgroup curves for log-logistic model does not fit well on later conversion dates. Furthermore, the curves seem to be better fitted for APOE 2 and APOE 1. There is a positive bias for APOE 1 and 2, and negative bias for APOE 0.

A possible explanation is that the model can not easily change its slope. In particular, the distance with the average curve is not large enough, and therefore, the influence of APOE is not strong enough.

Effect of truncation on model fitting Instead of learning the model on the original dataset, subjects with later conversion time can be truncated. More specifically, let τ be the follow-up date. If the subject i 's conversion is $T_i > \tau$, then we set his observed date is τ , and the conversion did not occur.

The effect of truncation makes the prediction more accurate for small conversion date, but more inaccurate for later conversion date.

FIGURE 6.12: Results for truncation time $\tau \in \{50, 80\}$

6.3.5 Discussion

The multilevel log-logistic model provides a parametric framework where the parameters depend on both clinical and genetic data, but does not consider. On the contrary of the Cox model and the original log-logistic model, clinical and genetic data on the same level. On the contrary, the multilevel log-logistic model defines a speed of conversion and the initial state of the subject.

On simulated data, the results show that the model can learn and reconstitute the correct values. On real data from ADNI1, individual predictions are slightly worse than the Cox model. The parameters given by both models are consistent, as the signs of the parameters are identical.

We notice a negative bias on the median survival time for the multilevel log-logistic model. The reason is mainly due to the log-logistic model itself; as shown on figure 6.3 with another parametrisation, the log-logistic model does not fit well for later conversion data for this dataset.

We also performed further analyses such as the effect of the gene APOE on the model,

and the effect of truncation on the model's performance. In particular, the experiments on subgroups of APOE show there is a positive bias for APOE 1 and 2, and negative bias for APOE 0. An explanation is that the model estimation for genetics can not change properly the slope (and the conversion speed), in particular for the subgroup APOE 0. Furthermore, the gap between the survival curve for APOE 0 and the survival curve for APOE 2 from survival curve for APOE 1 is not identical; as there are much less converters for the subgroup APOE 0.

The truncation shows that the model is more accurate on earlier conversion date, and less accurate on later conversion dates. A possible explanation is that there are less data and less conversions for later conversion dates.

6.4 Conclusion

In the first part of this chapter, we compared different survival models, and a typical classification model at fixed time T for the prediction of conversion to Alzheimers Disease. This models (classification at fixed time T , Cox proportional hazard model, Aalen additive model, log-logistic model) provide similar results and their coefficients are equivalent.

In the second part of this chapter, we developed a multilevel log-logistic model, where the coefficients of the model define a speed of conversion and an initial state. This model is simple to implement and test. However, it has a strong bias for long conversion dates, due to the log-logistic model, and strong biases for subgroups of population based on their APOE status.

Chapter 7

Multilevel Cox Proportional hazard Model with Structured Penalties for Imaging Genetics data

This chapter is a modified version of:

Lu Pascal, Colliot Olivier. Multilevel Survival Analysis with Structured Penalties for Imaging Genetics data. Accepted at *SPIE 2020 Medical Imaging*.

7.1 Introduction

Predicting the future occurrence of Alzheimer's Disease (AD) in patients with mild cognitive impairment (MCI) is a topic of active research. Many papers have formulated this question as a classification problem: one considers a fixed time of conversion and aims to discriminate between the patients who have converted to AD at that time and those who have not. However, a clinically more relevant question is to predict the date at which a patient will convert to AD. Survival analysis is an adequate statistical framework for such a task.

Multimodal data (imaging and genetic) provide complementary information for the prediction. While imaging data provides an estimate of the current patient's state, genetic variants can be associated to the speed of progression to AD. Although they do not provide the same type of information, most papers in classification or regression put imaging and genetic variables on the same level in order to predict the current or future patient's state.

In this chapter, we propose a survival model using multimodal data to estimate the conversion date to AD, by considering joint effects between the imaging and genetic modalities. We use an adapted penalty in the survival model, the group lasso penalty, over joint groups of genes and brain regions, as in the chapter

The model is evaluated on genetic (single nucleotide polymorphisms) and imaging (anatomical MRI measures) data from the ADNI1 database, and compared to a standard Cox PH model.

7.2 State of the art

Early diagnosis of Alzheimer’s disease (AD) is an active issue in medical imaging. In Alzheimer’s disease, the group of patients with mild cognitive impairment is heterogeneous, some are more likely than others to convert to AD; and therefore giving an accurate diagnosis can be difficult [Marinescu et al., 2018]. In this paper, we focus on predicting accurately the conversion to AD given only one time-point.

Most existing approaches to predict the conversion define the problem as a classification task at fixed time [Rathore et al., 2017, Cheng et al., 2015, Davatzikos et al., 2011]. The problem of using classification at a fixed time is that we have to arbitrarily choose a date at which the conversion to AD is observed, we create two groups of homogeneous patients (converting before or after the fixed date), and we do not ensure that conversion in time is monotonous.

Survival models provide a regression framework which directly estimates the conversion date using only one time-point. However, by using survival models, we make several hypotheses: the conversion event is sometimes not observed (because the study ended before conversion), sometimes never occurs (some patients will never convert to AD), and only happens once. For all patients in the study, the starting point is the entry in the study, and the conversion is the time of progression of the disease to AD.

In a single visit, several kind of data can be acquired (clinical data, neuroimages, fluid biomarkers, genetic data. . .). Some data, such as clinical data or neuroimaging data provide a picture of the patient’s state at the time they were acquired [Rathore et al., 2017]. On the contrary, others data, such as genetic data, help to identify whether or not a patient could develop AD in the future (for instance, some alleles of APOE increase the risk of developing AD). An issue raised by collecting data from different sources concerns their combination. In the area of imaging genetics, most papers focus either on the association between neuroimaging covariates and genetic data [Liu and Calhoun, 2014, Batmanghelich et al., 2016], or on building machine learning predictors for a disease at fixed time using classification (logistic regression, SVM [Peng et al., 2016]). All these models combine neuroimaging and genetic data by using an additive framework. However, adding the effect of both modalities and putting them on the same level is not optimal, as these modalities do not provide the same type of information. We proposed a multilevel framework for combining imaging and genetic data for classification.

In this paper, we propose a survival model, based on the Cox Proportional Hazard model and using a multilevel framework, as in chapter 4. Survival models (Cox Proportional Hazard model) have been applied for combining multimodal data [Bøvelstad et al., 2009] and for predicting the conversion to AD [Li et al., 2017, Anderson et al., 2016, Liu et al., 2017]; in both case using an additive framework. Learning the conversion date to AD with modalities taken separately shows that the genetic modality has weaker predictive value than the neuroimaging modality. Instead of summing both genetic and neuroimaging contributions (which could lead to a weaker contribution of the genetic modality in the model), we propose that the parameters, combined with the neuroimaging covariates, could be modulated by the patient’s genetic data. This hypothesis leads to a multilevel model where

genetic data express themselves through interactions with neuroimaging covariates.

Adding interactions leads to high-dimensional models, and adapted penalties for each modality is essential to avoid overfitting. For instance, SNPs can be grouped by genes [Silver et al., 2012] and a group lasso penalty can be applied on the groups formed by genes. In this paper, we will use the group lasso penalty on the interactions for the parameters coupling (genes, brain region). We use a proximal gradient descent algorithm to learn all the parameters. This model is evaluated on genetic single-nucleotide polymorphisms (SNPs) and neuroimaging data (MRI modality) from the ADNI database, and is compared to standard Cox models.

7.3 Methods

7.3.1 Model set-up

We aim to model the time to conversion to Alzheimer's Disease (AD) for MCI patients. For all patients, the starting point is the entry in the study, and the conversion is the time of progression to AD.

Notations For each patient i , we denote T_i^* his real conversion date from MCI to AD, C_i the date of his final visit and $T_i = \min(T_i^*, C_i)$ the duration observed in the study. We introduce $\delta_i = \mathbb{I}\{T_i^* \leq C_i\}$ indicating if the conversion has occurred.

We denote $\mathbf{x}_{\mathcal{G}}$ the vector of single-polymorphism nucleotides (SNP) counted by number of minor variants, \mathbf{x}_I the vector of imaging variables (brain regions), $|\mathcal{G}|$ the number of SNPs and $|I|$ the number of imaging variables.

The conversion date T is a continuous random variable with cumulative distribution function $F : t \mapsto \mathbb{P}\{T < t | \mathbf{x}_{\mathcal{G}}, \mathbf{x}_I\} = 1 - \exp\left(-\int_0^t h(u | \mathbf{x}_{\mathcal{G}}, \mathbf{x}_I) du\right)$ where h is the hazard function, representing the instantaneous rate of occurrence of the event.

Multilevel framework We propose the multilevel framework, based on the Cox proportional hazard assumption, defined by $h(t | \mathbf{x}_{\mathcal{G}}, \mathbf{x}_I) = h_0(t) e^{\beta(\mathbf{x}_{\mathcal{G}})^{\top} \mathbf{x}_I}$, where $\beta(\mathbf{x}_{\mathcal{G}})$ is the parameter vector depending on genetic data $\mathbf{x}_{\mathcal{G}}$ and h_0 is the baseline hazard function describing the risks for individuals whose covariates are null.

We make the assumption that β is an affine function depending on genetic data: $\beta(\mathbf{x}_{\mathcal{G}}) = \mathbf{W}^{\top} \mathbf{x}_{\mathcal{G}} + \beta_I$, where $\mathbf{W} \in \mathcal{M}_{|\mathcal{G}|, |I|}(\mathbb{R})$. Then,

$$h(t | \mathbf{x}_{\mathcal{G}}, \mathbf{x}_I) = h_0(t) e^{(\mathbf{x}_{\mathcal{G}})^{\top} \mathbf{W} \mathbf{x}_I + \beta_I^{\top} \mathbf{x}_I}$$

The survival function is given by:

$$S(t | \mathbf{x}_{\mathcal{G}}, \mathbf{x}_I) = (S_0(t))^{(\mathbf{x}_{\mathcal{G}})^{\top} \mathbf{W} \mathbf{x}_I + \beta_I^{\top} \mathbf{x}_I} \text{ where } S_0(t) = \exp\left(-\int_0^t h_0(u) du\right)$$

The effect of the covariate $(\mathbf{x}_{\mathcal{G}}, \mathbf{x}_I)$ on the survival function is to raise it to a power given by the prognostic index $\text{PI}(\mathbf{x}_{\mathcal{G}}, \mathbf{x}_I) = e^{(\mathbf{x}_{\mathcal{G}})^{\top} \mathbf{W} \mathbf{x}_I + \beta_I^{\top} \mathbf{x}_I}$.

The genetic modality has a much weaker predictive power compared to imaging or clinical features. And separately taken, the genetic modality provides poor results (see

table 7.1). By combining SNPs and imaging features in that way, we ensure that SNPs will add a significant contribution to the model.

7.3.2 Optimization

Given the dataset $\{(\mathbf{x}_{\mathcal{G}}^i, \mathbf{x}_I^i, T_i, \delta_i), i = 1, \dots, N\}$ where the covariates $\mathbf{x}_{\mathcal{G}}, \mathbf{x}_I$ and $\mathbf{x}_{\mathcal{G}}\mathbf{x}_I^\top$ are centered and normalized, the negative partial log-likelihood is given by:

$$\ell(\mathbf{W}, \boldsymbol{\beta}_I) = -\frac{1}{N} \sum_{i=1}^N \delta_i \left((\mathbf{x}_{\mathcal{G}}^i)^\top \mathbf{W} \mathbf{x}_I^i + \boldsymbol{\beta}_I^\top \mathbf{x}_I^i - \log \sum_{j \in \mathcal{R}(T_i)} e^{(\mathbf{x}_{\mathcal{G}}^i)^\top \mathbf{W} \mathbf{x}_I^j + \boldsymbol{\beta}_I^\top \mathbf{x}_I^j} \right)$$

where $\mathcal{R}(T_i)$ is the set of patients j such that $T_j \geq T_i$.

Penalties As the number of parameters to estimate is much larger than the number of patients, we need to add penalties on \mathbf{W} and $\boldsymbol{\beta}_I$. For imaging parameters $\boldsymbol{\beta}_I$, we considered the ridge penalty, as Alzheimer's Disease has a diffuse anatomical pattern of alteration. For the matrix \mathbf{W} , we start by mapping SNPs to genes \mathcal{G}_ℓ ($\ell \leq L$, where L is the number of genes), and we use a group lasso with overlap penalty, where groups are (genes, imaging covariate). This penalty enforces sparsity between groups and regularity inside the same group. Finally, we add the following penalty to the negative partial log-likelihood:

$$\Omega(\mathbf{W}, \boldsymbol{\beta}_I) = \lambda \sum_{i=1}^{|\mathcal{I}|} \sum_{\ell=1}^L \sqrt{|\mathcal{G}_\ell|} \|\mathbf{W}_{\mathcal{G}_\ell, i}\|_{\ell_2} + \lambda_I \|\boldsymbol{\beta}_I\|_{\ell_2}^2$$

where $\lambda > 0, \lambda_I > 0$ are the hyperparameters.

The parameters $\mathbf{W}, \boldsymbol{\beta}_I$ are obtained by minimizing the quantity $\ell(\mathbf{W}, \boldsymbol{\beta}_I) + \Omega(\mathbf{W}, \boldsymbol{\beta}_I)$. The usual approach for dealing with the penalty Ω is to use a proximal gradient descent on the convex set defined by Ω [Hastie et al., 2015, Beck and Teboulle, 2009].

Algorithm 3: Optimization procedure

```

1 Input:  $\{(\mathbf{x}_{\mathcal{G}}^i, \mathbf{x}_I^i, T_i, \delta_i), i = 1, \dots, N\}, \delta = 0.5, \varepsilon_0 = 0.01, \eta = 10^{-5}$ ;
2 Initialization:  $\mathbf{W} = \mathbf{0}, \boldsymbol{\beta}_I = \mathbf{0}, \text{converged} = \text{False}$ 
3 while not(converged) do
4    $\boldsymbol{\gamma} = (\text{flatten}(\mathbf{W}), \boldsymbol{\beta}_I)$  and  $\boldsymbol{\omega} = \boldsymbol{\gamma} - \varepsilon \nabla \ell$ ;
5    $\widehat{\mathbf{W}}_{\mathcal{G}_\ell, i} = \max\left(0, 1 - \frac{\varepsilon \lambda_{\mathcal{G}} \theta_{\mathcal{G}_\ell}}{\|\boldsymbol{\omega}_{\mathcal{G}_\ell, i}\|_2}\right) \boldsymbol{\omega}_{\mathcal{G}_\ell, i}$  for  $(i, \ell) \in \llbracket 1, |\mathcal{I}| \rrbracket \times \llbracket 1, L \rrbracket$ ;
6    $\widehat{\boldsymbol{\beta}}_I = \frac{\boldsymbol{\omega}_I + |\mathcal{I}| \lambda_I}{1 + 2\varepsilon \lambda_I}$  (imaging modality);
7   if  $(\ell + \Omega)(\widehat{\mathbf{W}}, \widehat{\boldsymbol{\beta}}_I) > (\ell + \Omega)(\mathbf{W}, \boldsymbol{\beta}_I)$  then
8      $\varepsilon = \delta \varepsilon$ 
9   else
10     $\text{converged} = \left| (\ell + \Omega)(\mathbf{W}, \boldsymbol{\beta}_I) - (\ell + \Omega)(\widehat{\mathbf{W}}, \widehat{\boldsymbol{\beta}}_I) \right| <^? \eta |(\ell + \Omega)(\mathbf{W}, \boldsymbol{\beta}_I)|$ ;
11     $\mathbf{W} = \widehat{\mathbf{W}}, \boldsymbol{\beta}_I = \widehat{\boldsymbol{\beta}}_I, \varepsilon = \varepsilon_0$ ;
12  end
13 end

```

Implementation We flatten the cross-product covariates $\mathbf{x}_G \mathbf{x}_I^\top$ and the matrix \mathbf{W} and transform them into a vector. We create a vector $\boldsymbol{\gamma} = (\text{flatten}(\mathbf{W}), \boldsymbol{\beta}_I)$ containing the coefficients of $\mathbf{W}, \boldsymbol{\beta}_I$. To recreate \mathbf{W} , we just need to unflatten $\boldsymbol{\gamma}$. The vector $\boldsymbol{\gamma}$ is updated using a proximal gradient descent described in Algorithm 3. The stopping criterion for this algorithm is

$$\left| (\ell + \Omega)(\mathbf{W}, \boldsymbol{\beta}_I) - (\ell + \Omega)(\widehat{\mathbf{W}}, \widehat{\boldsymbol{\beta}}_I) \right| < \eta |(\ell + \Omega)(\mathbf{W}, \boldsymbol{\beta}_I)|$$

7.4 Experiments and results

7.4.1 Dataset

We worked with a subset of ADNI1, in which all subjects have been genotyped. This subset contains 326 MCI genotyped patients with 172 MCI patients at baseline who progressed to AD during the study and 150 MCI patients who remained stable (censored data).

Covariates In this dataset, 620,901 SNPs have been genotyped. Based on the 44 first top genes related to AD (from AlzGene, <http://www.alzgene.org>) and on the Illumina annotation using the Genome build 36.2, we select 1,107 SNPs. For cross-validation purposes, SNPs for whose the variance among the dataset is smaller than 0.01 are removed, leading to 679 SNPs. Regarding the MRI modality, we use the segmentation of FreeSurfer which gives the volume of subcortical regions (44 features) and the average thickness in cortical regions (68 features).

Baseline survival function S_0 The baseline survival function S_0 is computed using the Kaplan-Meier estimate. On figure 7.1, is shown on the right the Kaplan-Meier estimated baseline survival function S_0 using the distribution of T^* and C displayed on the left. The follow-up date is $\tau_{\text{hor}} = 100$ months; patients who convert after this date are truncated. The median survival time (the smallest survival time for which the survivor function is less than or equal to 0.5) is 36 months.

7.4.2 Evaluation

Baseline models We compare the multilevel framework to the Cox Proportional Hazard model using one modality or using an additive framework. In this later case, the hazard function for patient i is given by $h(t|\mathbf{x}_G, \mathbf{x}_I) = h_0(t)e^{\boldsymbol{\beta}_G^\top \mathbf{x}_G + \boldsymbol{\beta}_I^\top \mathbf{x}_I}$.

Metrics We define the three following measures to assess the quality of the prediction:

- the concordance index (or C-index) [Steck et al., 2008] which checks if the model orders the conversion dates in the same order as the ground truth. As a generalization of AUC, the range of the C-index is $[0, 1]$, but typical values are between 0.55 and 0.7.

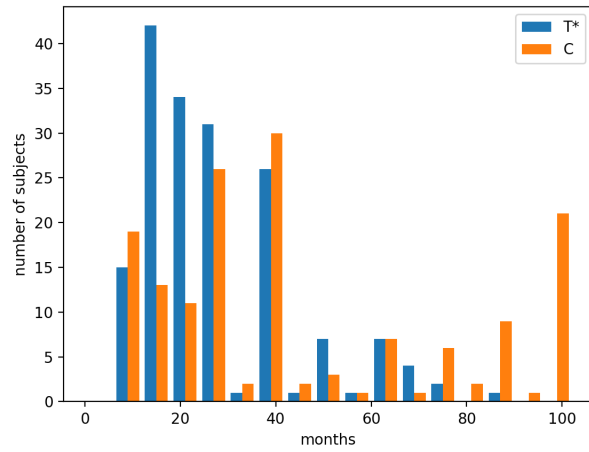
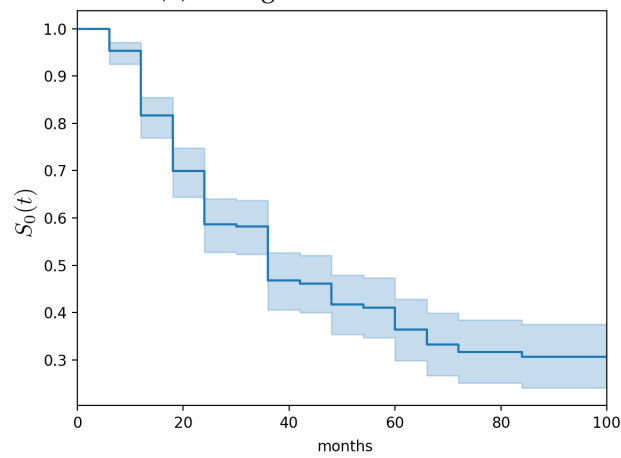
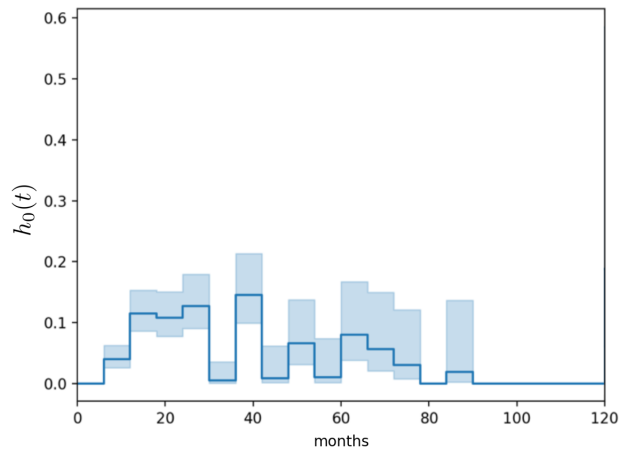
(A) Histogram of T^* and C (B) Kaplan-Meier estimated baseline survival function S_0 (c) Nelson-Aalen estimated hazard function h_0 (bandwidth = 6 months)

FIGURE 7.1: ADNI1 Dataset: baseline survival function and hazard function

- the integrated Brier score [Graf et al., 1999], for uncensored data:

$$\text{Brier} = \frac{1}{\tau_{\text{hor}}} \int_0^{\tau_{\text{hor}}} \left[\frac{1}{N} \sum_{i=1}^N \delta_i (\mathbf{1}(T_i^* > t) - \widehat{S}(t|\mathbf{x}_G, \mathbf{x}_I))^2 \right] dt$$

The Brier score measures the accuracy of probabilistic predictions. The model performs better when the Brier score is lower.

- the integrated Area Under Curve, defined as $iAUC = \int_0^{\tau_{\text{hor}}} AUC(t)f(t)dt$ where τ_{hor} is the follow-up date, f is the probability density function of T and $AUC(t)$ is the cumulative AUC [Chambless and Diao, 2006]. The cumulative/dynamic AUC plays the same role as the classical Area Under Curve in classification. As for the AUC, the range of the iAUC is $[0, 1]$, and the higher the iAUC is, the more predictive the model is.

Cross validation To determine the hyperparameters, we use a nested cross validation. We perform a 5-fold cross validation, and within each fold, we find the optimal hyperparameters using a 5-fold cross validation on the training set and taking the hyperparameters that maximize the C-index over the inner test set.

The hyperparameters are optimized between $\{10^{-4}, 10^{-3}, \dots, 10, 10^2\}$.

Modality	Method	C-index	Brier score	iAUC
SNPs only	Cox PH model (ℓ_1 penalty)	0.521 ± 0.040	0.166 ± 0.009	0.515 ± 0.031
MRI only	Cox PH model (no penalty)	0.636 ± 0.034	0.190 ± 0.017	0.636 ± 0.050
MRI only	Cox PH model (ℓ_2 penalty)	0.671 ± 0.022	0.149 ± 0.008	0.663 ± 0.044
All	Additive Cox PH model (ℓ_1 penalty)	0.677 ± 0.020	0.148 ± 0.006	0.680 ± 0.030
All	Multilevel model (ours)	0.681 ± 0.018	0.147 ± 0.006	0.686 ± 0.031

TABLE 7.1: Results for different modalities and methods (mean value across the test folds \pm standard deviation)

Results on Table 7.1 show that genetic modality, taken alone, have a much weaker predictive value than the imaging modality. The imaging modality already provides good performances, and adding the genetic modality improves the model performance, but not significantly (in both additive and multilevel frameworks). Adding a penalty on the imaging parameter also increases the performances. Finally, the multilevel model provides slightly better results than the additive model.

7.4.3 Effect of cross-product covariates on conversion

For interpretation purposes, we compute a new reduced matrix $\tilde{\mathbf{W}} \in \mathcal{M}_{|I|,L}(\mathbb{R})$, shown on figure 7.2, where for each brain region j and gene ℓ , $\tilde{\mathbf{W}}_{j,\ell} = \max_{s \in \mathcal{G}_\ell} |\mathbf{W}_{s,j}|$.

The matrix on figure 7.2 shows that some rows and columns have more non-null coefficients than others.

The strongest effects are found for the ventricles, which are enlarged in AD but also in aging and other degenerative diseases, and several medial temporal lobes (MTL) structures

(entorhinal cortex, hippocampus, amygdala) which are altered early in AD. It is interesting to note that the ventricles have a strong interaction with all genes except APOE, while the MTL structures have interactions with a more restricted but quite consistent set of genes. Regarding the genes, the strongest effects are found for the single-nucleotide polymorphism rs6503018 (TNK1), rs429358 (APOE) and rs3093662 (TNF).

7.5 Conclusion

We proposed a novel approach to estimate the conversion date to AD for MCI patients, using the Cox Proportional Hazard model, from genetic and neuroimaging data. On the contrary of additive models, the multilevel model captures interactions between genes and brain regions. The use of adapted penalties avoids overfitting by providing a sparse matrix and highlighting brain regions and genes both related to the progression to AD.

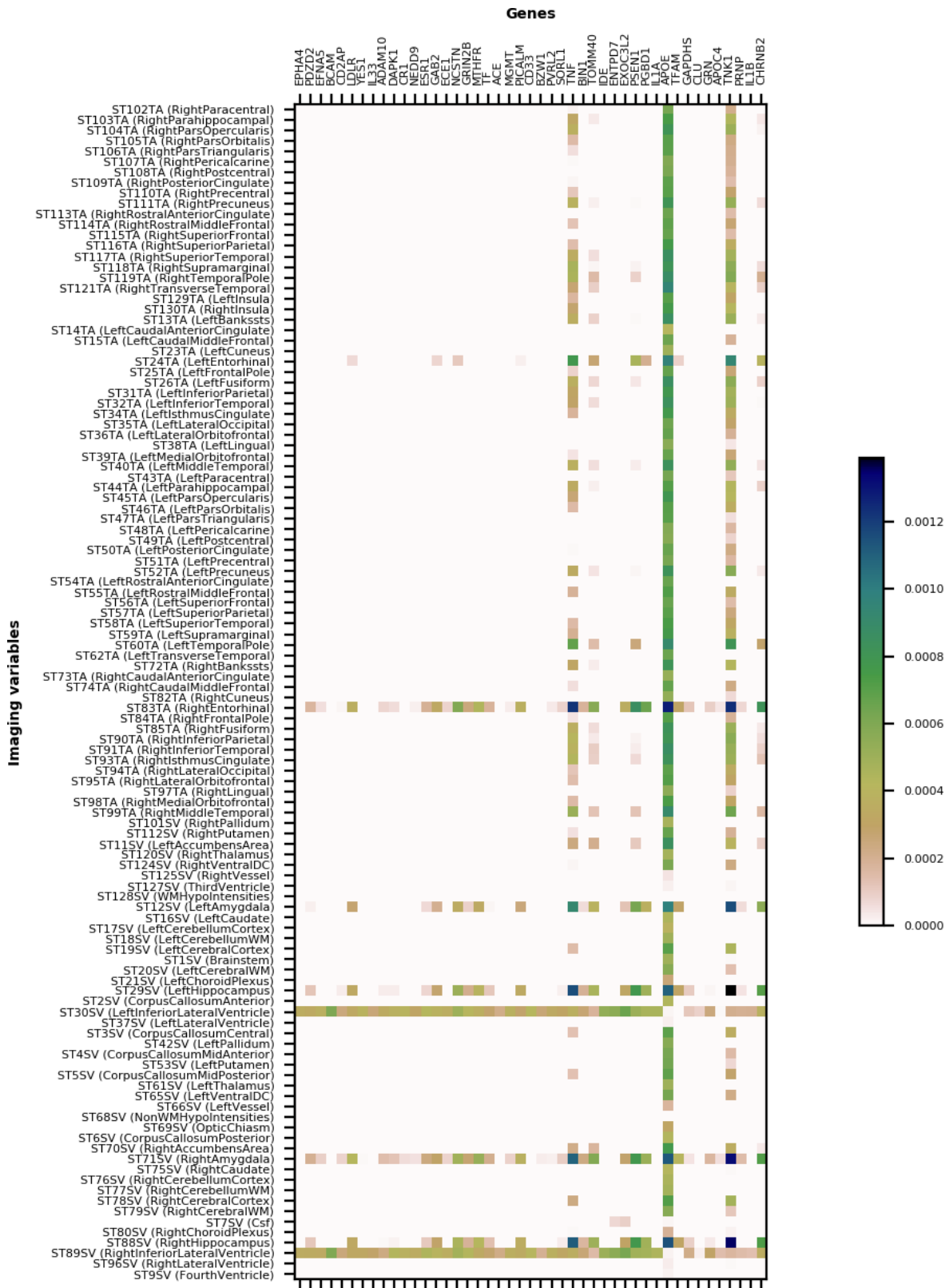


FIGURE 7.2: Reduced matrix \tilde{W}

Conclusion

In this PhD, we have explored the idea that modalities from different sources, instead of being simply added together in an additive model, could be multiplied using a multilevel framework in order to capture interactions between modalities. A multilevel framework will allow to hierarchise modalities and to use different level of decisions. In chapter 4, we proposed a multilevel classification framework for the diagnosis of Alzheimer's Disease. As we worked with high-dimensional data, we used adapted penalties for each modality and their joint interactions. In addition to providing a moderately higher predictive power, this approach is able to capture potential interactions between the genetic covariates, the brain regions, and their relation to the disease. The model uses both genetic and neuroimaging data together in order for assisting diagnosis, but also relates genetic covariates and brain regions as in an association study in imaging genetics.

Then, we focused on the prediction of the conversion date to Alzheimer's Disease. This question has recently become a major interest of the community. In particular, it was the objective of the TADPOLE challenge launched in 2017. We proposed a contribution to this challenge using a standard survival model, namely Aalen's additive model. This contribution ranked high on one of the outcomes of the challenge. Surprisingly, we were one of the only teams to use a survival model, even though this approach is well suited to the problem. We believe that it is due to the fact that such models, while classical in the statistics community, are less well known in the medical image analysis and machine learning communities. The interest of survival approaches were confirmed by experiments done in chapter 6 which showed that doing a classification at repeated fixed times and using some basic survival models provide similar results in terms of area under the ROC curve, but survival models as a whole are slightly better when it comes to consider the balanced classification accuracy. We then proposed several extensions of survival analysis to multimodal data. We first tried to extend the log-logistic survival model, by setting a speed and a position that depend on clinical and genetic data. However, results showed that this log-logistic model tended to be biased and resulted in lower performance compared to basic survival approaches. Then in chapter 7, we proposed a Cox multiplicative survival model which extends the approach of Chapter 4 to survival analysis. The performance was slightly higher than that of an additive Cox model. More interestingly, as in Chapter 4, our method allowed revealing interactions between imaging and genetic data.

Future work

There are multiple perspectives to this work.

First, the proposed multilevel methodology for combining genetic covariates and anatomical brain regions can be applied to other modalities, for instance using the PET neuroimaging modality and/or clinical scores. From a practical point of view, this has the potential to increase the classification accuracy. Indeed, recent papers have reported that PET is superior to MRI for AD classification [Samper-González et al., 2018] and that the combination of clinical and imaging data is superior to imaging alone for predicting progression to AD [Samper-Gonzalez et al., 2019].

In the present manuscript, the results of the TADPOLE challenge concern prediction at one year. Indeed, the evaluation is based on the future visits in the ADNI study which become available progressively. Thus, evaluations will continue every year until 2022. It will be interesting to continue to observe the ranking of the proposed method compared to other submissions.

During our experiments, we noticed a clear relationship, in terms of variations and signs, between the parameter coefficients of a fixed-time logistic regression and of the Aalen additive model. Although these parameter coefficients of both model are not learnt in the same way, as the logistic regression does not use temporal information, whereas the Aalen additive model. It can be interesting to focus on a theoretical relation between both models.

Using other datasets on Alzheimer's Disease (that provide genetic data and clinical data) would have been interesting. First, it would be interesting to check if all the previous models have the same behaviour on different datasets and how well they would generalise across datasets. More specifically, it would be interesting to see if the log-logistic model is still biased.

Finally, in our work, we did not consider death as a competing risk with Alzheimer's Disease. It could be interesting to take it into account. This could be done using the framework of multistate survival models. More generally, we could consider other competing risks such as other forms of dementia or even other diseases. This could pave the way to more general predictive models.

Appendix

Scientific production

Journal papers

- Samper-Gonzalez Jorge, Burgos Ninon, Bottani Simona, Fontanella Sabrina, Lu Pascal, Marcoux Arnaud, Routier Alexandre, Guillon Jérémy, Bacci Michael, Wen Junhao, Bertrand Anne, Bertin Hugo, Habert Marie-Odile, Durrleman Stanley, Evgeniou Theodoros, and Colliot Olivier, for the ADNI & the AIBL, Reproducible evaluation of classification methods in Alzheimers disease: Framework and application to MRI and PET data, *NeuroImage*, 183, 504521, 2018. <https://hal.inria.fr/hal-01858384>

Conference papers

- Lu Pascal, Colliot Olivier. Multilevel Modeling with Structured Penalties for Classification from Imaging Genetics data. *3rd MICCAI Workshop on Imaging Genetics (MICGen 2017)*, Sep 2017, Quebec City, Canada. pp.230-240.
- Lu Pascal, Colliot Olivier. Multilevel Survival Analysis with Structured Penalties for Imaging Genetics data. Accepted at *SPIE 2020 Medical Imaging*.

Posters and talks

- Lu Pascal, Colliot Olivier. Multilevel Survival Analysis with Structured Penalties for Imaging Genetics data. Poster at GDR Statistiques & Santé, Paris, October 2019.
- Lu Pascal, Colliot Olivier. A log-logistic survival model from multimodal data for prediction of Alzheimer's Disease. Poster at SAFJR 2019, Copenhagen, April 2019.
- Lu Pascal, Colliot Olivier. Multilevel Modeling with Structured Penalties for Classification from Imaging Genetics data. Poster at ICM - IoN Workshop, London, UK, October 2017.
- Lu Pascal, Colliot Olivier. Multilevel Modeling with Structured Penalties for Classification from Imaging Genetics data. Oral presentation at MICGen 2017, Quebec, Canada, September 2017.

List of Figures

1.1	Progression of Alzheimer’s Disease (reproduced from https://www.brightfocus.org/alzheimers-disease/infographic/progression-alzheimers-disease)	23
1.2	Simplified biological pathway leading to Alzheimer’s Disease	24
1.3	Coronal slides of the T1-weighted MRI for normal brain, MCI brain and AD brain. The bounding box represents the hippocampus region (reproduced from [Ahmed et al., 2017]).	25
1.4	PIB-PET scan for Normal brain versus AD brain (reproduced from [Ng et al., 2007])	26
1.5	Nucleus, chromosome and DNA	28
1.6	Histogram of the number of visits at each month among MCI patients at baseline	32
1.7	Histogram of the conversion date T^* and the censored date C	32
1.8	Kaplan-Meier estimator for the survival function for MCI patients from ADNI1 dataset	33
1.9	Nelson-Aalen estimator for the cumulative hazard function (top) and estimated hazard function \widehat{h} with bandwidth of 6 months (bottom)	34
2.1	Two topics of research in imaging genetics	36
2.2	Four categories of analyses in imaging genetics, reproduced from [Liu and Calhoun, 2014]	37
2.3	Real vs hypothetic distribution (N is the number of subjects, N_A (resp. N_a) the number of subjects with allele A (resp. a), N_{CN} (resp. N_{AD}) the number of subjects who are CN (resp. have AD))	38
2.4	Genome wide meta-analysis results in AD with 524,993 SNPs. Manhattan plot showing the p -values obtained in the meta-analysis. The end and beginning of a chromosome is denoted by the change of colour pattern of the SNPs (black, grey and brown dots). Genome-wide significance threshold is denoted by a red line. The Y-axis has been truncated (reproduced from [Perez et al., 2014])	38
2.5	(left) Main PLS eigen-component of \mathbf{V} for the phenotype features. (right) Chromosome representativeness among the set of most informative SNPs associated to the main PLS eigen-component of \mathbf{U} [Lorenzi et al., 2016].	40
2.6	Comparison between a univariate association test and a LASSO regression with an imaging-derived measure of temporal lobe volume. (reproduced from [Kohannim et al., 2012b])	41
2.7	Relationship between genetic, imaging and clinical measure (synthesised from [Batmanghelich et al., 2016])	42

2.8	Posterior relevance of the SNPs $p(\mathbf{a}_k \mathbf{x}, \mathbf{y}_k)$ with respect to (a) average thickness of the left entorhinal cortex, and (b) volume of the left hippocampus . . .	43
2.9	Pathways, genes and SNPs	44
2.10	A tree in a random forest [Silver and Montana, 2012]	45
2.11	Illustration of β learnt using the LASSO regression (top), and the Group LASSO regression where $L = 3$ (bottom). The LASSO regression selects only some variables, where the Group LASSO regression selects only some groups.	48
2.12	The problem of overlapping genes/pathways	48
2.13	Parameter \mathbf{w} learnt using the $\ell_{1,p}$ -MKL	50
3.1	Different type of censorships	54
3.2	Survival and hazard functions for exponential model	58
3.3	Survival and hazard functions for log-logistic model	59
3.4	Order graphs representing the ranking constraints. (a) No censored data and (b) with censored data. The empty circle represents a censored point. The points are arranged in the increasing value of their survival times with the lowest being at the bottom (reproduced from [Steck et al., 2008]).	62
3.5	Illness-death model (reproduced from [Joly et al., 2002], [Leffondré et al., 2013], [Yu et al., 2010])	66
4.1	The disease status y is predicted from imaging data \mathbf{x}_I and the parameters $\beta_0(\mathbf{x}_G), \beta(\mathbf{x}_G)$ (which are computed from genetic data \mathbf{x}_G)	73
4.2	Overview of the reduced parameters $\bar{\mathbf{W}} \in \mathbb{R}^{I \times L}$, $\bar{\beta}_I \in \mathbb{R}^I$ and $\bar{\beta}_G \in \mathbb{R}^L$ (learnt through the task "pMCI vs CN" for the whole model). For brain region i and gene ℓ , $\bar{\mathbf{W}}[i, \ell] = \max_{g \in \mathcal{G}_i} \mathbf{W}[i, g] $, $\bar{\beta}_I[i] = \beta_I[i] $ and $\bar{\beta}_G[\ell] = \max_{g \in \mathcal{G}_i} \beta_G[g] $. Only some brain regions are shown in this figure.	79
4.3	Intercept β_I and slope \mathbf{W} in the function $\alpha(\mathbf{x}_G) = \mathbf{W}\mathbf{x}_G + \beta_I$	80
4.4	Rows of \mathbf{W}	81
4.5	Rows of \mathbf{W}	82
5.1	Submission file format for TADPOLE (where Ventri. stands for ventricle volume and CI for confidence interval)	84
5.2	TADPOLE Challenge design. Participants are required to train a predictive model on a training dataset (D1 and/or others) and make forecasts for different datasets (D2, D3) by the submission deadline. Evaluation will be performed on a test dataset (D4) that is acquired after the submission deadline (reproduced from [Marinescu et al., 2018]).	85
5.3	Venn diagram of the ADNI datasets for training (D1), longitudinal prediction (D2), cross-sectional prediction (D3) and the test set (D4). D3 is a subset of D2, which in turn is a subset of D1. Other non-ADNI data can also be used for training (reproduced from [Marinescu et al., 2018]).	85
5.4	Possible paths	87
5.5	Proportion of each possible path based on D1 dataset	88
5.6	Estimators for the survival, cumulative hazard, and hazard functions	90

5.7	Cumulative regression functions of $\widehat{A}(t)$	91
5.8	Proportion of each possible paths for D4 dataset	92
6.1	Coefficients learnt for the Cox PH model, with confidence intervals	99
6.2	Coefficients of the log-logistic model	100
6.3	Comparison between the survival function estimated using the Kaplan-Meier estimator and the log-logistic regression; and between the cumulative hazard estimated using the Nelson-Aalen estimator and the log-logistic regression	101
6.4	Coefficients of the parameter vector β of the logistic regression at fixed time t (left); Cumulative regression functions in the Aalen additive model (right)	102
6.5	Assesment of the predictive value for the different models: Balanced Accuracy (top), AUC (middle), C-index (bottom)	103
6.6	Estimator of Kaplan-Meier of the survival function $S(t) = \mathbb{P}\{T > t \text{APOE}\}$. .	104
6.7	Log-logistic model	105
6.8	Effect of $\beta_I \mathbf{x}_I$ and $\beta_G \mathbf{x}_G$ on $S(t)$. It can be seen that $\beta_I \mathbf{x}_I$ only translates the curve of $S(t)$ (for fixed $\beta_G \mathbf{x}_G$), whereas $\beta_G \mathbf{x}_G$ changes both the shape and translation of the curve of $S(t)$	106
6.9	Example of domain \mathcal{D} in two-dimensions	107
6.10	Estimated survival function \widehat{S} for the Kaplan Meier estimate and for the multilevel log-logistic model; Estimated hazard function \widehat{h} for the Nelson-Aalen estimated hazard and for the multilevel log-logistic model	109
6.11	Median survival time, estimated survival function \widehat{S} for the Kaplan Meier estimate and for the multilevel log-logistic model	110
6.12	Results for truncation time $\tau \in \{50, 80\}$	111
7.1	ADNI1 Dataset: baseline survival function and hazard function	118
7.2	Reduced matrix $\widetilde{\mathbf{W}}$	121

List of Tables

1.1	Sequence of SNPs	29
1.2	Example of table defining major/minor variant in the entire population	29
1.3	Descriptive statistics for variables measured at study entry of ADNI1 participants who are AD or CN at baseline, [†] means that this feature is computed on a subset of ADNI1	31
1.4	Descriptive statistics for variables measured at study entry of ADNI1 participants with mild cognitive impairment (MCI), [†] means that this feature is computed on a subset of ADNI1	31
4.1	Classification results for different modalities and methods	78
5.1	TADPOLE timeline	84
5.2	Descriptive statistics of D1, D2, D3 datasets (reproduced from https://tadpole.grand-challenge.org/Results/)	86
5.3	Descriptive statistics of D4 dataset (219 subjects, reproduced from https://tadpole.grand-challenge.org/Results/)	92
5.4	Ranking based on mAUC	94
5.5	Ranking based on BCA	95
6.1	Number of positif labels and total number of labels throughout time	99
6.2	Testing the proportional hazard hypothesis	100
6.3	Coefficients for the Cox model and log-logistic model (↗ means that the higher this covariates is, the risk to have AD increases; whereas ↘ means that the higher this covariates is, the risk to have AD decreases)	108
6.4	Average results on ADNI1	109
6.5	AUC for each date	109
7.1	Results for different modalities and methods (mean value across the test folds ± standard deviation)	119

Bibliography

- [Aalen, 1980] Aalen, O. (1980). A Model for Nonparametric Regression Analysis of Counting Processes. In Klonecki, W., Kozek, A., and Rosiski, J., editors, *Mathematical Statistics and Probability Theory*, Lecture Notes in Statistics, pages 1–25, New York, NY. Springer.
- [Ahmed et al., 2017] Ahmed, O. B., Benois-Pineau, J., Allard, M., Catheline, G., and Amar, C. B. (2017). Recognition of Alzheimer’s disease and Mild Cognitive Impairment with multimodal image-derived biomarkers and Multiple Kernel Learning. *Neurocomputing*, 220:98–110.
- [Aiolli and Donini, 2015] Aiolli, F. and Donini, M. (2015). EasyMKL: a scalable multiple kernel learning algorithm. *Neurocomputing*, 169:215–224.
- [Anderson et al., 2016] Anderson, E. D., Wahoske, M., Huber, M., Norton, D., Li, Z., Kosciak, R. L., Umucu, E., Johnson, S. C., Jones, J., Asthana, S., and Gleason, C. E. (2016). Cognitive variability A marker for incident MCI and AD: An analysis for the Alzheimer’s Disease Neuroimaging Initiative. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, 4:47–55.
- [Batmanghelich et al., 2016] Batmanghelich, N. K., Dalca, A., Quon, G., Sabuncu, M., and Golland, P. (2016). Probabilistic Modeling of Imaging, Genetics and Diagnosis. *IEEE Transactions on Medical Imaging*, 35(7):1765–1779.
- [Beck and Teboulle, 2009] Beck, A. and Teboulle, M. (2009). Gradient-based algorithms with applications to signal-recovery problems.
- [Binder and Schumacher, 2016] Binder, N. and Schumacher, M. (2016). Incidence of Dementia over Three Decades in the Framingham Heart Study. *New England Journal of Medicine*, 375(1):92–94.
- [Bøvelstad et al., 2009] Bøvelstad, H. M., Nygård, S., and Borgan, . (2009). Survival prediction from clinico-genomic models - a comparative study. *BMC Bioinformatics*, 10(1):413.
- [Chambless and Diao, 2006] Chambless, L. E. and Diao, G. (2006). Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine*, 25(20):3474–3486.
- [Chekouo et al., 2016] Chekouo, T., Stingo, F. C., Guindani, M., and Do, K.-A. (2016). A Bayesian predictive model for imaging genetics with application to schizophrenia. *The Annals of Applied Statistics*, 10(3):1547–1571.

- [Cheng et al., 2015] Cheng, B., Liu, M., Zhang, D., Munsell, B. C., and Shen, D. (2015). Domain Transfer Learning for MCI Conversion Prediction. *IEEE transactions on bio-medical engineering*, 62(7):1805–1817.
- [Davatzikos et al., 2011] Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N., and Trojanowski, J. Q. (2011). Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging*, 32(12):2322.e19–27.
- [Du et al., 2014] Du, L., Jingwen, Y., Kim, S., Risacher, S. L., Huang, H., Inlow, M., Moore, J. H., Saykin, A. J., Shen, L., and Alzheimer’s Disease Neuroimaging Initiative (2014). A novel structure-aware sparse learning algorithm for brain imaging genetics. *Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 17(Pt 3):329–336.
- [Dubois and Albert, 2004] Dubois, B. and Albert, M. L. (2004). Amnestic MCI or prodromal Alzheimer’s disease? *The Lancet Neurology*, 3(4):246–248.
- [Dubois et al., 2007] Dubois, B., Feldman, H. H., Jacova, C., Dekosky, S. T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., Meguro, K., O’Brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Stern, Y., Visser, P. J., and Scheltens, P. (2007). Research criteria for the diagnosis of Alzheimer’s disease: revising the NINCDS-ADRDA criteria. *The Lancet. Neurology*, 6(8):734–746.
- [Dubois et al., 2014] Dubois, B., Feldman, H. H., Jacova, C., Hampel, H., Molinuevo, J. L., Blennow, K., DeKosky, S. T., Gauthier, S., Selkoe, D., Bateman, R., Cappa, S., Crutch, S., Engelborghs, S., Frisoni, G. B., Fox, N. C., Galasko, D., Habert, M.-O., Jicha, G. A., Nordberg, A., Pasquier, F., Rabinovici, G., Robert, P., Rowe, C., Salloway, S., Sarazin, M., Epelbaum, S., de Souza, L. C., Vellas, B., Visser, P. J., Schneider, L., Stern, Y., Scheltens, P., and Cummings, J. L. (2014). Advancing research diagnostic criteria for Alzheimer’s disease: the IWG-2 criteria. *The Lancet. Neurology*, 13(6):614–629.
- [Dubois et al., 2016] Dubois, B., Hampel, H., Feldman, H. H., Scheltens, P., Aisen, P., Andrieu, S., Bakardjian, H., Benali, H., Bertram, L., Blennow, K., Broich, K., Cavedo, E., Crutch, S., Dartigues, J.-F., Duyckaerts, C., Epelbaum, S., Frisoni, G. B., Gauthier, S., Genthon, R., Gouw, A. A., Habert, M.-O., Holtzman, D. M., Kivipelto, M., Lista, S., Molinuevo, J.-L., O’Bryant, S. E., Rabinovici, G. D., Rowe, C., Salloway, S., Schneider, L. S., Sperling, R., Teichmann, M., Carrillo, M. C., Cummings, J., and Jack, C. R. (2016). Preclinical Alzheimer’s disease: Definition, natural history, and diagnostic criteria. *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association*, 12(3):292–323.
- [Duyckaerts et al., 2009] Duyckaerts, C., Delatour, B., and Potier, M.-C. (2009). Classification and basic pathology of Alzheimer disease. *Acta Neuropathologica*, 118(1):5–36.
- [Graf et al., 1999] Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

- [Gönen and Alpaydn, 2011] Gönen, M. and Alpaydn, E. (2011). Multiple Kernel Learning Algorithms. *Journal of Machine Learning Research*, 12(Jul):2211–2268.
- [Hampel et al., 2014] Hampel, H., Lista, S., Teipel, S. J., Garaci, F., Nisticò, R., Blennow, K., Zetterberg, H., Bertram, L., Duyckaerts, C., Bakardjian, H., Drzezga, A., Colliot, O., Epelbaum, S., Broich, K., Lehericy, S., Brice, A., Khachaturian, Z. S., Aisen, P. S., and Dubois, B. (2014). Perspective on future role of biological markers in clinical therapy trials of Alzheimer’s disease: a long-range point of view beyond 2020. *Biochemical Pharmacology*, 88(4):426–449.
- [Hastie et al., 2015] Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity - The Lasso and Generalizations*, volume 143 of *Monographs on Statistics and Applied Probability* 143. Crc press edition.
- [Jacob et al., 2009] Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group Lasso with Overlap and Graph Lasso. *Proceedings of the 26 th International Conference on Machine Learning*.
- [Janssen et al., 2000] Janssen, J. C., Hall, M., Fox, N. C., Harvey, R. J., Beck, J., Dickinson, A., Campbell, T., Collinge, J., Lantos, P. L., Cipelotti, L., Stevens, J. M., and Rossor, M. N. (2000). Alzheimer’s disease due to an intronic presenilin-1 (PSEN1 intron 4) mutationA clinicopathological study. *Brain*, 123(5):894–907.
- [Joly et al., 2002] Joly, P., Commenges, D., Helmer, C., and Letenneur, L. (2002). A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics (Oxford, England)*, 3(3):433–443.
- [Kloft et al., 2011] Kloft, M., Brefeld, U., Sonnenburg, S., and Zien, A. (2011). Ip-Norm Multiple Kernel Learning. *Journal of Machine Learning Research*, 12(Mar):953–997.
- [Kohannim et al., 2012a] Kohannim, O., Hibar, D. P., Stein, J. L., Jahanshad, N., Hua, X., Rajagopalan, P., Toga, A. W., Jack, C. R., Weiner, M. W., de Zubicaray, G. I., McMahon, K. L., Hansell, N. K., Martin, N. G., Wright, M. J., Thompson, P. M., and Alzheimers Disease Neuroimaging Initiative (2012a). Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression. *Frontiers in Neuroscience*, 6:115.
- [Kohannim et al., 2012b] Kohannim, O., Hibar, D. P., Stein, J. L., Jahanshad, N., Hua, X., Rajagopalan, P., Toga, A. W., Jack, C. R., Weiner, M. W., de Zubicaray, G. I., McMahon, K. L., Hansell, N. K., Martin, N. G., Wright, M. J., Thompson, P. M., and Alzheimers Disease Neuroimaging Initiative (2012b). Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression. *Frontiers in Neuroscience*, 6:115.
- [Laird and Lange, 2011] Laird, N. M. and Lange, C. (2011). Introduction to Statistical Genetics and Background in Molecular Genetics. In *The Fundamentals of Modern Statistical Genetics*, Statistics for Biology and Health, pages 1–13. Springer New York.
- [Lambert et al., 2013] Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., DeStafano, A. L., Bis, J. C., Beecham, G. W., Grenier-Boley, B., Russo, G., Thorton-Wells, T. A., Jones, N., Smith, A. V., Chouraki, V., Thomas, C., Ikram,

- M. A., Zelenika, D., Vardarajan, B. N., Kamatani, Y., Lin, C. F., Gerrish, A., Schmidt, H., Kunkle, B., Dunstan, M. L., Ruiz, A., Bihoreau, M. T., Choi, S. H., Reitz, C., Pasquier, F., Cruchaga, C., Craig, D., Amin, N., Berr, C., Lopez, O. L., De Jager, P. L., Deramecourt, V., Johnston, J. A., Evans, D., Lovestone, S., Letenneur, L., Morón, F. J., Rubinsztein, D. C., Eiriksdottir, G., Sleegers, K., Goate, A. M., Fiévet, N., Huentelman, M. W., Gill, M., Brown, K., Kamboh, M. I., Keller, L., Barberger-Gateau, P., McGuinness, B., Larson, E. B., Green, R., Myers, A. J., Dufouil, C., Todd, S., Wallon, D., Love, S., Rogaeva, E., Gallacher, J., St George-Hyslop, P., Clarimon, J., Lleo, A., Bayer, A., Tsuang, D. W., Yu, L., Tsolaki, M., Bossù, P., Spalletta, G., Proitsi, P., Collinge, J., Sorbi, S., Sanchez-Garcia, F., Fox, N. C., Hardy, J., Deniz Naranjo, M. C., Bosco, P., Clarke, R., Brayne, C., Galimberti, D., Mancuso, M., Matthews, F., European Alzheimer's Disease Initiative (EADI), Genetic and Environmental Risk in Alzheimer's Disease, Alzheimer's Disease Genetic Consortium, Cohorts for Heart and Aging Research in Genomic Epidemiology, Moebus, S., Mecocci, P., Del Zompo, M., Maier, W., Hampel, H., Pilotto, A., Bullido, M., Panza, F., Caffarra, P., Nacmias, B., Gilbert, J. R., Mayhaus, M., Lannefelt, L., Hakonarson, H., Pichler, S., Carrasquillo, M. M., Ingelsson, M., Beekly, D., Alvarez, V., Zou, F., Valladares, O., Younkin, S. G., Coto, E., Hamilton-Nelson, K. L., Gu, W., Razquin, C., Pastor, P., Mateo, I., Owen, M. J., Faber, K. M., Jonsson, P. V., Combarros, O., O'Donovan, M. C., Cantwell, L. B., Soininen, H., Blacker, D., Mead, S., Mosley, T. H., Bennett, D. A., Harris, T. B., Fratiglioni, L., Holmes, C., de Bruijn, R. F., Passmore, P., Montine, T. J., Bettens, K., Rotter, J. I., Brice, A., Morgan, K., Foroud, T. M., Kukull, W. A., Hannequin, D., Powell, J. F., Nalls, M. A., Ritchie, K., Lunetta, K. L., Kauwe, J. S., Boerwinkle, E., Riemenschneider, M., Boada, M., Hiltunen, M., Martin, E. R., Schmidt, R., Rujescu, D., Wang, L. S., Dartigues, J. F., Mayeux, R., Tzourio, C., Hofman, A., Nöthen, M. M., Graff, C., Psaty, B. M., Jones, L., Haines, J. L., Holmans, P. A., Lathrop, M., Pericak-Vance, M. A., Launer, L. J., Farrer, L. A., van Duijn, C. M., Van Broeckhoven, C., Moskvina, V., Seshadri, S., Williams, J., Schellenberg, G. D., and Amouyel, P. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45(12):1452–1458.
- [Leffondré et al., 2013] Leffondré, K., Touraine, C., Helmer, C., and Joly, P. (2013). Interval-censored time-to-event and competing risk with death: is the illness-death model more accurate than the Cox model? *International Journal of Epidemiology*, 42(4):1177–1186.
- [Li et al., 2017] Li, K., O'Brien, R., Lutz, M., Luo, S., and ADNI, T. (2017). A prognostic model of Alzheimer's disease relying on multiple longitudinal measures and time-to-event data. *Alzheimers & Dementia*.
- [Liu and Calhoun, 2014] Liu, J. and Calhoun, V. D. (2014). A review of multivariate analyses in imaging genetics. *Frontiers in Neuroinformatics*, 8.
- [Liu et al., 2017] Liu, K., Chen, K., Yao, L., and Guo, X. (2017). Prediction of Mild Cognitive Impairment Conversion Using a Combination of Independent Component Analysis and the Cox Model. *Frontiers in Human Neuroscience*, 11:33.

- [Lorenzi et al., 2016] Lorenzi, M., Gutman, B., Hibar, D. P., Altmann, A., Jahanshad, N., Thompson, P. M., and Ourselin, S. (2016). Partial least squares modelling for imaging-genetics in Alzheimer’s disease: Plausibility and generalization. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 838–841.
- [Mahley et al., 2006] Mahley, R. W., Weisgraber, K. H., and Huang, Y. (2006). Apolipoprotein E4: A causative factor and therapeutic target in neuropathology, including Alzheimers disease. *Proceedings of the National Academy of Sciences*, 103(15):5641–5643.
- [Marinescu et al., 2018] Marinescu, R. V., Oxtoby, N. P., Young, A. L., Bron, E. E., Toga, A. W., Weiner, M. W., Barkhof, F., Fox, N. C., Klein, S., Alexander, D. C., and Consortium, t. E. (2018). TADPOLE Challenge: Prediction of Longitudinal Evolution in Alzheimer’s Disease. *arXiv:1805.03909 [q-bio, stat]*. arXiv: 1805.03909.
- [McKhann et al., 1984] McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer’s disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer’s Disease. *Neurology*, 34(7):939–944.
- [McKhann et al., 2011] McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., Mohs, R. C., Morris, J. C., Rossor, M. N., Scheltens, P., Carrillo, M. C., Thies, B., Weintraub, S., and Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer’s disease: recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease. *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association*, 7(3):263–269.
- [Meier et al., 2008] Meier, L., van de Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. pages 53–71.
- [Ming and Yi, 2006] Ming, Y. and Yi, L. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B (2006) 68, Part 1, pp. 4967*, pages 49–67.
- [Ng et al., 2007] Ng, S., Villemagne, V. L., Berlangieri, S., Lee, S.-T., Cherk, M., Gong, S. J., Ackermann, U., Saunderson, T., Tochon-Danguy, H., Jones, G., Smith, C., O’Keefe, G., Masters, C. L., and Rowe, C. C. (2007). Visual Assessment Versus Quantitative Assessment of 11c-PIB PET and 18f-FDG PET for Detection of Alzheimer’s Disease.
- [Peng et al., 2016] Peng, J., An, L., Zhu, X., Jin, Y., and Shen, D. (2016). Structured Sparse Kernel Learning for Imaging Genetics Based Alzheimers Disease Diagnosis. In Ourselin, S., Joskowicz, L., Sabuncu, M. R., Unal, G., and Wells, W., editors, *MICCAI 2016*, number 9901 in Lecture Notes in Computer Science, pages 70–78. Springer International Publishing.
- [Perez et al., 2014] Perez, E., Bustos, B., Villaman, C., Alarcón, M., Avila, M., Ugarte, G., Reyes, A., Opazo, C., and De Ferrari, G. (2014). Overrepresentation of Glutamate Signaling in Alzheimer’s Disease: Network-Based Pathway Enrichment Using Meta-Analysis of Genome-Wide Association Studies. *PLoS one*, 9:e95413.

- [Rathore et al., 2017] Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., and Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, 155:530–548.
- [Samper-Gonzalez et al., 2019] Samper-Gonzalez, J., Burgos, N., Bottani, S., Habert, M.-O., Evgeniou, T., Epelbaum, S., and Colliot, O. (2019). Reproducible evaluation of methods for predicting progression to Alzheimer's disease from clinical and neuroimaging data. In *Medical Imaging 2019: Image Processing*, volume 10949, page 109490V. International Society for Optics and Photonics.
- [Samper-González et al., 2018] Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., Bertrand, A., Bertin, H., Habert, M.-O., Durrleman, S., Evgeniou, T., Colliot, O., Alzheimer's Disease Neuroimaging Initiative, and Australian Imaging Biomarkers and Lifestyle flagship study of ageing (2018). Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. *NeuroImage*, 183:504–521.
- [Satizabal et al., 2016] Satizabal, C. L., Beiser, A. S., Chouraki, V., Chêne, G., Dufouil, C., and Seshadri, S. (2016). Incidence of Dementia over Three Decades in the Framingham Heart Study. *New England Journal of Medicine*, 374(6):523–532.
- [Scheltens et al., 1992] Scheltens, P., Leys, D., Barkhof, F., Huglo, D., Weinstein, H. C., Vermersch, P., Kuiper, M., Steinling, M., Wolters, E. C., and Valk, J. (1992). Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. *Journal of Neurology, Neurosurgery, and Psychiatry*, 55(10):967–972.
- [Silver et al., 2012] Silver, M., Janousova, E., Hua, X., Thompson, P. M., Montana, G., and Alzheimer's Disease Neuroimaging Initiative (2012). Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. *NeuroImage*, 63(3):1681–1694.
- [Silver and Montana, 2012] Silver, M. and Montana, G. (2012). Fast Identification of Biological Pathways Associated with a Quantitative Trait Using Group Lasso with Overlaps. *Statistical applications in genetics and molecular biology*, 11(1):Article–7.
- [Sim et al., 2013] Sim, A., Tsagkrasoulis, D., and Montana, G. (2013). Random Forests on Distance Matrices for Imaging Genetics Studies. *arXiv:1309.6158 [stat]*. arXiv: 1309.6158.
- [Steck et al., 2008] Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P., and Raykar, V. C. (2008). On Ranking in Survival Analysis: Bounds on the Concordance Index. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1209–1216. Curran Associates, Inc.
- [van der Laan et al., 2007] van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super Learner.

- [van Houwelingen et al., 2006] van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., vant Veer, L. J., and Wessels, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. pages 3201–3216.
- [van Houwelingen and Putter, 2012] van Houwelingen, H. C. and Putter, H. (2012). *Dynamic Prediction in Clinical Survival Analysis*, volume Monographs on Statistics and Applied Probability 123. CRC Press.
- [Wang et al., 2012] Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., Saykin, A. J., and Shen, L. (2012). Identifying quantitative trait loci via group-sparse multi-task regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics*, 28(2):229–237.
- [Yu et al., 2010] Yu, B., Saczynski, J. S., and Launer, L. J. (2010). Multiple Imputation for Estimating the Risk of Developing Dementia and Its Impact on Survival. *Biometrical journal. Biometrische Zeitschrift*, 52(5):616–627.
- [Yu et al., 2011] Yu, C.-N., Greiner, R., Lin, H.-C., and Baracos, V. (2011). Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 1845–1853. Curran Associates, Inc.