

The symphony of cacophony

Understanding the order in neurodegenerative diseases



Vikram
Venkatraghavan

The symphony of cacophony
Understanding the order in neurodegenerative diseases

Vikram Venkatraghavan

Cover design by Douwe Hoendervanger and Vikram Venkatraghavan.

The work in this thesis was conducted at the department of Radiology & Nuclear Medicine of the Erasmus MC, Rotterdam, the Netherlands. This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 666992 and from TKI-LSH Health Holland Alzheimer Nederland project (No. LSHM18049).



This work was carried out in the ASCI graduate school. ASCI dissertation series number 415.

For financial support for the publication of this thesis, the following organizations are gratefully acknowledged: the ASCI graduate school, Erasmus MC.

ISBN: 978-94-6423-248-6

Printed by Proefschriftmaken

© 2021 Vikram Venkatraghavan

All rights reserved. No part of this thesis may be reproduced or transmitted in any form or by any means without prior permission of the copyright owner.

THE SYMPHONY OF CACOPHONY
Understanding the order in neurodegenerative diseases

De symfonie van kakofonie
Het begrijpen van de orde in neurodegeneratieve ziekten

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof. dr. F.A. van der Duijn Schouten

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
dinsdag 08 juni 2021 om 10:30 uur

door

Vikram Venkatraghavan
geboren te Srirangam, India

Promotiecommissie

Promotor: Prof. dr. W.J. Niessen

Overige leden: Prof. dr. M. Smits

Prof. dr. W.M. van der Flier

Dr. M. Lorenzi

Copromotoren: Dr. ir. S. Klein

Dr. E.E. Bron

Contents

| | Page |
|--|--|
| Chapter 1 | General introduction |
| | 7 |
| Part I - Discriminative Event Based Modeling (DEBM) | |
| Chapter 2 | Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling |
| | 15 |
| Chapter 3 | Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease |
| | 55 |
| Chapter 4 | Modelling the cascade of biomarker changes in progranulin related frontotemporal dementia |
| | 83 |
| Chapter 5 | The sequence of structural, functional and cognitive changes in multiple sclerosis |
| | 109 |
| Chapter 6 | The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge: Results after 1 Year Follow-up |
| | 135 |
| Part II - Extensions of DEBM: high-dimensional biomarkers, stratified populations, and subtypes | |
| Chapter 7 | Event-based modeling with high-dimensional imaging biomarkers for estimating spatial progression of dementia |
| | 173 |
| Chapter 8 | Analyzing the effect of <i>APOE</i> in Alzheimer's disease progression timeline using event-based model with coupled training in stratified data |
| | 187 |
| Chapter 9 | Progression along <i>APOE</i> -specific data-driven temporal cascades is predictive of Alzheimer's disease in a population-based cohort |
| | 205 |
| Chapter 10 | Subtyping in sporadic Creutzfeldt-Jakob disease with discriminative event-based modeling |
| | 221 |
| General discussion & summary | |
| Chapter 11 | General discussion |
| | 241 |
| | Summary |
| | 249 |
| | Nederlandse samenvatting |
| | 253 |
| | Bibliography |
| | 257 |
| | Acknowledgements |
| | 287 |
| | Publications |
| | 291 |
| | PhD portfolio |
| | 295 |
| | About the author |
| | 297 |

Chapter 1

General introduction

Neurodegenerative diseases such as Alzheimer's disease are notoriously heterogeneous; pathologically as well as in their clinical presentation in patients. There are differences between different patients in terms of the pathways of progression, the speed of progression, and the effect the progression has on a patient's cognition. These myriad of differences not only makes clinical diagnosis of these diseases very challenging, but also has major implications for the efficacy of drug trials. As heterogeneous as these diseases are, there is an underlying order in their progression. An underlying method to their disruption of homeostasis. An underlying symphony leading to the cacophony.

This thesis is about developing data-driven methods for understanding the orderly progression of neurodegenerative diseases and validating their utility in providing insights into the progression patterns of four such diseases: Alzheimer's disease (AD), frontotemporal dementia (FTD), Creutzfeldt-Jakob disease (CJD) and multiple sclerosis (MS).

1.1 Neurodegenerative diseases

Neurodegenerative diseases are characterized by a cascade of changes in the structure and function of the central nervous system. This section provides an overview of the above mentioned four diseases and their underlying processes resulting in neurodegeneration.

AD is a fatal progressive brain disease that gradually deteriorates memory, thinking, and other cognitive skills. AD is associated with accumulation of amyloid- β plaques and hyperphosphorylated tau tangles in the brain [1, 2, 3]. These neuropathological alterations starts to occur up to 10 – 20 years before symptom onset [4]. These abnormalities eventually cascade to disrupt the structural integrity of the brain [5]. The structural disruptions in the brain eventually impair the cognition of the patient, eventually leading to their death.

FTD is the second most common form of presenile dementia (symptom onset age < 65) accounting for ~10% of all presenile dementia cases [6]. This is also a fatal

progressive brain disease that gradually causes deterioration in personality and behavior. Pathologically, FTD is associated with abnormal forms of Tau, TAR DNA-binding protein 43 (*TDP-43*) or FET proteins [7]. These abnormalities eventually cascade to disrupt the white matter microstructural integrity of the brain [8] as well as the integrity of the gray matter [9], causing cognitive problems. The earliest structural changes in the brain were identified ~ 10 years before symptoms onset, while the earliest signs of cognitive decline were identified ~ 5 years before symptoms onset [10]. The pattern of structural degeneration in the brain as well as in cognition is distinct from that in AD [11, 12].

CJD is a rare neurodegenerative disease caused by abnormal prion proteins in the brain. CJD has a very prolonged incubation period [13, 14], but after the initial onset of symptoms the lesions in the brain and the clinical symptoms, such as memory impairment and poor coordination, cascade rapidly. The survival time after the symptom onset varies from a few weeks to a few months [15], although a few slow progressing variants could take decades [13]. The lesions in the brain are characterized by spongiform abnormalities.

MS is a non-fatal chronic inflammatory disease of the central nervous system [16]. MS is associated with a cascading accumulation of demyelinating lesions that occur in the brain's gray matter and white matter regions, as well as in the spinal cord [17, 18]. These abnormalities frequently result in cognitive [19] and physical disabilities in patients [20].

1.2 Biomarkers of neurodegeneration

Biomarkers are measures of different biological states or processes used for objectively quantifying different aspects of a disease or susceptibility to it. They are standardized so as to be comparable among different subjects. The biomarkers of neurodegeneration can be broadly categorized as follows: i) fluid-based biomarkers extracted from blood or cerebrospinal fluids (CSF), ii) imaging biomarkers such as those extracted from a magnetic resonance image (MRI) or positron emission tomography (PET), iii) cognitive biomarkers obtained from neuropsychological examinations, and iv) genetic biomarkers. The commonly used biomarkers for the aforementioned four neurodegenerative diseases are as follows:

AD: Biomarkers for Amyloid- β and hyperphosphorylated tau protein abnormalities can be measured in blood [21, 22], CSF [23], or can be observed using brain PET imaging [24, 25]. MRI based biomarkers such as the volumes of different brain regions (volumetric biomarkers) quantify the structural integrity of the brain. Cognitive biomarkers such as measures quantifying attention, episodic memory etc., can be quantified by neuropsychological tests [26]. Genetic biomarkers such as mutations in the amyloid

precursor protein (*APP*), presenilin 1 (*PSEN1*), and presenilin 2 (*PSEN2*) genes are used as genetic biomarkers to identify familial AD [3]. Alleles of the Apolipoprotein E (*APoE*) gene are used to identify genetic risk factors of sporadic AD [27].

FTD: Specific biomarkers to identify *TDP-43* or FET pathology *in-vivo* are still lacking [28, 29]. In the absence of such biomarkers, neurofilament lightchain (NfL) obtained from blood or CSF is used as a non-specific biomarker of neuroaxonal degeneration [30]. Diffusion tensor imaging (DTI) based biomarkers are used to quantify microstructural white matter integrity in the brain [8] and volumetric biomarkers obtained from MRI are used to quantify structural integrity. Cognitive biomarkers such as measures quantifying behavior, language etc., can be quantified by neuropsychological tests [31]. Genetic biomarkers such as mutations in the progranulin (*GRN*), microtubule-associated protein tau (*MAPT*), and chromosome 9 open reading frame 72 (*C9orf72*) genes are used as genetic biomarkers to identify familial FTD [7].

CJD: The characteristic spongiform abnormalities in the brain can be observed using diffusion weighted MRI [15]. CSF biomarkers of Tau and 14-3-3 proteins are also used for diagnostic purposes [32]. Homozygosity at the *PRNP* gene is used to identify a genetic risk factor for sporadic CJD [33].

MS: MRI scans of brain and spine are often used for observing the demyelinating abnormalities that are characteristic of MS [34]. The volumetric biomarkers derived from MRIs quantify the structural integrity of the brain [35]. CSF biomarkers of oligoclonal bands (OCB) are also used to confirm the diagnosis of MS [36]. Neuropsychological tests and expanded disability status scale [37] are used to assess the cognitive and physical disabilities in patients.

1.3 Need for understanding neurodegenerative diseases

In spite of the similarity in the pathological processes that helps us categorize each of these neurodegenerative diseases, there are multiple pathological pathways through which these diseases progress, resulting in distinct subtypes [38, 39, 40, 41]. Clinical trials in AD, FTD, CJD have failed so far to be effective in altering the natural course of these diseases. Although there have been a few drugs for MS that alter its natural course [42], they are only effective for the earliest stages of the disease [43].

Understanding the progression in the different subtypes of these neurodegenerative diseases is pivotal for a multitude of reasons. Understanding the earliest biomarkers in the neuropathologic cascade in different subtypes of these diseases can help identify individuals at-risk of developing symptoms, and monitor their progression effectively in the pre-clinical stages of the diseases. This would help in selecting patients in the earliest stages of the disease for clinical trials, which could be crucial for their success [39, 43, 44, 45].

Grasping the various pathologic pathways can help in identifying novel therapeutic targets for clinical trials [46, 47, 48, 49] and formulating evidence-based personalized treatment strategies. Until effective treatments are discovered, understanding the earliest symptoms in different clinical subtypes of the diseases can also help in formulating supportive therapies for efficient patient care.

One of the earliest attempts to understand the progression of neurodegenerative diseases was done for AD by Braak and Braak [50], where they used neuropathological data from deceased AD patients to understand the spatial spread of Amyloid- β plaques and Tau tangles. Almost two decades later, Jack Jr. *et al.* [5] proposed a hypothetical model of AD's pathological cascade of key biomarkers, using meta-analysis from literature. While such approaches give a bird's eye view of the diseases, there is a need for computational approaches to understand the heterogeneous progression of such diseases from *in-vivo* data.

1.4 Data-driven disease progression models

Disease progression models are data-driven approaches to understand the temporal evolution of multi-modal biomarkers and can be used to understand the temporal sequence of disease events in neurodegeneration. Several data-driven approaches have emerged in the last decade [51, 52, 53]. Such data-driven disease progression models can be largely classified into two categories: i) models that estimate the trajectories of biomarkers, and the pathways of progression using longitudinal datasets ii) models that estimate the cascade of biomarker changes after the onset of the disease using cross-sectional datasets.

Longitudinal datasets of neurodegenerative diseases are usually created by including subjects with or at-risk of developing the disease and repeatedly measuring a combination of imaging, fluid, cognitive biomarkers over a period of time. Such datasets are available for AD [54], familial AD [55], and familial FTD [10].

One of the requirements for modeling the biomarker trajectories through regression is to have an independent variable (or an x-axis) that causes the changes in biomarker values. In this case, since the onset of the disease causes the observed changes, time since onset of the disease is the required variable for this purpose. However, this variable is unobservable in practice, as clinical symptoms typically arise years or decades after the onset of such diseases. One of the main challenges of such modeling approaches is therefore to construct the biomarker trajectories using robust proxy measures of this latent independent variable. Some of the different approaches developed for constructing biomarker trajectories in such situations are differential equation models [56, 57], Bayesian mixed-effect models [58], and self-modeling regression [52].

While longitudinal datasets are rich in temporal information of neurodegeneration, they are also time-consuming and expensive to collect. In fast-progressing diseases like CJD, collecting longitudinal datasets is often not feasible. To circumvent this problem, several disease progression models have emerged that only require *cross-sectional* data [53, 59, 60].

Event-based models (EBM) [53, 59, 61] are one such class of models that were developed to estimate a temporal sequence of biomarker abnormality events from case-control cross-sectional data. A biomarker abnormality event is defined as the moment when a biomarker goes from a normal state to an abnormal state, after disease onset. Consequently, such an estimation of temporal progression patterns from cross-sectional data is feasible only when the biomarkers are (or can be approximated to be) monotonically increasing or decreasing in neurodegeneration. Under such conditions, selecting a large cohort of patients in different stages of the disease would result in sampling more abnormal biomarker values for early biomarkers than for late biomarkers. EBMs use data-driven probabilistic methods to exploit this for estimating the temporal sequence of biomarker abnormality events.

Some of the open challenges in data-driven disease progression modeling are: i) robustly dealing with disease heterogeneity, ii) accurately estimating the effect of genetics on disease progression, iii) exploiting disease progression models for diagnosis and prognosis of a patient.

1.5 Research aim

The research described in this thesis aims at the development and validation of novel data-driven disease progression models that provide comprehensive insight into the orderly progression of neurodegenerative diseases using multi-modal in-vivo biomarker data from cross-sectional studies. I focused particularly on EBMs in this thesis since the concept of estimating a temporal sequence of biomarker abnormality events from cross-sectional data intrigued me, and because of the high practical impact of models that do not rely on longitudinal data.

1.6 Outline

This thesis is divided into two parts: The first part focuses on developing a novel and robust disease progression model called discriminative EBM (DEBM) and validating it in AD, FTD, and MS. The second part focuses on further methodological extensions of DEBM and their validations in AD and CJD.

Part I: In Chapter 2, we developed a novel and robust disease progression model called

discriminative EBM (DEBM), which estimates a mean disease progression timeline in a cohort with heterogeneous disease progression patterns. Furthermore, we developed a novel patient-staging approach that estimates the severity of the disease in an individual using the estimated disease progression timeline. We validated the utility of each of these innovations in a synthetic dataset simulating the progression of AD, as well in a large AD cohort. In Chapter 3, we validated the generalizability of the temporal cascades provided by DEBM as well as the original EBM methods in multiple clinical cohorts for AD.

In Chapter 4, we obtained novel insights into the progression of progranulin related FTD, a fast-progressing form of familial FTD, using DEBM. In Chapter 5, we obtained novel insights into the progression of relapse-onset MS using structural, functional and cognitive biomarkers, using DEBM.

We participated in a global challenge to predict the future clinical diagnosis of subjects, volume of ventricles in the brain of these subjects, as well as their cognitive summary scores, in a large AD cohort. We developed novel approaches for these challenges using DEBM, which ended up as the winning entry for the ventricular volume prediction and came second in the overall challenge. The details of the challenge, comparative analysis of all the submitted approaches, as well their brief algorithmic details are provided in Chapter 6.

Part II: In Chapter 7, we developed a novel extension of DEBM to effectively estimate the spatio-temporal disease progression timeline using high-dimensional imaging biomarkers, which was validated using a new deep-learning based simulation framework as well as in a large AD cohort. In Chapter 8, we further developed a novel extension of DEBM to estimate orderings in stratified populations and used it to estimate the effect of *APOE* genotypes on the disease progression timeline of AD. In Chapter 9, we investigated if the *APOE*-specific disease progression timelines of AD constructed in a case-controlled setting are generalizable to a population-based cohort, and can be used there to identify preclinical and prodromal AD cases.

In Chapter 10, we obtained novel insights into the progression of seven molecular subtypes of sporadic CJD and developed a novel approach for *ante-mortem* identification of these subtypes using their disease progression timelines.

Lastly, Chapter 11 discusses the novel contributions in this thesis, and provides a roadmap for further research in this field.

PART I
Discriminative Event Based Modeling (DEBM)

Chapter 2

Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling

This chapter contains the content of the manuscript 'Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. **Vikram Venkatraghavan**, Esther E. Bron, Wiro J. Niessen, Stefan Klein, for the Alzheimer's Disease Neuroimaging Initiative. *NeuroImage*, 186: 518-532, 2019.' An earlier version of this chapter was presented (oral) at the International Conference on Information Processing in Medical Imaging (IPMI) 2017 and at the Dutch Biomedical Engineering Conference (BME) 2017.

Vikram Venkatraghavan's contributions in this chapter are: Conceptualization, methodology, software, validation, and writing.

Abstract

Alzheimer’s Disease (AD) is characterized by a cascade of biomarkers becoming abnormal, the pathophysiology of which is very complex and largely unknown. Event-based modeling (EBM) is a data-driven technique to estimate the sequence in which biomarkers for a disease become abnormal based on cross-sectional data. It can help in understanding the dynamics of disease progression and facilitate early diagnosis and prognosis by staging patients. In this work we propose a novel discriminative approach to EBM, which is shown to be more accurate than existing state-of-the-art EBM methods. The method first estimates for each subject an approximate ordering of events. Subsequently, the central ordering over all subjects is estimated by fitting a generalized Mallows model to these approximate subject-specific orderings based on a novel probabilistic Kendall’s Tau distance. We also introduce the concept of relative distance between events which helps in creating a disease progression timeline. Subsequently, we propose a method to stage subjects by placing them on the estimated disease progression timeline. We evaluated the proposed method on Alzheimer’s Disease Neuroimaging Initiative (ADNI) data and compared the results with existing state-of-the-art EBM methods. We also performed extensive experiments on synthetic data simulating the progression of Alzheimer’s disease. The event orderings obtained on ADNI data seem plausible and are in agreement with the current understanding of progression of AD. The proposed patient staging algorithm performed consistently better than that of state-of-the-art EBM methods. Event orderings obtained in simulation experiments were more accurate than those of other EBM methods and the estimated disease progression timeline was observed to correlate with the timeline of actual disease progression. The results of these experiments are encouraging and suggest that discriminative EBM is a promising approach to disease progression modeling.

2.1 Introduction

Dementia is considered a major global health problem as the number of people living with dementia was estimated to be about 46.8 million in 2015. It is expected to increase to 131.5 million in 2050 [62]. Alzheimer’s Disease (AD) is the most common form of dementia. There is a gradual shift in the definition of AD from it being a clinical-pathologic entity (based on clinical symptoms), to a biological one based on neuropathologic change (change of imaging and non-imaging biomarkers from normal to abnormal) [63]. The latter definition is more useful for understanding the mechanisms of disease progression.

Preventive and supportive therapy for patients at risk of developing dementia due to AD could improve their quality of life and reduce costs related to care and lifestyle changes. To identify the at-risk individuals as well as monitor the effectiveness of these preventive and supportive therapies, methods for accurate patient staging (estimating the disease severity in each individual) are needed. To enable accurate patient staging in an objective and quantitative way, it is important to understand how the different imaging and non-imaging biomarkers progress after disease onset.

Longitudinal models of disease progression, such as [51], reconstruct long-term biomarker trajectories using short term data. [52] estimate these trajectories based on self-modeling

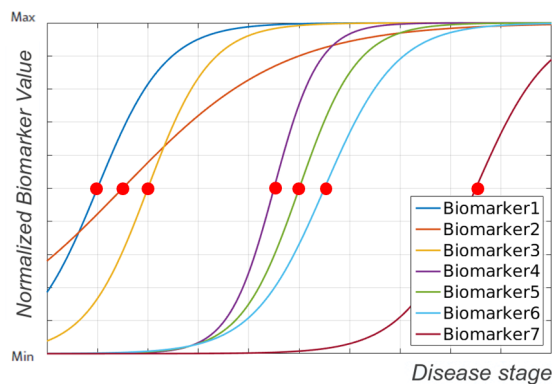


Figure 2.1: Illustration of the output expected in an EBM. The biomarker trajectories shown here are hypothetical trajectories representing a change of biomarker value from normal state. The dots on these trajectories are biomarker events as defined in an EBM. Output of an EBM is the ordering of such events.

regression, whereas Cox regression was used in [64]. Rather than focussing only on a mean trajectory for the entire population, [65] estimate percentile curves based on quantile regression. [66, 67] estimate subject-specific trajectories using a mixed model. [68] provide a probabilistic estimate of biomarker trajectories. While such models are useful for understanding disease progression, their utility in identifying at-risk individuals is restricted. This is due to the fact that selecting a cohort of at-risk individuals for clinical trials based on a longitudinal dataset is not feasible [69]. The utility of these models in studying other forms of dementia is also restricted because longitudinal data in large groups of patients is often scarce.

To circumvent this problem, methods to infer the order in which biomarkers become abnormal during disease progression using cross-sectional data have been proposed [53, 59, 60]. The model used in [60] relies on stratification of patients into several subgroups based on symptomatic staging, for inferring the aforementioned ordering. However, the problem with using symptomatic staging is that it is very coarse and qualitative. The models used in [53, 59] are variants of Event-Based Models (EBM). EBM algorithms neither rely on symptomatic staging nor on the presence of longitudinal data for inferring the temporal ordering of events, where an event is defined by a biomarker becoming abnormal. Figure 2.1 shows these biomarker events on hypothetical trajectories as expected in a typical neuropathologic change.

An important assumption made in [53] is that the ordering of events is common for all the subjects in a dataset. AD is known to be a heterogeneous disease with multiple disease subtypes. The assumptions in Fonteijn’s EBM may therefore be too restrictive. The assumptions in Huang’s EBM on the other hand are more realistic, as they do assume that the disease is heterogeneous. However the algorithm does not scale well to a large number of biomarkers [70].

To make EBM more scalable to large number of biomarkers and subjects, as well as make it robust to variations in ordering, we propose a novel approach to EBM, discriminative event-based model (DEBM), for estimating the ordering of events. We also introduce the concept of relative distance between events which helps in creating a disease progression timeline. Subsequently, we propose a method to stage subjects by placing them on the estimated disease progression timeline. The other contributions of this paper include an optimization technique for Gaussian mixture modeling that helps in accurate estimation of event ordering in DEBM as well as improving the accuracies of other EBMs, and a novel probabilistic distance metric between event orderings (probabilistic Kendall’s Tau).

The remainder of the paper is organized as follows: An introduction to the existing EBM models is given in Section 2.2. In Section 2.3, we propose our novel method for estimating central ordering of events. We perform extensive sets of experiments on ADNI data as well as on simulation data, the details of which are in Section 8.3. Section 9.3 summarizes the results of the experiments. Section 9.4 discusses the implications of these findings followed by concluding remarks in Section 8.6.

2.2 Event-Based Models

EBM assumes monotonic increase or decrease of biomarker values with increase in disease severity (with the exception of measurement noise). It considers disease progression as a series of events, where each event corresponds to a new biomarker becoming abnormal. Fonteijn’s EBM [53] finds the ordering of events (S) such that the likelihood that a dataset was generated from subjects following this event ordering is maximized. S is a set of integer indices of biomarkers, which represents the order in which they become abnormal. Thus, disease progression is defined by $\{E_{S(1)}, E_{S(2)}, \dots, E_{S(N)}\}$, where N is the number of biomarkers per subject in the dataset and $E_{S(i)}$ is the i -th event that is associated with biomarker $S(i)$ becoming abnormal.

In a cross-sectional dataset (X) of M subjects, X_j denotes a measurement of biomarkers for subject $j \in [1, M]$, consisting of N scalar biomarker values $x_{j,i}$. Probabilistic formulation of an EBM, as proposed in [53], can be given by $\argmax_S(p(S|X))$, where $p(S|X)$ can be written using Bayes’ rule as:

$$p(S|X) = \frac{p(S)p(X|S)}{p(X)} \quad (2.1)$$

An important assumption in [53] is that $p(S)$ is uniformly distributed. This makes inferring S ,

equivalent to the maximum likelihood problem of maximizing $p(X|S)^*$. This can be further written in terms of X_j as follows:

$$p(X|S) = \prod_{j=1}^M p(X_j|S) \quad (2.2)$$

where $p(X_j|S)$ can be written as:

$$p(X_j|S) = \sum_{k=0}^N p(k|S)p(X_j|k, S) \quad (2.3)$$

where $p(k|S)$ is the prior probability of a subject being at position k of the event ordering, which is assumed to be equal for each position. The k which maximizes $p(X_j|S)$ denotes subject j 's disease stage. This method of identifying disease severity for a subject results in discrete set of stages, where the number of stages is one more than the number of biomarkers used for creating the model. $p(X_j|k, S)$ can be expressed as:

$$p(X_j|k, S) = \prod_{i=1}^k p(x_{j,S(i)}|E_{S(i)}) \times \prod_{i=k+1}^N p(x_{j,S(i)}|\neg E_{S(i)}) \quad (2.4)$$

where $p(x_{j,S(i)}|E_{S(i)})$ is the likelihood of observing $x_{j,S(i)}$ in subject j , conditioned on event i having already occurred. $p(x_{j,S(i)}|\neg E_{S(i)})$, on the other hand, computes a similar likelihood, given that event i has not occurred.

With the assumption that all the biomarkers in the control population are normal and that the biomarker values follow a Gaussian distribution, $p(x_{j,S(i)}|\neg E_{S(i)})$ is computed. Abnormal biomarker values in the patient population are assumed to follow a uniform distribution but not all biomarkers of a patient could be assumed to be abnormal. For this reason, the likelihoods were obtained using a mixture model of a Gaussian and a uniform distribution, where only the parameters of the uniform distribution were allowed to be optimized.

This method was modified in [61] to estimate the optimal ordering in a sporadic AD dataset with significant proportions of controls expected to have presymptomatic AD [71]. A Gaussian distribution was used to describe both the control and patient population, and

*Fonteinj's EBM uses Markov Chain Monte Carlo (MCMC) sampling to estimate the posterior distribution $P(S|X)$. Average position of events in all the MCMC samples was used as a way for selecting the mean ordering by [53] whereas further extensions of the work such as [61] prefer the maximum likelihood solution.

the mixture model allowed for optimization of parameters for the Gaussians describing both control and patient population. The Gaussian mixture model was also used to incorporate more subjects from the dataset with clinical diagnosis of mild cognitive impairment (MCI).

After obtaining the central ordering S which maximizes the likelihood $p(X|S)$, staging of patients is done by finding a disease stage k for subject j , such that $p(X_j|k, S)$ is maximized.

The assumption that subjects follow a unique event ordering was relaxed by [59], who estimate a distribution of event orderings with a central event ordering (S) and a spread (ϕ) as per a generalized Mallows model [72] using an expectation maximization algorithm. The E-step estimates the likelihood of patients' biomarker value measurements following subject-specific event order s_j , given S and ϕ . In the M-step, S and ϕ are estimated based on s_j estimated in the E-step. This is done iteratively to maximize the likelihood of generation of patients' data based on S and ϕ . Patient staging in Huang's EBM is also a maximum likelihood estimate, but unlike Fonteijn's EBM, the staging is done on the subject-specific event ordering s_j .

In both Fonteijn's and Huang's EBM, relative distances between events, that can be observed in Figure 2.1, are not captured[†]. Some events can be closer to each other than others and using these relative distance between events could help create a more informative disease progression model.

2.3 Discriminative Event-Based Model

Fonteijn's and Huang's EBM are generative models where the likelihood $p(X|S)$ is maximized. Huang's EBM also estimates subject-specific ordering based on a generative approach. Here, we propose our novel method for estimating central ordering of events (S), a discriminative event-based model (DEBM).

The proposed framework is discriminative in nature, since we estimate s_j directly based on the posterior probabilities of individual biomarkers becoming abnormal. We also introduce a new concept of relative distance between events. This subsequently leads to a novel continuous patient staging algorithm. Figure 2.2 shows the different steps involved in our approach.

In Section 2.3.1, we present the method to robustly estimate biomarker distributions in pre-event and post-event classes, given a single cross-sectional measurement of biomarkers. In Section 2.3.2, we present a way for estimating s_j , and we address the problem of estimating a

[†] [53] briefly mention the idea of capturing relative distance between events, but it was not validated or used in any of the experiments.

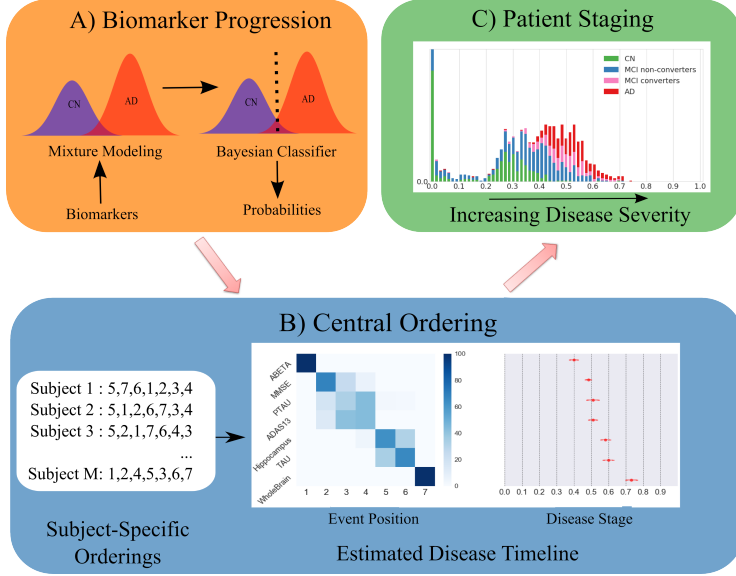


Figure 2.2: Overview of the steps in DEBM. A) Biomarkers measured from different subjects are converted to probabilities of abnormality for individual biomarkers. This is done by estimating normal and abnormal distributions using Gaussian mixture modeling before classifying individual biomarkers using a Bayesian classifier. B) Subject-specific orderings of biomarker abnormalities are inferred from these probabilities which are then used to estimate the central ordering and for creating the disease progression timeline. C) This is then used to stage subjects based on disease severity.

disease timeline from noisy estimates of s_j . In Section 2.3.3, we present the continuous patient staging method.

2.3.1 Biomarker Progression

In this section, we propose a method to robustly convert $x_{j,i}$ to $p(E_i|x_{j,i})$, which denotes the posterior probability of a biomarker measurement being abnormal. Assuming a paradigm similar to that in previous EBM variants [59, 61], the probability density functions (PDF) of pre-event ($p(x_{j,i}|\neg E_i)$) and post-event ($p(x_{j,i}|E_i)$) classes in the biomarkers are assumed to be represented by Gaussians, independently for each biomarker. There are two reasons why constructing these PDFs is non-trivial. Firstly, the labels (clinical diagnoses) for the subjects do not necessarily represent the true labels of all the biomarkers extracted from the subject. Not all biomarkers are abnormal for subjects with AD diagnosis, while some of the cognitively normal (CN) subjects could have undiagnosed pre-symptomatic conditions. Secondly, the clinical diagnosis can be non-binary and include classes such as MCI, with significant number of biomarkers in normal and abnormal classes.

In our approach we address these two issues independently. We make an initial estimate of

the PDFs using biomarkers from easily classifiable CN and easily classifiable AD subjects and later refine the estimated PDF using the entire dataset.

A Bayesian classifier is trained for each biomarker using CN and AD subjects, based on the assumption that there are no biomarkers in the pre-symptomatic stage for CN subjects and all the biomarkers are abnormal for AD subjects. This classifier is subsequently applied to the training data, and the predicted labels are compared with the clinical labels. The misclassified data in the dataset could either be outliers in each class resulting from our aforementioned assumption or could genuinely belong to their respective classes and represent the tails of the true PDFs. Irrespective of the reason of misclassification, we remove them for initial estimation of the PDFs. This procedure thus results, for each biomarker, in a set of easily classifiable CN subjects (whose biomarker values represent normal values) and easily classifiable AD subjects (whose biomarker values represent abnormal values). This is shown in the top part of Figure 2.3.

As we use Gaussians to represent the PDFs, we calculate initial estimates for mean and standard deviation for both normal ($\mu_i^{\neg E}, \sigma_i^{\neg E}$) and abnormal classes (μ_i^E, σ_i^E) based on ‘easy’ CN and ‘easy’ AD subjects for each biomarker i . As these means and standard deviations are estimated based on truncated Gaussians, these are biased estimates. The initial estimates of standard deviations are always smaller than the expected unbiased estimates whereas the initial estimates of means are underestimated for Gaussians with smaller means (as compared to the other class for corresponding biomarkers) and overestimated for Gaussians with larger means.

We refine the initial estimates using a Gaussian mixture model (GMM) and include all the available data, including MCI subjects and previously misclassified cases. To obtain a robust GMM fit, a constrained optimization method is used, with bounds on the means, standard deviations and mixing parameters, based on the aforementioned relationship between the initial estimates and their corresponding expected unbiased estimates. The objective function for optimization for biomarker i is a summation of log-likelihoods, for all subjects:

$$C_i = \sum_{\forall j} \log f(x_{j,i}) \quad (2.5)$$

where the likelihood function $f(x_{j,i})$ is computed as a function of mixing parameters ($\theta_i^E, \theta_i^{\neg E}$) for the groups corresponding to post-event and pre-event respectively and their

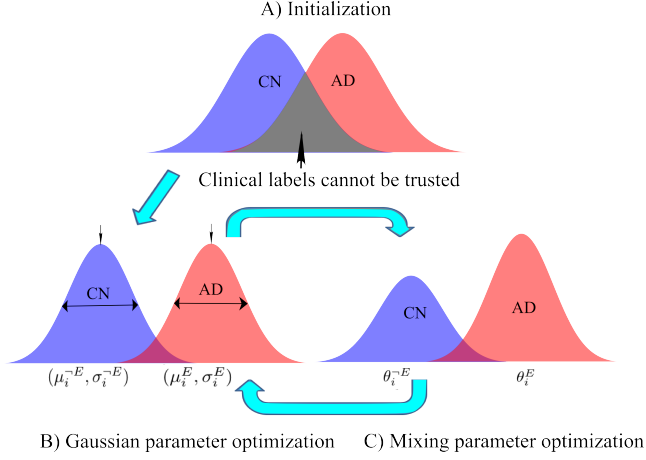


Figure 2.3: Overview of the steps involved in the proposed Gaussian Mixture Model optimization strategy. A) Illustration of the initialization step for Gaussian Mixture Model. Rejecting the tails of the Gaussian distribution in CN and AD class is done to account for the fact that some of the CN subjects could be in pre-symptomatic stage of disease progression and some of the biomarkers could still be normal in AD subjects. B and C) This is followed by iterative estimation of Gaussian parameter optimization and Mixing parameter optimization.

corresponding Gaussian distributions (μ_i^E, σ_i^E) and $(\mu_i^{-E}, \sigma_i^{-E})$:

$$f(x_{j,i}) = \theta_i^E p(x_{j,i} | \mu_i^E, \sigma_i^E) + \theta_i^{-E} p(x_{j,i} | \mu_i^{-E}, \sigma_i^{-E}) = \theta_i^E p(x_{j,i} | E_i) + \theta_i^{-E} p(x_{j,i} | \neg E_i) \quad (2.6)$$

θ_i^E and θ_i^{-E} are selected such that $\theta_i^E + \theta_i^{-E} = 1$. The mixing parameters and the Gaussian parameters are optimized alternately, until convergence of the mixing parameters. The initialization and optimization strategy in GMM is illustrated in Figure 2.3.

The strategy of alternating between optimizing for mixing parameter and optimizing for Gaussian parameters in combination with the initialization strategy and the subsequent constraints is different from all previous versions of EBM and it will be shown in Section 9.3 that this results in more accurate central ordering of events in most cases.

2.3.2 Estimating a disease progression timeline

Estimating Subject-Specific Orderings

The PDF thus obtained is used for classification of the biomarkers using a Bayesian classifier, where the mixing parameters (θ_i^E and θ_i^{-E}) are used as the prior probabilities ($p(E_i)$ and $p(\neg E_i)$ respectively) when estimating posterior probabilities for each biomarker. We assume

these posterior probabilities to be a measure of progression of a biomarker. Thus, s_j is established such that:

$$s_j \ni p(E_{s_j(1)}|x_{j,s_j(1)}) > p(E_{s_j(2)}|x_{j,s_j(2)}) > \dots > p(E_{s_j(N)}|x_{j,s_j(N)}) \quad (2.7)$$

Missing biomarker values are implicitly handled in this definition of s_j , as s_j only consists of events for which biomarkers are present for subject j . The posterior probabilities in Equation 8.3 are influenced not only by progression of the biomarker values to their abnormal states, but also by inherent variability in normal and abnormal biomarker values across subjects, and by measurement noise. Disentangling measurement noise and inherent variability in normal biomarker values from progression of the biomarker to its abnormal state can only be done based on longitudinal data. This makes s_j a noisy estimate.

Estimating a central ordering

Since the event ordering for each subject is estimated independently, any heterogeneity in disease progression is captured in the estimates of s_j . The central event ordering (S) is the mean of the subject-specific estimates of s_j . To describe the distribution of s_j , we make use of a generalized Mallows model. The generalized Mallows model is parameterized by a central ('mean') ordering as well as spread parameters (analogous to the standard deviation in a normal distribution). The central ordering is defined as the ordering that minimizes the sum of distances to all subject-wise orderings s_j . To measure distance between orderings, an often used measure is Kendall's Tau distance [59]. Kendall's Tau distance between a subject specific event ordering (s_j) and central ordering (S) can be defined as:

$$K(S, s_j) = \sum_{i=1}^{N-1} V_i(S, s_j) \quad (2.8)$$

where $V_i(S, s_j)$ is the number of adjacent swaps needed so that event at position i is the same in s_j and S . In case of missing biomarkers, $K(S, s_j)$ is computed for a subset of S consisting only of the events corresponding to the available biomarkers for subject j .

Since the estimates of s_j are based on rankings of posterior probabilities, it would be desirable to penalize certain swaps more than others, based on how close the posterior probabilities are to each other. To this end, we introduce a probabilistic Kendall's Tau distance, which penalizes

each swap based on the difference in posterior probabilities of the corresponding events.

$$\hat{K}(S, s_j) = \sum_{i=1}^{N-1} \hat{V}_i(S, s_j) \quad (2.9)$$

$\hat{V}_i \forall i \in [1, N-1]$ is computed sequentially using the following algorithm[‡]:

Algorithm 1 Probabilistic Kendall Tau distance between Subject-specific event orderings and central event ordering

```

1: for  $i \in [1, N-1]$  do
2:    $k \leftarrow s_j^{-1}(S(i))$ 
3:   if  $k > i$  then
4:      $\hat{V}_i(S, s_j) \leftarrow \sum_{l=i+1}^k p_l - p_i$ 
5:     Move  $s_j(k)$  to position  $i$  and update  $s_j$ 
6:   else
7:      $\hat{V}_i(S, s_j) \leftarrow 0$ 

```

where p_a is shortened notation for $p(E_{s_j(a)} | x_{j, s_j(a)})$.

This variant of Kendall's Tau distance is quite close to the weighted Kendall's Tau distance defined in the permutation space introduced in [73]. The difference stems from the fact that since the probabilistic Kendall's Tau distance is between individual estimates and a central-ordering, the penalization of each swap is weighted asymmetrically as $\hat{V}_i(S, s_j) \neq \hat{V}_i(s_j, S)$.

The optimum S is the one that minimizes $\sum_{j \in J} \hat{K}(S, s_j)$. However, computing a global optimum S based on subject-wise orderings is NP-hard. Thus getting a good initial estimate of S is important to ensure the estimated S is not a suboptimal local optimum. In our implementation the initial estimate of S is based on ordering θ_i^{-E} . The motivation for this is discussed in Section 2.3.3. S was further optimized based on the algorithm introduced by [72] to estimate the central ordering.

Estimating Event Centers

The S that has been derived in this manner, is an estimate of the sequence in which the biomarkers become abnormal during the progression of a disease. However, it falls short of being a disease timeline, because it does not provide information about the proximity of

[‡]The summation symbol in step 4 was missed accidentally in [70].

consecutive events. To address this issue, we estimate distances between events by computing the cost of adjacent swaps in the event ordering, as measured by summation of probabilistic Kendall's Tau distance over all subjects.

$$\Gamma_{i+1,i} = \sum_{\forall j} \widehat{K}(S_{i+1,i}, s_j) - \widehat{K}(S, s_j) \quad (2.10)$$

where $S_{i+1,i}$ is identical to S except for the swap between events at locations i and $i + 1$, and $\Gamma_{i+1,i}$ is the cost of the swap. This represents the cost for the central ordering to be $S_{i+1,i}$ instead of S . We hypothesize that the closer the events $i + 1$ and i are to each other, the lower the swapping cost would be. Hence we consider these costs to be proportional to distance between events in terms of biomarker progression.

To estimate the distance of the first biomarker being abnormal (event) in S to a hypothetical disease-free individual, we introduce a pseudo-event which becomes abnormal at the beginning of the disease timeline and hence is abnormal for all the subjects in the database *i.e.* $p(E_0|x_{j,0}) = 1 \forall j$. Similarly, we introduce another pseudo-event which becomes abnormal at the end of the disease timeline and hence is normal for all the subjects in the database *i.e.* $p(E_{N+1}|x_{j,N+1}) = 0 \forall j$. We scale $\Gamma_{i+1,i} \forall i \in [0, N]$ such that $\sum \bar{\Gamma}_{i+1,i} = 1$. Event center (λ_k) of event k in S for $k > 0$, is computed as follows:

$$\lambda_k = \sum_{i=0}^{k-1} \bar{\Gamma}_{i+1,i} \quad (2.11)$$

In fact, the concept of event centers can also be extended to Fonteijn's EBM by computing the cost of adjacent swaps in the event ordering as the difference in log-likelihoods as follows:

$$\Gamma_{i+1,i} = \log(p(X|S)) - \log(p(X|S_{i+1,i})) \quad (2.12)$$

Extension of this concept to Huang's EBM is not straightforward and is beyond this paper's scope.

The set of event centers $\lambda_{1,2,\dots,N}$, will henceforth be referred to as Λ . This results in a disease timeline, with S giving information about the order of progression of biomarkers and Λ giving information about the event centers in this timeline.

2.3.3 Patient Staging

Once the central ordering of events (S) and event centers (Λ) have been determined, we propose a patient staging algorithm where a patient stage (Υ_j) is interpreted as an expectation of λ_k with respect to the conditional distribution $p(k|S, X_j)$. Thus, Υ_j can be written as given below:

$$\Upsilon_j = \frac{\sum_{k=1}^N \lambda_k p(k|S, X_j)}{\sum_{k=1}^N p(k|S, X_j)} \quad (2.13)$$

Multiplying $p(S, X_j)$ in both numerator and denominator and using the chain rule of probability results in:

$$\Upsilon_j = \frac{\sum_{k=1}^N \lambda_k p(k, S, X_j)}{\sum_{k=1}^N p(k, S, X_j)} \quad (2.14)$$

Using chain rule of probability, we can write $p(k, S, X_j)$ as:

$$p(k, S, X_j) = p(X_j|k, S)p(k, S) \quad (2.15)$$

If we assume a uniform distribution of $p(k|S)$ and $p(S)$ as in [53], $p(k, S, X_j)$ becomes equal to $p(X_j|k, S)$, which was used for patient staging in Fonteijn's EBM as discussed in Section 2.2. However we use prior knowledge in order to define a more informative distribution $p(k, S)$:

$$p(k, S) = \frac{\prod_{i=1}^k \theta_{S(i)}^E \prod_{i=k+1}^N \theta_{S(i)}^{-E}}{Z} \quad (2.16)$$

where Z is a normalizing factor, chosen so as to make this a probability. This choice of $p(k, S)$ can be justified because biomarkers which become abnormal earlier in the disease process are more likely to have a higher value of θ_i^E than the biomarkers which become abnormal later. Hence it is far more likely to have a central-ordering based on ascending values of θ_i^{-E} than an ordering with ascending values of θ_i^E . It should be noted that, the choice of $p(k, S)$ is not unique. For example, it could also be any n -th power of the above equation $\forall n > 0$. Thus, from Equations 8.5, 2.16 and 2.4, we get:

$$p(k, S, X_j) \propto \prod_{i=1}^k p(x_{j,S(i)}|E_{S(i)}) \theta_{S(i)}^E \times \prod_{i=k+1}^N p(x_{j,S(i)}|\neg E_{S(i)}) \theta_{S(i)}^{-E} \quad (2.17)$$

Using the above value of $p(k, S, X_j)$ in Equation 7.2, results in continuous patient stages.

2.4 Experiments

This section describes the experiments performed to benchmark the accuracy of the proposed DEBM algorithm and compare it with state-of-the-art EBM methods. The EBM methods used for comparison in these experiments are Huang’s EBM [59] and the variant of Fonteijn’s EBM that is suited for AD disease progression modeling [61]. The source code for DEBM and Fonteijn’s EBM, with different mixture modeling techniques and patient staging techniques discussed in this paper have been made publicly available online under the GPL 3.0 license: <https://github.com/88vikram/pyebm/>. The source code for Huang’s EBM used in our experiments was provided by the authors of the method.

For brevity, Fonteijn’s EBM and Huang’s EBM will henceforth be referred to as FEBM and HEBM, respectively. The mixture model used with an EBM model (as the one described in Section 2.3.1) will be denoted by a subscript. For example, FEBM with the Gaussian mixture model proposed in [61] will be referred to as FEBM_{ay}. The Gaussian mixture model optimization techniques in [59], [70] and the one introduced in this paper will be denoted with subscripts ‘jh’, ‘vv1’ and ‘vv2’ respectively.[§]

Data used in the experiments were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu)[¶]. We begin with the details of the experiments performed on ADNI data to estimate the event ordering in Section 8.3.2. Since the ground-truth event ordering is unknown for clinical datasets, we resort to using the ability of patient staging to classify AD and CN subjects, as an indirect way of measuring the reliability of the event ordering. We also measure the accuracy of event ordering and relative distance between events more directly by performing extensive experiments on synthetic data simulating the progression of AD. The details of these experiments are given in Section 7.4.2.

[§]Mixture model ‘ay’ optimizes for Gaussian and mixing parameters together. Initialization of Gaussian parameters for optimization is done without rejecting the overlapping part of Gaussians in CN and AD classes. ‘vv1’ also optimizes for Gaussian and mixing parameters together (although with much stricter bounds) but the initialization of Gaussian parameters is similar to the one in this paper. ‘jh’ couples mixture modeling with estimation of subject-specific ordering to estimate a combined optimum solution.

[¶]The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD). For up-to-date information, see www.adni-info.org.

| Demographics | | | | |
|--------------|-----|---------|------------------|------------------|
| Diagnosis | n | Sex M/F | Age [yrs.] | Edu. [yrs.] |
| CN | 417 | 209/208 | 74.76 ± 5.72 | 16.28 ± 2.73 |
| SMC | 106 | 44/62 | 72.20 ± 5.53 | 16.76 ± 2.51 |
| MCI | 872 | 515/357 | 73.00 ± 7.61 | 15.90 ± 2.83 |
| AD | 342 | 189/153 | 75.02 ± 7.78 | 15.17 ± 2.98 |

Table 2.1: Demographics for the whole population.

2.4.1 ADNI Data

We considered 1737 subjects from ADNI 1, Go and 2 (417 CN, 106 with Significant Memory Concern (SMC), 872 MCI and 342 AD subjects) who had a structural MRI (T1w) scan at baseline. Study subject demographics are summarized in Table 8.1. The T1w scans were non-uniformity corrected using the N3 algorithm [74]. This was followed by multi-atlas brain extraction using the method described in [9]. Multi-atlas segmentation was performed [75, 76] using the structural MRI scans to obtain a region-labeling for 83 brain regions in each subject using a set of 30 atlases. Probabilistic tissue segmentations were obtained for white matter, gray matter (GM), and cerebrospinal fluid on the T1w image using the unified tissue segmentation method [77] of SPM8 (Statistical Parametric Mapping, London, UK). The probabilistic GM segmentation was then combined with region labeling to obtain GM volumes in the extracted regions. We also downloaded CSF ($A\beta_{1-42}$ (ABETA), TAU and p-TAU) and cognitive score (MMSE, ADAS-Cog) values from the ADNI database, making the total number of features equal to 88.

The features TAU and p-TAU were transformed to logarithmic scales to make the distributions less skewed. GM volumes of segmented regions were regressed with age, sex and intracranial volume (ICV) and the effects of these factors were subsequently corrected for, before being used as biomarkers. The effect of age and sex was regressed out of CSF based features, whereas effects of age, sex and education was regressed out of cognitive scores.

We retained 52 biomarkers (GM volume based biomarkers of 47 regions, 3 CSF and 2 cognitive scores) having significant differences between CN and AD subjects using Student’s t-test with $p < 0.005$, after Bonferroni correction. These biomarker values were used to perform Experiments 1 and 2.

Experiment 1(a): A subset of 7 biomarkers including the 3 CSF features, MMSE score, ADAS-Cog score, gray matter volume of the hippocampus (combined volume of left and right hippocampi) and gray matter volume in whole brain was created. Event ordering of these 7 biomarkers was inferred using DEBM. We studied the positional variance of central ordering

| Demographics | | | | |
|--------------|-----|---------|------------------|------------------|
| Diagnosis | n | Sex M/F | Age [yrs.] | Edu. [yrs.] |
| CN | 160 | 83/77 | 73.56 ± 5.81 | 16.38 ± 2.66 |
| MCI | 414 | 249/165 | 73.20 ± 7.11 | 16.01 ± 2.79 |
| AD | 216 | 125/91 | 74.36 ± 8.06 | 15.45 ± 2.94 |

Table 2.2: Demographics for the homogeneous subset of subjects.

and variance of event centers inferred by DEBM by creating 100 bootstrapped samples of the data.

Experiment 1(b): The Biomarkers were ranked based on their aforementioned p -value and the above experiment was repeated with top 25 and top 50 biomarkers to investigate if the event-centers estimated for the subset of Biomarkers used in Experiment 1(a), remain comparable to the ones estimated in Experiment 1(a).

Experiment 2: As an indirect way of measuring the accuracy of the estimated event ordering, we use patient staging based on the estimated event orderings as a way to classify CN and AD subjects in the database. 10-fold cross validation was used for this purpose. AUC measures were used to measure the performance of these classifications and thus indirectly hint at the reliability of the event ordering based on which the corresponding patient staging were performed.

We used varying number of biomarkers (ranked based on their p -value) ranging from 5 to 50 in steps of 5 for this experiment. We used the methods FEBM_{ay} , HEBM_{jh} , DEBM_{vv1} and DEBM_{vv2} for inferring the ordering. Patient staging was done based on the methods described in their respective papers. Since the earlier version of DEBM [70] had not introduced a patient staging method, we use the patient staging method described in this paper for evaluating the method.

Experiment 3(a): To study disease progression in a homogeneous population showing signs of typical AD progression, Experiment 1(a) was repeated with a subset of subjects, selected based on their CSF ABETA values. For this experiment, we selected ABETA positive MCI and AD subjects ($\text{ABETA} < 192 \text{ pg/ml}$) and ABETA negative CN subjects ($\text{ABETA} \geq 192 \text{ pg/ml}$). This cut-off was chosen according to the results of [78]. Moreover, we excluded all SMC subjects and subjects with missing ABETA biomarker values. This subset of subjects will henceforth be referred to as the ‘homogeneous subset’. Demographics for the homogeneous subset are summarized in Table 2.ST1. We excluded ABETA biomarker when inferring the event ordering using DEBM.

Experiment 3(b): We retained 49 biomarkers (GM volume based biomarkers of 45 regions, 2 CSF biomarkers excluding ABETA and 2 cognitive scores) having significant differences between CN and AD subjects in the homogeneous subset using Student's t-test with $p < 0.05$, after Bonferroni correction. The biomarkers were ranked based on their aforementioned p -value and the above experiment was repeated with top 24 and top 49 biomarkers, to investigate if the event-centers estimated for the subset of biomarkers used in Experiment 3(a), remain comparable to the ones estimated in Experiment 3(a).

2.4.2 Simulation Data

We used the framework developed by [79] for simulating cross-sectional data consisting of scalar biomarker values for CN, MCI and AD subjects. In this framework, disease progression in a subject is modeled by a cascade of biomarkers becoming abnormal and individual biomarker trajectories are represented by a sigmoid. The equation for generating biomarker values for different subjects is given below:

$$x_{j,i}(\Psi) = \frac{R_i}{1 + \exp(-\rho_i(\Psi - \xi_{j,i}))} + \beta_{j,i} \quad (2.18)$$

Ψ denotes disease stage of a subject which we take to be a random variable distributed uniformly throughout the disease timeline. ρ_i signifies the rate of progression of a biomarker, which we take to be equal for all subjects. $\xi_{j,i}$ denotes the disease stage at which the biomarker becomes abnormal. $\beta_{j,i}$ denotes the value of the biomarker when the subject is normal and R_i denotes the range of the sigmoidal trajectory of the biomarker, which we take to be equal for all subjects.

In our experiments, $\beta_{j,i}$ and $\xi_{j,i} \forall j$ are assumed to be random variables with Normal distribution $\mathbb{N}(\mu_{\beta_i}, \Sigma_{\beta_i})$ and $\mathbb{N}(\mu_{\xi_i}, \Sigma_{\xi_i})$ respectively. μ_{β_i} is equal to the mean value of the corresponding biomarker in the CN group of the selected ADNI data. R_i is equal to the difference between the mean values of the biomarker in the CN and AD groups of the selected ADNI data. Σ_{β_i} represents the variability of biomarker values in the CN group. We consider a relative scale for Σ_{β_i} , where 1 refers to the observed variation among the CN subjects in ADNI data. Variation in $\xi_{j,i}$ is controlled by Σ_{ξ_i} and results in variation in ordering among subjects in population and could be seen as a parameter controlling the disease heterogeneity within a simulated population. $\Sigma_{\xi_i} \forall i$ is varied in multiples of $\Delta\xi$, where $\Delta\xi$ is the average difference between adjacent μ_{ξ_i} . μ_{ξ_i} refers to the event centers of various biomarkers. The set of $\mu_{\xi_i} \forall i$ will collectively be referred to as Λ_{gt} and they will be used to assess the accuracy of estimated event centers (λ_i).

The parameters in the simulation framework that could have an effect on the performance of

EBMs are Σ_{β_i} , μ_{ξ_i} , Σ_{ξ_i} , and ρ_i . Apart from this, the number of subjects (M) and the number of biomarkers (N) in the dataset could also have an effect on the performance of EBMs. Using this simulation framework, we study the effect of the aforementioned parameters on the ability of different variants of EBM algorithms to accurately infer the ground-truth central ordering in the population. Change in μ_{β_i} results only in a translational effect on biomarker values and change in R_i results only in a scaling effect on biomarker values. These factors do not affect the performance of the EBMs and hence were not evaluated in our experiments.

Performance of an EBM method can be measured using error in estimation of either S or Λ . Error in estimating S (ϵ_S) will henceforth be referred to as ‘ordering error’ whereas the error in estimating Λ (ϵ_Λ) will henceforth be referred to as ‘event-center error’. ϵ_S is computed using the following equation:

$$\epsilon_S = \frac{K(S, S_{gt})}{\binom{N}{2}} \quad (2.19)$$

where S_{gt} is the ground truth ordering. ϵ_S is effectively a normalized Kendall’s Tau distance between S and S_{gt} . The normalization factor for $\binom{N}{2}$, was chosen to make the accuracy measure interpretable for different number of biomarkers.

For comparing Λ and Λ_{gt} , Λ were scaled and translated such that the mean and standard deviation of Λ were equal to that of Λ_{gt} . This is done because we are only interested in evaluating the errors in estimating relative distance between events and not the absolute position of event-centers. The choice of scale in event-centers are arbitrary and the chosen scale for the estimated event-centers was based on pseudo-events, which need not necessarily coincide with the simulation framework’s ground-truth event-centers.

$$\epsilon_\Lambda = \sum_{\forall i} |\lambda_i^{st} - \mu_{\xi_i}| \quad (2.20)$$

where λ_i^{st} is the scaled and translated version of λ_i .

As mentioned before, the factors that can have an effect on the performance of EBMs are Σ_{β_i} , μ_{ξ_i} , Σ_{ξ_i} , ρ_i , M and N . In each of the following 5 experiments, a few of these factors were varied while the others were set to their default values. The default value for Σ_{β_i} was taken to be 1 as this corresponds to the observed variation among CN subjects in ADNI. μ_{ξ_i} were spaced equidistantly, i.e., $\mu_{\xi_{i+1}} - \mu_{\xi_i} = 1/(N + 1)$. As the actual variation in event centers among different subjects is not known in a clinical dataset, the default value of Σ_{ξ_i} was taken to be $2\Delta\xi$. For the sake of simplicity of notation $\Delta\xi$ will be omitted henceforth, and the values

of Σ_{ξ_i} are implicitly in multiples of $\Delta\xi$. ρ_i was considered to be equal for all biomarkers by default. The default values for M and N were 1737 and 7 respectively, mimicking the dataset used in Experiment 1(a). For each simulation setting, 50 repetitions of simulation data were created and used for benchmarking the performance of EBMs on synthetic data.

Experiment 4: The first simulation experiment was performed to study the effect of $\Sigma_{\beta} \in [0.2, 1.8]$ and $\Sigma_{\xi} \in [0, 4]$, varying one at a time while keeping the other at its mean value. The ϵ_S of FEBM_{ay}, FEBM_{vv2}, HEBM_{jh}, HEBM_{vv2}, DEBM_{vv1} and DEBM_{vv2} were determined.

Experiment 5: The above experiment was repeated for DEBM_{vv2} and FEBM_{vv2} and the ϵ_{Λ} were measured for the two methods.

Experiment 6: This experiment was performed to study the effect of a non-uniform distribution of μ_{ξ_i} . Σ_{β} and Σ_{ξ} combinations of (0.6, 1), (1.0, 2), (1.4, 3) and (1.8, 4) were tested to study their effect in non-uniformly spaced biomarkers. ϵ_S of DEBM_{vv2}, FEBM_{vv2} and HEBM_{vv2} were measured. Additionally, ϵ_{Λ} of DEBM_{vv2} and FEBM_{vv2} were measured. To also study the effect of unequal rates of progression of biomarkers (ρ_i), the above experiment was performed once with equal ρ_i for all biomarkers and once when they were unequal. The experiment with unequal biomarker rates had the same mean biomarker progression rate as the the experiment with equal biomarker rates. The progression rates of different biomarkers has been included as supplementary material (Figure 2.SF1).

Experiment 7: This experiment was performed to study the influence of the number of subjects (M). M was varied from 100 to 2100 in steps of 200. ϵ_S of DEBM_{vv2}, FEBM_{vv2} and HEBM_{vv2} were measured. DEBM_{vv2} and FEBM_{vv2} were also assessed based on ϵ_{Λ} .

Experiment 8: This experiment was performed to study the influence of the number of biomarkers (N). N was varied from 7 to 52 in steps of 5. In each random generation of a dataset, we randomly selected (with replacement) the biomarkers to be used in the iteration. This was done to study the effect of N on the EBM models and separate it from the effect of adding weaker biomarkers. ϵ_S of DEBM_{vv2}, FEBM_{vv2} and HEBM_{vv2} were measured. DEBM_{vv2} and FEBM_{vv2} were also assessed based on ϵ_{Λ} .

2.5 Results

2.5.1 ADNI Data

Experiment 1: Figure 2.4 shows the positional variance and event-center variance obtained using DEBM_{vv2} with 7 events. The Gaussian mixture model parameters for every biomarker have been tabulated in the supplementary material, Table 2.ST1

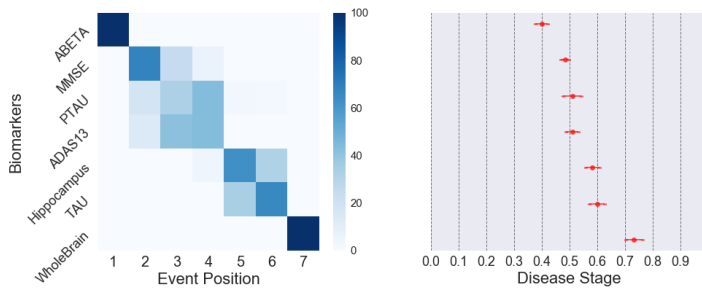


Figure 2.4: Experiment 1(a): DEBM_{vv2} with 7 Events. The positional variance diagram (left) shows the uncertainty in estimating the central event ordering. The event-center variance diagram (right) shows the standard error of estimated event centers. These were measured by 100 repetitions of bootstrapping.

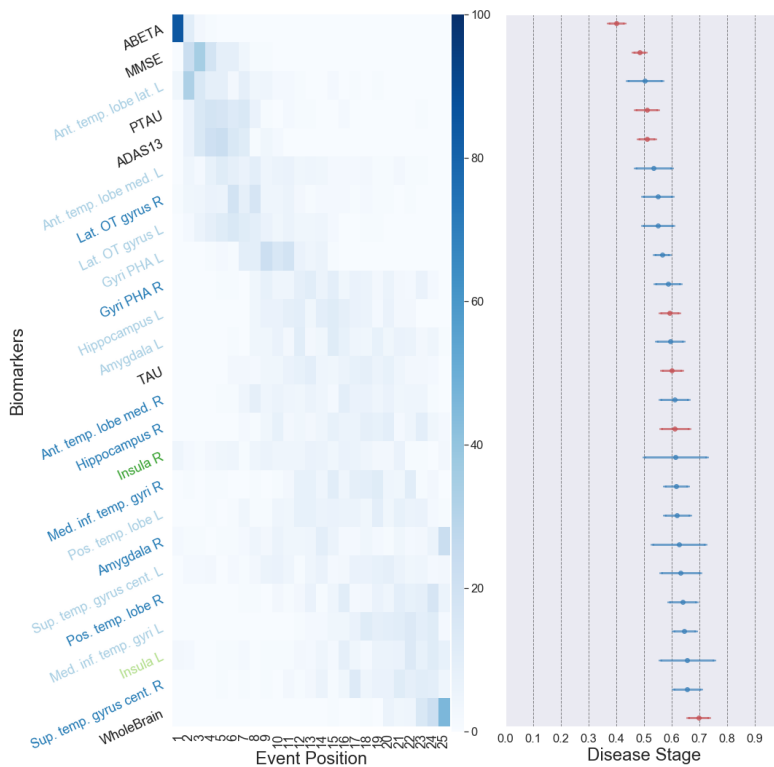


Figure 2.5: Experiment 1(b): DEBM_{vv2} with 25 Events. The positional variance diagram (left) shows the uncertainty in estimating the central event ordering and the event-center variance diagram (right) shows the standard error of estimated event centers. These were measured by 100 repetitions of bootstrapping. The event centers of the biomarkers used in Figure 2.4 are marked in red. Table 2.3 shows the full forms of the abbreviations used in the y-axis labels. Figure 2.7 maps the colors used for y-axis labels to different lobes in the brain.



Figure 2.6: Experiment 1(b): DEBM_{vv2} with 50 Events. Positional variance diagram (left) shows the uncertainty in estimating the central event ordering and event center variance diagram (right) shows the standard error of estimated event-centers. These were measured by 100 repetitions of bootstrapping. The event-centers of the biomarkers used in Figure 2.4 are marked in red, whereas the ones used in Figure 2.5 are marked in blue. Table 2.3 shows the full forms of the abbreviations used in the y-axis labels. Figure 2.7 maps the colors used for y-axis labels to different lobes in the brain.

It can be seen from Figure 2.4 (left) that CSF-based biomarkers ABETA becomes abnormal before MMSE and CSF-based p-TAU. This is followed by ADAS13, Hippocampal volume, TAU and whole brain volume events. However Figure 2.4 (right) shows that the event centers

| Abbreviation | Full name |
|--------------|------------------------------|
| L | Left |
| R | Right |
| PHA | Parahippocampalis et Ambiens |
| Med. | Medial |
| Inf. | Inferior |
| Sup. | Superior |
| Temp. | Temporal |
| Pos. | Posterior |
| Lat. | Lateral |
| Ant. | Anterior |
| OT | Occipitotemporal |
| Cent. | Central |
| Mid. | Middle |
| Rem. | Remainder |
| Occ. | Occipital |
| PS | Pre-subgenual |

Table 2.3: Abbreviations used in Figures 2.5 and 2.6 along with their full names [75].

| | |
|-------------------------|--------------------------|
| — Temporal Lobe (Left) | — Temporal Lobe (Right) |
| — Central Lobe (Left) | — Central Lobe (Right) |
| — Frontal Lobe (Left) | — Frontal Lobe (Right) |
| — Parietal Lobe (Left) | — Parietal Lobe (Right) |
| — Occipital Lobe (Left) | — Occipital Lobe (Right) |

Figure 2.7: Legend for the colors used in Figures 2.5 and 2.6. The colors map different biomarker labels to lobes in the brain.

for MMSE, ADAS13, p-TAU are close to each other and so are the event-centers of TAU and hippocampus volume. The event associated with the TAU biomarker seems closer to the whole brain volume event as they are in positions 6 and 7 of Figure 2.4 (left). However, the centers of these two events are quite far apart in Figure 2.4 (right) and the p-TAU event (position 2) is closer to the TAU event than whole brain volume event.

As the number of biomarkers increases, the variation in the positions also increases considerably, as seen in Figures 2.5 (left) and 2.6 (left). The event centers of the biomarkers used in Experiment 1(a) remain fairly consistent (± 0.05) in Experiment 1(b). It can also be seen that biomarkers with lower p -values (biomarkers included in the model with 50 biomarkers and not in the model with 25 biomarkers), have larger variance in their event-center estimation.

Experiment 2: Figure 2.8 (a) shows the mean AUC when using patient stages for classifying

CN versus AD subjects using DEBM and other variants of EBM methods. It can be observed that the AUC of all the methods decreases as the number of events increases. The proposed method DEBM_{vv2} followed by the proposed patient staging algorithm outperforms all the existing EBM variants consistently.

Figure 2.8 (b) shows the distribution of patient stages for the whole population when the most significant 25 features were given as input to DEBM_{vv2}. This graph shows a peak at disease stage 0 dominated by CN and MCI non-converters, which shows that these subjects are not progressing towards AD. The non-zero lower disease stages are dominated by CN subjects and MCI non-converters, whereas MCI converters¹¹ and the subjects with AD have higher disease stages.

Experiment 3: Figure 2.9 shows the positional variance and event-center variance obtained using DEBM_{vv2} with 6 events, in the homogeneous subset of subjects. It can be seen from Figure 2.9 that in the homogeneous subset of subjects, p-TAU event occurs before ADAS13 and MMSE events as opposed to p-TAU event occurring after ADAS13 and MMSE in Figure 2.4. It can also be seen from Figure 2.9 that the TAU event precedes Hippocampus volume event as opposed to Hippocampus event preceding the TAU event in Figure 2.4.

The results of Experiment 3(b) with 24 and 49 have been included as supplementary material (Figures 2.SF2 and 2.SF3).

2.5.2 Simulation Data

Experiment 4: Figures 2.10 shows the ordering errors of DEBM, FEBM and HEBM models with different mixture models as Σ_β and Σ_ξ increase. The error-bars depict mean and standard deviation of the errors obtained in 50 repetitions of simulations. It can be seen that the proposed optimization technique improves the performance of all three EBM models. The change is particularly evident when comparing the performance of FEBM_{vv2} and FEBM_{ay}.

It can also be seen that FEBM_{vv2} performs slightly better than DEBM_{vv2} when Σ_ξ is low, but as Σ_ξ increases, the performance of FEBM_{vv2} degrades significantly. The performance of HEBM is almost always worse than its FEBM or DEBM counterpart.

Experiment 5: Figure 2.11 (a) and (b) shows the event-center errors in DEBM_{vv2} and FEBM_{vv2} as the variability in population (Σ_β) and disease heterogeneity (Σ_ξ) increases respectively. It should be noted from Figure 2.10(b) and Figure 2.11 (b) that, even when the FEBM_{vv2} gets the ordering more accurately than DEBM_{vv2} in cases of low Σ_ξ , the event-center estimation of DEBM_{vv2} is on par with or better than its FEBM counterpart.

¹¹ MCI converters are subjects who convert to AD within 3 years of baseline measurement

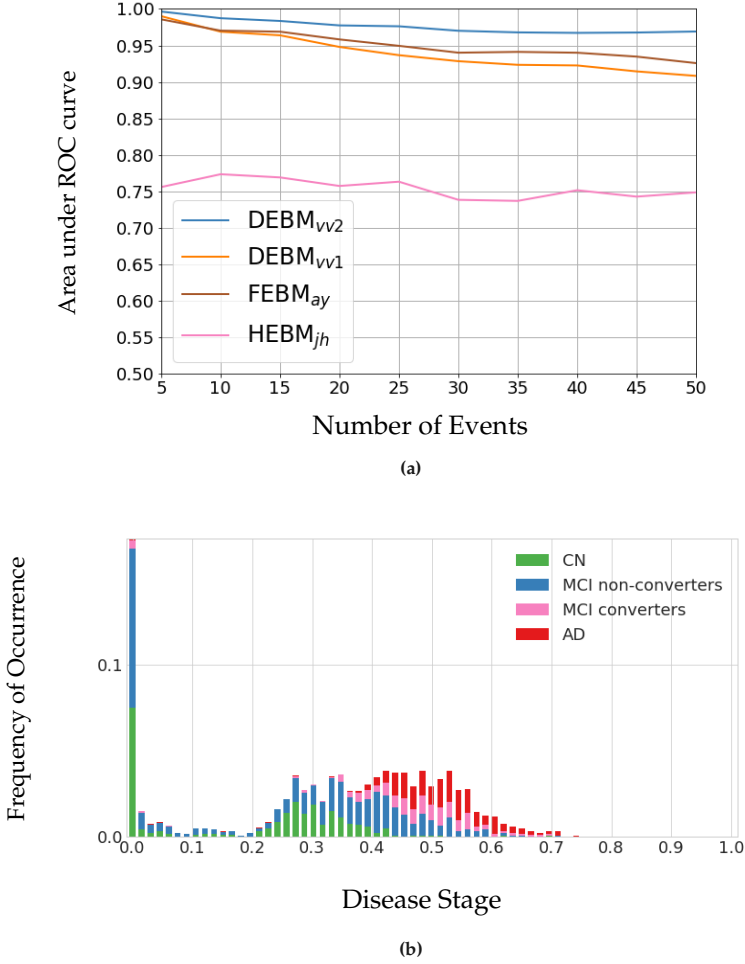


Figure 2.8: Experiment 2: In (a) we see the variation of AUC with respect the number of biomarkers used for building the model using DEBM, when the obtained patient stages were used for classification of CN versus AD subjects. The AUC measure was obtained using 10-fold cross-validation. In (b) we see the frequency of occurrence of subjects in different disease stages, when the most significant 25 features were given as input to DEBM_{vv2} for inferring the ordering as well as for patient staging.

Figure 2.11 (c) shows the estimated event-center locations for $\Sigma_\beta = 1.0$ and $\Sigma_\xi = 2$ and the ground truth event-centers.

Experiment 6: Figure 2.12 (a) shows the ordering errors of DEBM_{vv2}, FEBM_{vv2} and HEBM_{vv2} as Σ_β and Σ_ξ increase, when the ground-truth event centers (μ_{ξ_i}) are non-uniformly spaced. The spacing of μ_{ξ_i} can be observed in Figure 2.12 (b), where the ground truth event-centers

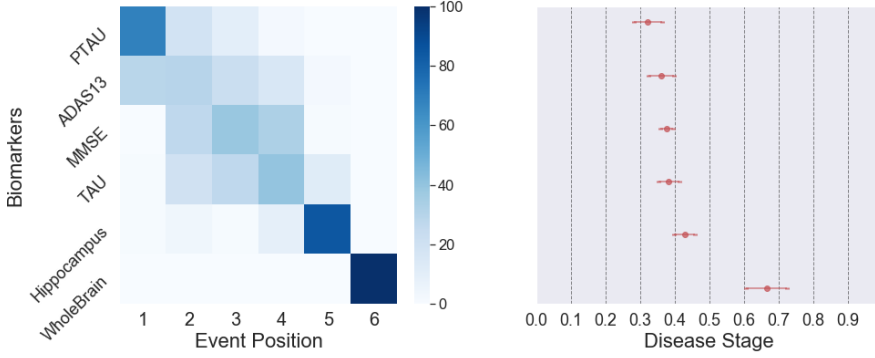


Figure 2.9: Experiment 3(a): DEBM_{vv2} with 6 Events, in the homogeneous subset of subjects. The positional variance diagram (left) shows the uncertainty in estimating the central event ordering. The event-center variance diagram (right) shows the standard error of estimated event centers. These were measured by 100 repetitions of bootstrapping.

as well as the estimated event-centers of DEBM_{vv2} and FEBM_{vv2} are shown for $\Sigma_\beta = 1.0$ and $\Sigma_\xi = 2$. It can be observed that the estimated event-centers for DEBM_{vv2} are much closer to the ground-truth event centers than those of FEBM_{vv2} and also have a much lower variance over different iterations of simulations.

Figure 2.12 (c) shows the ordering errors as Σ_β and Σ_ξ increases, when μ_{ξ_i} is non-uniformly spaced and ρ_i is not identical for all biomarkers. It should also be noted that the mean of ρ_i over all i has not changed between (a) and (c). The variation of errors in (c) is quite similar to the one in (a). This shows that performance of EBM methods that are reported in other experiments (where ρ_i is equal for all biomarkers) can be expected to not deteriorate in the more realistic scenario of ρ_i not being equal for all biomarkers. The event-center variance for $\Sigma_\beta = 1.0$ and $\Sigma_\xi = 2$ for the case of unequal ρ_i is very similar to (b) and has been included as supplementary material (Figure 2.SF4).

Experiment 7: Figure 2.13 shows the mean ordering errors of DEBM_{vv2}, FEBM_{vv2} and HEBM_{vv2} as a function of number of subjects in the dataset on one vertical axis and shows the mean event-center errors of DEBM_{vv2} and FEBM_{vv2} on the other vertical axis. As expected, the models perform better as the number of subjects increases. DEBM_{vv2} is slightly better at inferring the central ordering than FEBM_{vv2} when the number of subjects is very low, but FEBM_{vv2} outperforms DEBM_{vv2} when the number of subjects is higher. However, when the accuracy of event centers are considered, DEBM_{vv2} consistently outperforms FEBM_{vv2}.

Experiment 8: Figure 2.14 shows the mean ordering errors of DEBM_{vv2}, FEBM_{vv2} and HEBM_{vv2} as a function of the number of events (biomarkers) in the dataset on one vertical axis

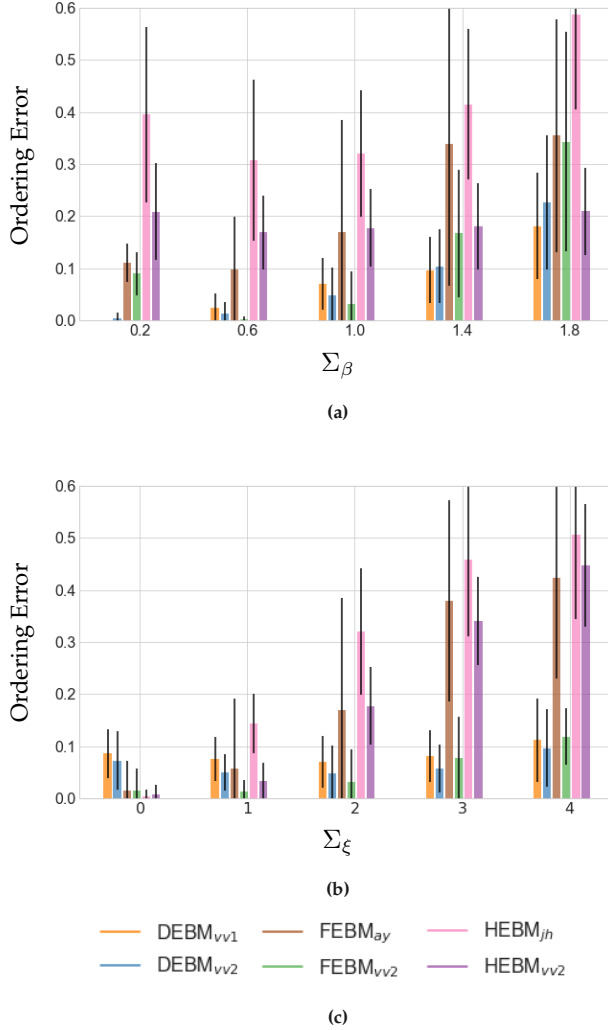
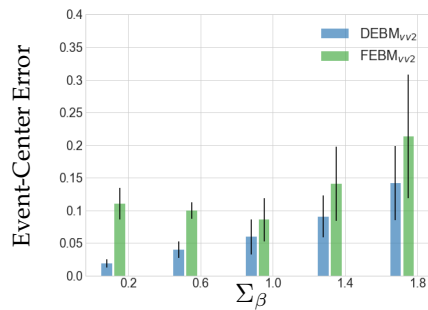
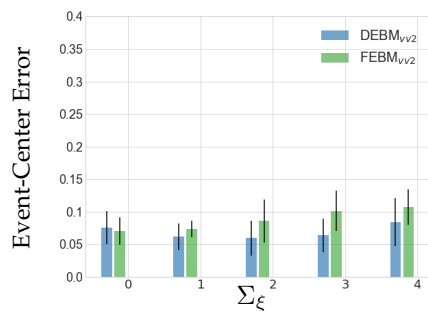


Figure 2.10: Experiment 4: Ordering errors of DEBM_{vv1}, DEBM_{vv2}, FEBM_{ay}, FEBM_{vv2}, HEBM_{jh} and HEBM_{vv2} for 50 repetitions of simulations. Figure (a) shows the ordering error as a function of variability in population (Σ_β). Figure (a) shows the ordering error as a function of variation in ordering (Σ_ξ). Error bars in (a) and (b) represent standard deviations over the 50 repetitions. Figure (c) shows the legend for the plots in (a) and (b).

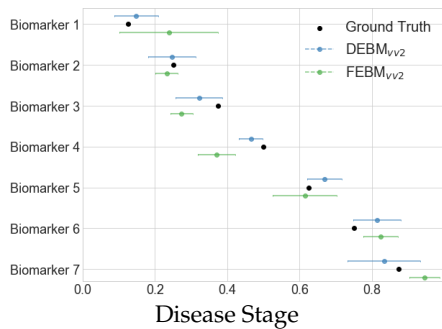
and shows the mean event-center errors of DEBM_{vv2} and FEBM_{vv2} on the other vertical axis. The biomarkers were selected randomly after replacement so that the chances of selecting a bad biomarker remain equal as the number of events increases. It can be noted that the errors of the EBM models increase as the number of events increases initially, even when the average



(a)



(b)



(c)

Figure 2.11: Experiment 5: Figures (a) and (b) show the event-center errors of DEBM_{vv2} and FEBM_{vv2} as a function of Σ_β and Σ_ξ respectively. Figure (c) shows the estimated event-center locations for both methods as well as the ground-truth event centers. Error bars in (a), (b) and (c) represent standard deviation over 50 repetitions of simulation.

quality of biomarkers remains the same. However the errors stabilize beyond a certain point and do not increase any more.

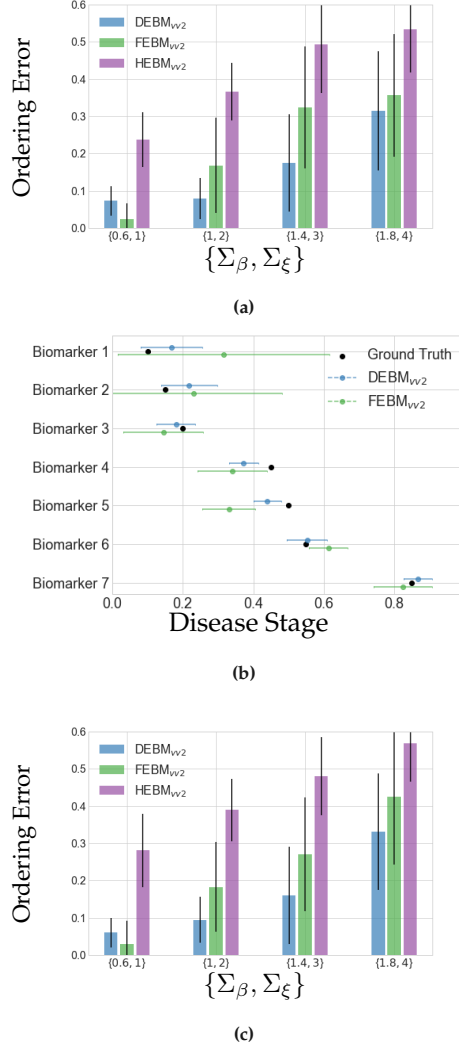


Figure 2.12: Experiment 6: Figures (a) and (c) show the ordering errors of DEBM_{vv2}, FEBM_{vv2} and HEBM_{vv2} when μ_{ξ_i} are not uniformly distributed. Σ_β and Σ_ξ increase as we move from left to right. Figure (a) shows the errors in the case when ρ_i are identical for all the biomarkers whereas (c) shows the errors when ρ_i are different. Figure (b) shows the non-uniform μ_{ξ_i} as well as the estimated event-centers by DEBM_{vv2} and FEBM_{vv2} for the case of ρ_i being equal. Error bars in (a), (b) and (c) represent standard deviation over 50 repetitions of simulation.

2.6 Discussion

We proposed a novel discriminative EBM framework to estimate the ordering in which biomarkers become abnormal during disease progression, based on a cross-sectional dataset.

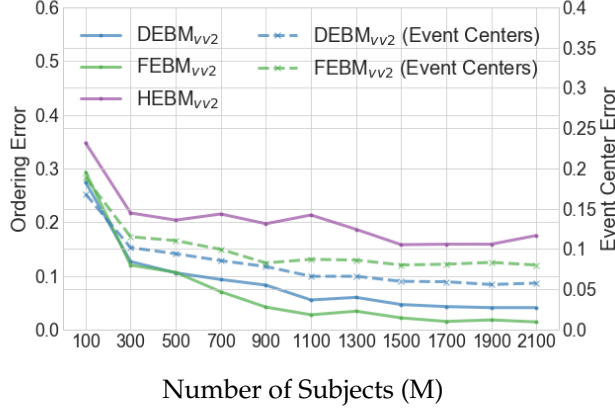


Figure 2.13: Experiment 7: Ordering errors of $DEBM_{vv2}$, $FEBM_{vv2}$ and $HEBM_{vv2}$ as a function of number of subjects (M) in the dataset. It also shows the event-center errors of $DEBM_{vv2}$ and $FEBM_{vv2}$ as a function of M .

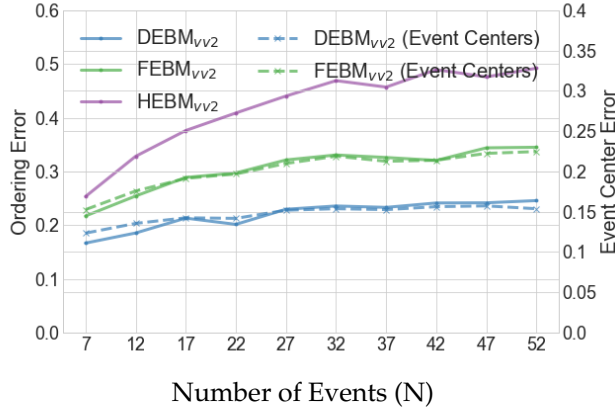


Figure 2.14: Experiment 8: Ordering errors of $DEBM_{vv2}$, $FEBM_{vv2}$ and $HEBM_{vv2}$ as a function of number of events (N) in the dataset. It also shows the event-center errors of $DEBM_{vv2}$ and $FEBM_{vv2}$ as a function of N .

The proposed framework outperforms state-of-the-art EBM techniques in estimating the event ordering. We also introduced the concept of relative distance between event-centers, which enables creating a disease progression timeline. This in turn led to the development of a new continuous patient staging mechanism. In addition to the framework, we also proposed a novel probabilistic Kendall's Tau distance metric and a robust biomarker distribution estimation algorithm. In this section, we discuss different aspects of the proposed algorithm.

2.6.1 Event Centers

Event-centers capture relative distance between events. This helps in creating the disease progression timeline from an ordering of events. If an event (Event A) leads to another event (Event B), this would be observed as event-center for A occurring before event-center for B. However EBMs cannot assess causality, and cannot distinguish the aforementioned case from the case when Event B is caused by some external factor which happened to occur after Event A.

Event centers are an intrinsic property of the biomarker used, for the selected population. This was observed in Experiment 1(b) where the event-centers estimated using $DEBM_{vv2}$ remained fairly consistent (± 0.05) across models using different number of biomarkers.

The estimated disease progression timeline can be used for inferring progression of the disease, with the event centers being synonymous to milestones of progression. A strict quantization of position in ordering of events (as reported in [80], [70], [81], [61], [53]) in the positional variance diagram can sometimes be non-intuitive in terms of inferring actual progression of the disease. This was seen in Experiment 1, where the event center variance diagram showed that the TAU event (at position 6) was closer to the p-TAU event (at position 2) than the whole brain event (position 7).

The approach of scaling the event-centers between $[0, 1]$ has its advantages and disadvantages. The advantage of such a scaling is that models built on different biomarkers, but within the same population, remain comparable. For example, a model built with CSF and MRI based biomarkers can be compared with a model built on MRI based biomarkers alone, as the event-centers of MRI based biomarkers would approximately be the same. On the other hand, the position of the first event relies heavily on the number of ‘true’ controls in the dataset (CN subjects who are not in an early asymptomatic stage of the disease). This is the result of introducing pseudo-events for scaling the events-centers.

Comparison of the event centers across different datasets with different number of controls (albeit with the same biomarkers) can be done in three ways. Event-centers can be scaled and translated such that the mean and standard deviation of event centers computed across different datasets are the same (similar to the comparisons between estimated and ground-truth event centers in this paper). Alternately, the event center of the first biomarker can be set as 0 and the event center of the last biomarker can be set to 1, before comparison. Lastly, in a dataset where controls (i.e., subjects whose biomarker values are all normal) can be easily identified, it would be better to exclude them for event-center computation.

The estimated event centers have a good correlation with the groundtruth disease timeline. This can be seen in the simulation experiments with and without uniform spacing of events

(Experiments 5 and 6). It must however be noted that, the disease stages Ψ of the simulated subjects were distributed uniformly throughout the disease timeline. If the distribution is not uniform, we expect it to have an effect on the estimation of event centers. Analyzing the exact effect of such non-uniform distributions on the estimation of event centers and ways to estimate event centers invariant to the distribution of subjects on the disease timeline could be an interesting extension of the current work.

Experiment 6 also showed that different biomarkers having different rates of progression does not degrade the performance of EBM models, as long as the mean rate of progression is the same. We did not perform an experiment to benchmark the accuracies by changing the mean rates of progression of biomarkers. This experiment was already performed in [79] and it was observed that FEBM ordering error decreases as the mean rates of progression increase.

FEBM assumes that the disease is homogeneous, as it expects all the subjects in the dataset to follow the same ordering. When the variability of ordering in different subjects is low, FEBM with the proposed mixture model ‘vv2’ outperforms DEBM with the proposed mixture model. This can be seen in the results of Experiments 4, 6 and 7. When the assumption becomes too restrictive, DEBM with the proposed mixture model outperforms FEBM. Even when the assumption holds true, estimation of event-centers with DEBM is more accurate than with FEBM.

2.6.2 Patient Staging

Existing patient staging algorithms discretize the patient stages based on event position, whereas the patient staging algorithm introduced in this paper takes relative distance between events into consideration while staging new subjects. This makes patient stages more useful for diagnosis and prognosis as they correlate more with the actual disease progression timeline. Discrete patient stages without considering the event centers could diminish the prognosis value of the obtained stages.

The cross-validation experiment on ADNI data (Experiment 2) showed that the CN and AD subjects are well separated after patient staging and that the AUC of the proposed method is better than that of the state-of-the-art EBM techniques. It also showed that MCI converters and non-converters are well separated after patient staging, without explicitly training the model to achieve this.

It must however be noted that even though heterogeneity of the disease was considered while inferring the central ordering, it was not considered for patient staging. Inferring multiple central orderings corresponding to different disease subtypes [81] and staging patients on one of these central orderings may help us overcome this drawback. Patient staging with respect to

subject-specific orderings (as done in HEBM) can also be considered when extending DEBM for longitudinal data, where the subject-specific orderings might be estimated with higher confidence.

2.6.3 Scalability of Event-Based Models

Understanding the progression of several imaging and non-imaging biomarkers after disease onset is important for assessing the severity of the disease. Hence it is desirable to have a model scalable to a large number of biomarkers. FEBM and DEBM are scalable to large number of events, whereas HEBM is not. This was seen in the simulation experiment on varying number of events (Experiment 8), where the errors of FEBM and DEBM increased asymptotically with increasing number of events. The ordering errors of HEBM reached 0.5 for large number of events, which is equivalent to random prediction.

In Experiment 7, we observed that the errors of the EBM decrease with increasing number of subjects in the dataset. We hence expect FEBM, DEBM and HEBM_{vv2} to be scalable to a large number of subjects.

The performance of HEBM_{jh} is seen to be consistently worse than FEBM_{ay} in Experiment 4. This is in contrast with the findings of [70], where HEBM_{jh} performed better than FEBM_{ay} when the number of biomarkers used were 7, while it performed worse when the number of biomarkers used were 42. One of the key differences between the experiment performed in [70] and Experiment 4 is the number of subjects in the simulation dataset. While the previous study considered 509 subjects, Experiment 4 considered 1737 subjects. HEBM_{jh} jointly estimates the subject-specific orderings of all the subjects and the mixture model to represent the biomarkers in different diagnostic groups. We think that while the joint estimation was good for low number of subjects, increasing the number of subjects had an adverse effect on the convergence of the algorithm. Hence HEBM_{jh} is not scalable to a large number of subjects.

We decoupled the mixture model and estimation of subject-specific orderings in HEBM_{vv2} (Experiments 4, 6, 7 and 8). This made HEBM more scalable as it improved the results in Experiment 4 with 1737 subjects, but the decoupling had an adverse effect on the algorithm when the number of subjects was low, as seen in Experiment 7, where HEBM_{vv2} performs worse than FEBM_{vv2} even when the number of subjects was low.

FEBM and HEBM are generative approaches for estimating the central ordering. Our results suggest that HEBM is not very scalable. Although FEBM is scalable, the assumptions made in FEBM are too restrictive for heterogeneous disease such as AD. DEBM is a discriminative approach to event-based modeling, which is both scalable and can robustly estimate central

ordering even when the disease is heterogeneous.

2.6.4 The Mixture Model

The optimization technique for the Gaussian mixture model that is presented in this paper decouples the optimization of Gaussian parameters and mixing parameters. When the Gaussians of the pre-event and post-event classes are highly overlapping, the optimum mixing parameter changes a lot even for small changes in the Gaussian parameters. By decoupling the optimizations for Gaussian parameters and mixing parameters, we get more stable mixing parameters. This helps in improving the accuracy of all EBMs. This was observed in Experiment 4.

2.6.5 The Importance of Good Biomarkers

Quality of biomarkers plays a huge role in the accuracy of the EBMs. This was seen in Experiment 8, where the mean error value for 7 biomarkers was considerably higher than the mean error value with the same number of biomarkers in Experiment 5 (for the same Σ_β and Σ_ξ parameters). The observed difference can be explained by the choice of the biomarkers used in those experiments. While the biomarkers chosen in Experiment 8 was at random, the ones chosen in Experiment 5 were the 7 best biomarkers.

2.6.6 Interpretation of model results on ADNI

Experiment 1(a) showed that CSF biomarker ABETA is the first biomarker to become abnormal, followed by MMSE, p-TAU and ADAS13. However, Experiment 3(a) showed that in the homogeneous subset of subjects showing signs of typical AD progression (with ABETA positive subjects in MCI and AD, and with ABETA negative CN subjects) p-TAU becomes abnormal before cognitive biomarkers of ADAS13 and MMSE, which is in agreement with Jack’s hypothetical model [82]. The earlier position of MMSE in Experiment 1 as compared to Experiment 3 can be attributed to the inclusion of SMC subjects as well, who need not necessarily be progressing towards AD. The ordering of p-TAU becoming abnormal before ADAS13 which is then followed by Hippocampus was also observed by [52] in ABETA positive subjects.

ADAS13 and MMSE are seen to become abnormal quite early in the disease progression timeline in Experiments 1 and 3. This is in agreement with other studies on prodromal Alzheimer’s Disease[83, 84]. Cognitive biomarkers becoming abnormal before abnormality in Hippocampus and other structural biomarkers, as seen in Experiments 1 and 3, could be due to the fact that the region-based volumes from structural MRI may not be sensitive enough to detect mild structural changes.

The event centers of Hippocampus volume and TAU are quite close to each other in both Experiment 1(a) and 3(a), which is also in agreement with the current understanding of the disease [85].

Figure 2.5 shows that abnormality in the anterior temporal lobe precedes that of the posterior temporal lobe. This was also observed by [67], where the anterior temporal lobe had a higher averaged acceleration factor than the posterior temporal lobe, in a study on AD patients and stable controls.

Nucleus accumbens right and left are the first biomarkers to become abnormal as seen in Figure 2.6. This was also observed by [86] in one of the subtypes of AD identified in their work. However, the large standard error of the event centers for the events before ABETA suggests that the exact position of those events are unreliable. Experiment 1(b) showed that weak biomarkers (biomarkers excluded in Figure 2.5, but included in Figure 2.6) could lead to greater uncertainty in event centers. This can be explained by the fact that weak biomarkers are the ones where there is a lot of overlap between the Gaussians of pre-event and post-event classes. Small variation in the sampling population during bootstrapping leads to large changes in the parameters estimated in the mixture modeling step of the algorithm. It also showed that majority of the early structural biomarkers are from Temporal lobe, followed by Central structures, Frontal lobe, Parietal lobe and Occipital lobe.

2.7 Conclusion

We proposed a new framework for event-based modeling, called discriminative event-based modeling (DEBM), which includes a new optimization strategy for Gaussian mixture modeling, a new paradigm for inferring the mean ordering, a way for estimating the proximity of events in the order to create a disease progression timeline, and a new way of staging patients that uses these relative proximities of events while placing new subjects on the estimated timeline. The source code for DEBM and FEBM was made publicly available online under the GPL 3.0 license: <https://github.com/88vikram/pyebm/>.

We applied the DEBM framework to a set of 1737 subjects from the baseline ADNI measurement, and also performed an extensive set of simulation experiments verifying the technical validity of DEBM. The experiment on ADNI data illustrated a number of advantages of the new approach. Firstly, we showed that strict quantization of position in ordering of events in the positional variance diagram can sometimes be non-intuitive in terms of inferring actual progression of a disease. Secondly, we showed that the patient staging based on the proposed approach separates CN and AD group of subjects much better than the previous EBM models. Thirdly, we showed that the patient staging can be used to identify individuals at-risk of developing AD as the MCI converters and non-converters were

well-separated. Staging patients based on the estimated disease progression timeline can thus make computer-aided diagnosis and prognosis more explainable. The results of these experiments are encouraging and suggest that DEBM is a promising approach to disease progression modeling.

Supplementary Material

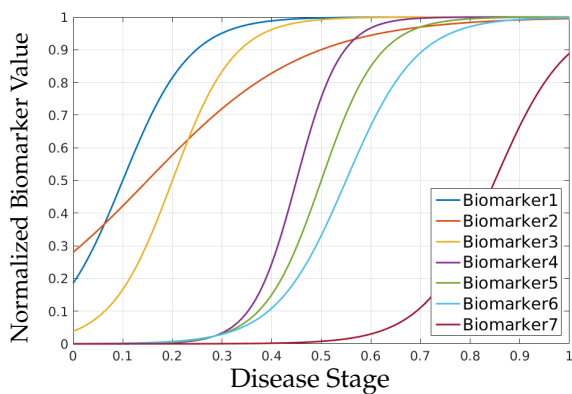


Figure 2.SF1: Biomarker progression rates for different biomarkers used in the second part of Experiment 6.

| Biomarker (unit) | Normal Parameters (Mean \pm Std. Dev.) | Abnormal Parameters (Mean \pm Std. Dev.) | Mixing Parameter (θ^{-E}) |
|--------------------|---|---|------------------------------------|
| ABETA (pg/ml) | 228.47 \pm 32.99 | 136.55 \pm 26.63 | 0.40 |
| P-TAU (log(pg/ml)) | 3.22 \pm 0.43 | 3.85 \pm 0.49 | 0.51 |
| TAU (log(pg/ml)) | 4.07 \pm 0.40 | 4.78 \pm 0.45 | 0.59 |
| ADAS13 | 10.77 \pm 4.21 | 23.28 \pm 8.99 | 0.51 |
| MMSE | 29.10 \pm 0.80 | 25.36 \pm 2.41 | 0.49 |
| Hippocampus (ml) | 3.64 \pm 0.40 | 3.22 \pm 0.52 | 0.58 |
| WholeBrain (ml) | 466.89 \pm 29.32 | 434.07 \pm 49.48 | 0.73 |

Table 2.ST1: Gaussian mixture model parameters for the biomarkers in Experiment 1(a).

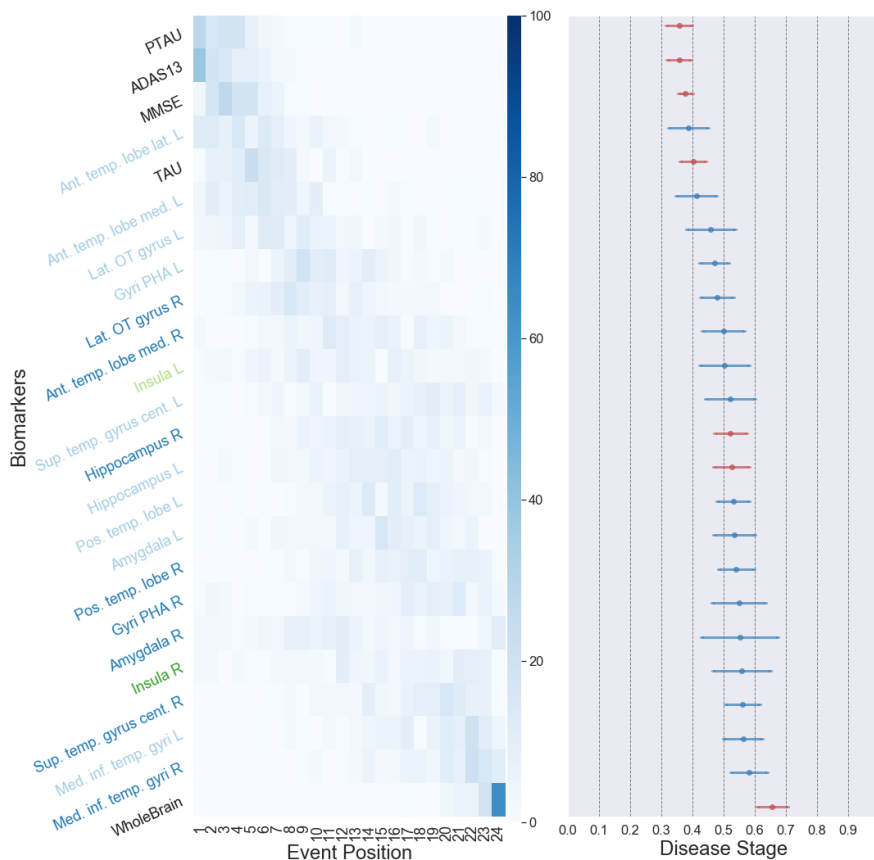


Figure 2.SF2: Experiment 3(b): DEBM_{vv2} with 24 Events, in the homogeneous subset of subjects. The positional variance diagram (left) shows the uncertainty in estimating the central event ordering and the event-center variance diagram (right) shows the standard error of estimated event centers. These were measured by 100 repetitions of bootstrapping. The event centers of the biomarkers used in Figure 9 are marked in red. Table 3 shows the full forms of the abbreviations used in the y-axis labels. Figure 7 maps the colors used for y-axis labels to different lobes in the brain.

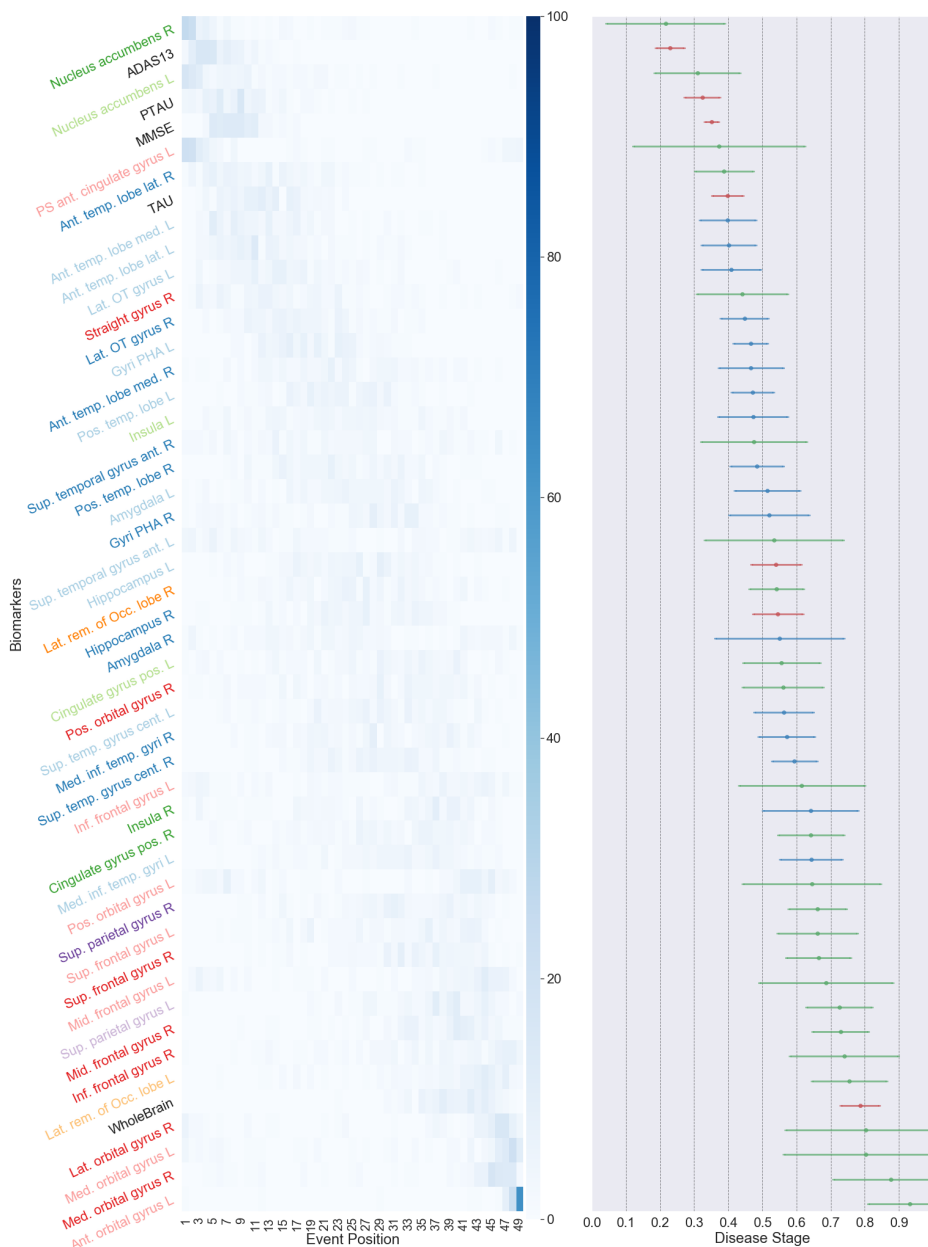


Figure 2.SF3: Experiment 3(b): $\text{DEBM}_{\text{VV}2}$ with 49 Events, in the homogeneous subset of subjects. Positional variance diagram (left) shows the uncertainty in estimating the central event ordering and event center variance diagram (right) shows the standard error of estimated event-centers. These were measured by 100 repetitions of bootstrapping. The event-centers of the biomarkers used in Figure 9 are marked in red, whereas the ones used in Figure 2.SF2 are marked in blue. Table 3 shows the full forms of the abbreviations used in the y-axis labels. Figure 7 maps the colors used for y-axis labels to different lobes in the brain.

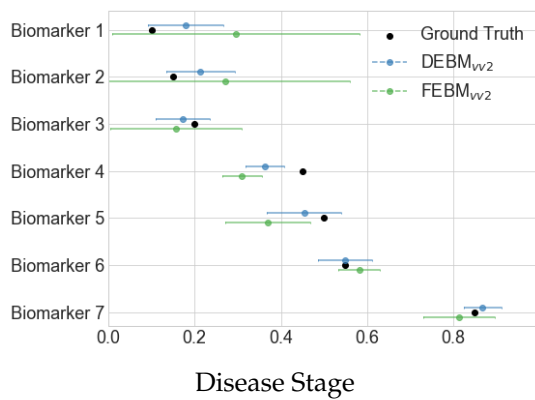


Figure 2.SF4: Experiment 6: The non-uniform μ_{ξ_i} as well as the estimated event-centers by DEBM_{vv2} and FEBM_{vv2} for the case of ρ_i being unequal for different biomarkers.

Chapter 3

Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease

This chapter contains the content of the manuscript 'Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease. Damiano Archetti, Silvia Ingala, **Vikram Venkatraghavan**, Viktor Wottschel, Alexandra L. Young, Maura Bellio, Esther E. Bron, Stefan Klein, Frederik Barkhof, Daniel C. Alexander, Neil P. Oxtoby, Giovanni B. Frisoni, Alberto Redolfi, Alzheimer's Disease Neuroimaging Initiative, & for EuroPOND Consortium. *NeuroImage Clinical*, 24, 2019.'

Vikram Venkatraghavan's contributions in this chapter are: Methodology, software, writing: review & editing

Abstract

Understanding the sequence of biological and clinical events along the course of Alzheimer’s disease provides insights into dementia pathophysiology and can help participant selection in clinical trials. Our objective is to train two data-driven computational models for sequencing these events, the Event Based Model (EBM) and discriminative-EBM (DEBM), on the basis of well-characterized research data, then validate the trained models on subjects from clinical cohorts characterized by less-structured data-acquisition protocols.

Seven independent data cohorts were considered totalling 2389 cognitively normal (CN), 1424 mild cognitive impairment (MCI) and 743 Alzheimer’s disease (AD) patients. The Alzheimer’s Disease Neuroimaging Initiative (ADNI) data set was used as training set for the construction of disease models while a collection of multi-centric data cohorts was used as test set for validation. Cross-sectional information related to clinical, cognitive, imaging and cerebrospinal fluid (CSF) biomarkers was used.

Event sequences obtained with EBM and DEBM showed differences in the ordering of single biomarkers but according to both the first biomarkers to become abnormal were those related to CSF, followed by cognitive scores, while structural imaging showed significant volumetric decreases at later stages of the disease progression. Staging of test set subjects based on sequences obtained with both models showed good linear correlation with the Mini Mental State Examination score ($R^2_{EBM} = 0.866$; $R^2_{DEBM} = 0.906$). In discriminant analyses, significant differences ($p\text{-value} \leq 0.05$) between the staging of subjects from training and test sets were observed in both models. No significant difference between the staging of subjects from the training and test was observed ($p\text{-value} > 0.05$) when considering a subset composed by 562 subjects for which all biomarker families (cognitive, imaging and CSF) are available.

Event sequence obtained with DEBM recapitulates the heuristic models in a data-driven fashion and is clinically plausible. We demonstrated inter-cohort transferability of two disease progression models and their robustness in detecting AD phases. This is an important step towards the adoption of data-driven statistical models into clinical domain.

3.1 Introduction

Alzheimer’s disease (AD) is a complex multifactorial neurodegenerative condition characterized by deposition of abnormal protein-aggregate, synaptic dysfunction, and eventually neuronal loss in the brain [50]. While progression of the disease invariably results in dementia, it has been estimated that clinically-overt manifestations are preceded by a latent phase with no measurable cognitive dysfunction lasting approximately 15-20 years [87]. As AD onset remains insidious in terms of clinical manifestations, biomarkers are the most accurate approach to track disease onset and progression [87].

A variety of biomarkers have been proposed to describe the different phases of the disease, each mirroring different biochemical, functional, or structural changes as the disease develops and progresses. The correct sequence of biomarker transitions to abnormality would allow an appropriate characterization of the different clinical and preclinical disease stages. In addition, this approach could inform the development of individualized treatments in the context of precision medicine or the identification of individuals at-risk of dementia for secondary prevention strategies [88, 89].

While the recently published research criteria [90, 91] for the definition of AD stages outlined robust principles [5, 82, 92], their operationalization in mathematical models and out-of-the-box algorithms has recently begun.

The event-based model (EBM) [53, 61] and the discriminative event-based model (DEBM) [93] are two among an increasing number [80] of probabilistic data-driven methods developed to understand evolution of biomarkers as disease develops and progresses [51, 52, 57, 68]. Their assumption is that the disease is characterized by an irreversible and monotonic change of biomarkers towards abnormality, which might track disease progression. Both algorithms are cross-sectional statistical models that use no strong a priori assumptions regarding the relationship among the different biomarkers or pre-defined cut-offs separating their normal and abnormal values. Both models estimate disease progression as a single average sequence, albeit in slightly different ways: the EBM estimates the maximum-likelihood sequence over all individuals, whereas the DEBM calculates the optimal event sequence as an average of estimations of patient-specific orderings.

Previous works demonstrated the EBM's capability to order biomarkers and stage subjects with a fine-grained ability in classification of Cognitively normal (CN) and AD subjects as well as to predict conversion from Mild Cognitive Impairment (MCI) to AD or from CN to MCI [53, 61].

So far, statistical models have been tested and validated exclusively on a few well-characterized research data sets, such as: Alzheimer's Disease Neuroimaging Initiative (ADNI) [53, 61, 93], Magnetic Resonance in Multiple Sclerosis (MAGNIMS) [94], GENetic Frontotemporal dementia Initiative (GENFI) [86] and TRACK-HD study of Huntington's disease [95], or on synthetic data. This work focusses on transferability of the models to clinical data in AD and provides new evidence that supports widespread clinical adoption of the EBM and DEBM.

Key steps in the validation for the adoption of this kind of models are: (i) ability to build robust disease models on the basis of well-phenotyped research data sets, such as ADNI; (ii) consistency of the disease models on less well-phenotyped clinical data sets in terms of model stability and subjects' staging; (iii) clear end-user interfaces to make model results accessible by clinicians.

In the next sections, we addressed the aforementioned points towards the definition of two valid models for disease progression. Our goal was to assess the transferability of EBM and DEBM's optimal sequence of biomarkers on independent clinical data coming from six different multi-centric initiatives spanning the entire AD spectrum.

3.2 Material and Methods

3.2.1 Participants

A total of 4556 subjects (CN=2389; MCI=1424; AD=743) from different cohorts were selected for this study. The initiatives and projects included in this study are described in Tab. 3.1 and Tab. 3.2. Each cohort had different proportions of subjects in different AD stages depending on the scope of the study. Each study was approved by the local medical ethics committee. Participants for our study were selected using of the following criteria: 1) availability of information on syndromic diagnosis at baseline; 2) availability of T1-weighted Magnetic Resonance Imaging (MRI) scans obtained by either 1.5T or 3T scanners at baseline;

| Data set | Full name | Description | Categories |
|----------|---|---|------------------|
| ADNI-1 | Alzheimer's disease neuroimaging initiative - 1 | The Alzheimer's Disease Neuroimaging Initiative [96] is a longitudinal multicentre study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD). ADNI was originally launched in 2003 as a public-private partnership; its primary goal has been to test whether magnetic resonance imaging (MRI), biological markers, clinical and neuropsychological assessments can be combined to measure the progression of MCI and Alzheimer's disease. The initial five-year study (ADNI-1) was extended by two years in 2009 by a Grand Opportunities grant (ADNI-GO), and in 2011 by a further competitive renewal of the ADNI-1 grant (ADNI-2). Through its 3 phases, it has targeted participants with AD, different stages of MCI, and CN. | CN, SMC, MCI, AD |
| ADNI-GO | Alzheimer's disease neuroimaging initiative – grand opportunities | | SMC, MCI |
| ADNI-2 | Alzheimer's disease neuroimaging initiative - 2 | | CN, SMC, MCI, AD |

Table 3.1: Characteristics of the training dataset

3) absence of any other major neurological, psychiatric or somatic disorders that could cause cognitive impairment at baseline. Subjects were divided in two subsets: training set, used to define the event sequences that serve as disease model, and test set, used for the validation of the disease models.

. The training set was composed of 1488 subjects from the ADNI data set of which 468 were CN, 753 were MCI and 267 were AD. The test set was formed by 3068 subjects from six independent data sets of which 1921 were CN, 671 were MCI and 476 were AD 3.3.

Subjects from ADNI and Amsterdam Dementia Cohort (ADC) with a diagnosis of subjective memory complaints (SMC) were assimilated to CN group, since Mini Mental State Examination (MMSE) score of these individuals was 28.1 ± 1.6 . Significant differences in demographical (age, sex and education) and genetic (carriers of Apolipoprotein E $\epsilon 4$ (APOE4)) information between diagnostic groups were observed for both training and test sets. Differences were observed in the estimated Total Intracranial Volume (eTIV) only in the training set. All demographic and genetic data of training set subjects were significantly different ($p\text{-value} \leq 0.05$) from demographic and genetic data of test subjects in the similar diagnostic group and for the totality of the populations (see Tab. 3.4 for full demographical information).

3.2.2 Biomarkers

When available, multimodal biomarkers collected at baseline tracking different aspects of disease biology were retrieved, i.e. (i) results of neuropsychological tests, (ii) cerebrospinal fluid (CSF) markers and (iii) imaging markers. All the selected subjects had imaging biomarkers, but some missed the results of neuropsychological tests and/or did not undergo lumbar puncture depending on the study cohort; in the latter case staging was performed on the basis of the available markers.

Cognitive biomarkers included MMSE, Alzheimer's Disease Assessment Scale - Cognitive (ADAS-Cog) and Rey's Auditory Verbal Learning Test - Immediate Recall (RAVLT).

| Data set | Full name | Description | Categories |
|-----------|--|--|--------------|
| ADC | Amsterdam dementia cohort | The ADC includes all patients who come to the Alzheimer Center in Amsterdam (since 2004) for diagnostic work-up and consent to give all their data collected for research [97]. The aim is to facilitate research into new and existing biomarkers in the broadest sense, to establish diagnostic, prognostic values and further insight into the pathogenesis of neurodegenerative dementias. The data consist of baseline and annual follow-up assessments. Clinical, neuropsychological, imaging, and biological markers are collected. Since it is conception it has grown into one of the largest clinical data sets in the dementia field. | SMC, MCI, AD |
| ARWiBo | Alzheimer's disease repository without borders | ARWiBo is a cross-sectional data set including data from > 2500 patients enrolled in Brescia (Italy) and nearby areas. The data set contains socio-demographic, clinical, genotype, bio-specimen information, MRI T1-weighted images [98]. | CN MCI AD |
| EDSD | European DTI study on dementia | EDSD [99] is a framework of nine European centres: Amsterdam (Netherlands), Brescia (Italy), Dublin (Ireland), Frankfurt (Germany), Freiburg (Germany), Milano (Italy), Mainz (Germany), Munich (Germany), and Rostock (Germany). It is a cross-sectional multi-centre study characterized by 474 volumetric MRI T1-weighted scans with socio-demographic, clinical, genetic, and biological variables. | CN MCI AD |
| OASIS | Open access series of imaging studies | OASIS [100] consists of (I) a cross-sectional collection of 416 subjects. 100 of the included subjects, over the age of 60, have been clinically diagnosed with very mild to moderate Alzheimer's disease (AD). (II) A longitudinal collection of 150 subjects aged from 60 to 96 years. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. In addition, the data set contains socio-demographic, clinical, genotype information. | CN MCI AD |
| PharmaCog | Prediction of cognitive properties of new drug candidates for neurodegenerative diseases in early clinical development | PharmaCog is an industry-academic European project (IMI) aimed at identifying biomarkers sensitive to symptomatic and disease modifying effects of drugs for Alzheimer's disease [101]. Several clinical sites participated in this study across Italy (Brescia, Verona, Milan, Perugia, and Genoa), Spain (Barcelona), France (Marseille, Lille, and Toulouse), Germany (Leipzig and Essen), Greece (Thessaloniki) and Netherland (Amsterdam). 151 MCI patients have been studied longitudinally collecting multimodal image scans, clinical variables, and bio-specimens. | MCI |
| ViTA | Vienna transdanube aging | ViTA is a population-based cohort-study of all 75-years old inhabitants of a geographically defined area of Vienna [102]. ViTA is composed of 606 subjects followed longitudinally for 4 years. Recruitment took place between May 2000 and October 2002. The primary focus of the ViTA work-group was to establish a prospective age cohort for evaluation of prognostic criteria for the development of AD. | CN MCI AD |

Table 3.2: Characteristics of the test datasets

| Diagnoses and biomarker availability | | | | | | | | |
|--------------------------------------|-------------|------|------|-----|-----------|------|-----|------------------|
| | Data Set | CN | MCI | AD | Sub-Total | MRI | CSF | Cognitive scores |
| Training set | ADNI 1/GO/2 | 468 | 753 | 267 | 1488 | 100% | 72% | 100% |
| | ADC | 125 | 80 | 129 | 334 | 100% | 83% | 99% |
| | ARWiBo | 1399 | 169 | 152 | 1720 | 100% | 3% | 59% |
| Test set | EDSD | 179 | 138 | 151 | 468 | 100% | 19% | 97% |
| | OASIS | 177 | 122 | 42 | 341 | 100% | NA | 100% |
| | PharmaCog | 0 | 147 | 0 | 147 | 100% | 99% | 100% |
| | ViTA | 41 | 15 | 2 | 58 | 100% | NA | 100% |
| | Total | 2389 | 1424 | 743 | 4556 | 100% | 36% | 77% |

Table 3.3: The number of cognitively normal (CN), mild cognitive impairment (MCI), Alzheimer’s disease (AD) and total subjects is reported for each data set. Biomarker availability is expressed as percentage related to the total subjects in each data set. No CSF biomarker is available for OASIS and ViTA data sets.

| Demographics and clinical characteristics | | | | | | |
|---|-------------------------|------------|------------|------------|-----------------------|------------|
| | | CN | MCI | AD | P-value | Total |
| Training set | Age | 73.9 ± 6.7 | 72.5 ± 7.3 | 73.9 ± 7.9 | 3.22×10^{-4} | 73.2 ± 7.0 |
| | Years of education | 16.4 ± 2.7 | 15.9 ± 2.8 | 15.2 ± 2.9 | 1.09×10^{-6} | 15.9 ± 2.8 |
| | eTIV (cm ³) | 1510 ± 180 | 1540 ± 160 | 1530 ± 160 | 4.20×10^{-3} | 1530 ± 160 |
| | MMSE | 29.1 ± 1.2 | 27.6 ± 1.8 | 23.2 ± 2.0 | 2.2×10^{-16} | 27.3 ± 2.6 |
| | Sex (% of females) | 52% | 42% | 48% | 1.43×10^{-3} | 46% |
| | APOE4-carrier | 34% | 49%* | 66% | 2.2×10^{-16} | 49% |
| Test set | Age | 56 ± 17 | 70.6 ± 7.7 | 73.7 ± 8.1 | 2.2×10^{-16} | 62 ± 16 |
| | Years of education | 10.8 ± 4.8 | 9.0 ± 4.5 | 8.7 ± 4.5 | 2.2×10^{-16} | 10.2 ± 4.8 |
| | eTIV (cm ³) | 1450 ± 160 | 1460 ± 170 | 1470 ± 170 | 0.157 | 1460 ± 160 |
| | MMSE | 28.7 ± 1.4 | 26.5 ± 2.4 | 21.0 ± 4.7 | 2.2×10^{-16} | 26.6 ± 3.9 |
| | Sex (% of females) | 61% | 49% | 63% | 1.50×10^{-5} | 58% |
| | APOE4-carrier | 21% | 43% | 49% | 2.2×10^{-16} | 43% |

Table 3.4: Data are expressed as mean values ± standard deviations. Acronyms: eTIV: estimated total intracranial volume; MMSE: Mini Mental State Examination; APOE4: apolipoprotein E ε4; CN: cognitively normal; MCI: mild cognitive impairment; AD: Alzheimer’s disease. P-values were calculated via chi square test for dichotomic variables and via ANOVA for non-dichotomic variables. Values of training set denoted with * are not significantly different from their corresponding values derived from the test subjects (p-value > 0.05).

The CSF concentrations of Amyloid- $\beta_{1,42}$ ($A\beta_{1,42}$) [23, 103, 104], total Tau (t-Tau) and phosphorylated Tau (p-Tau) proteins [23, 103, 104] were collected, and the ratio between the concentrations of $A\beta_{1,42}$ and p-Tau was calculated [104].

The selected imaging biomarkers were: volumetric measures of the hippocampus, entorhinal cortex, fusiform gyrus, middle-temporal gyrus and precuneus, together with whole brain volume and ventricles [105, 106]. Imaging biomarkers were estimated from MRI 3D-T1 sequences analysed with FreeSurfer software v5.3 cross-sectional stream (<http://surfer.nmr.mgh.harvard.edu>) and outputs were visually checked. We assumed a symmetric pattern of atrophy in AD and selected imaging biomarkers were averaged between the left and right hemisphere.

Imaging biomarkers and cognitive scores were available for the totality of subjects from the training set, while CSF biomarkers were available for 72% of these individuals. Imaging biomarkers were available for the totality of test subjects while cognitive scores were available for 84% of test subjects. Within the test set, ADAS-Cog and RAVLT scores were available only for subjects from the PharmaCog data set. CSF biomarkers were available for 18% of test subjects. See Tab. 3.3 for full information on biomarker availability.

CSF biomarkers were obtained with different assays across different cohorts, i.e. Multiplex xMAP Luminex platform with Innogenetic immunoassay kit-based reagents [107] for ADNI subjects and Enzyme Linked Immunosorbent Assay (ELISA) [108] for subjects from all other cohorts, which led to different CSF biomarkers distributions. In order to tackle this issue and to correct for possible acquisition-related differences across datasets, all biomarkers (cognitive scores, CSF, imaging) from subjects from ADC, ARWiBo (Alzheimer disease Repository Without Borders), EDSO (European DTI Study on Dementia), OASIS (Open Access Series of Imaging Studies), PharmaCog and ViTA (Vienna Transdanube Aging) cohorts were rescaled to match the mean and standard deviation of biomarkers distribution of ADNI subjects. In order to ensure Gaussianity, we performed a log-transformation of p-tau and t-tau as their values were non-normally distributed.

All biomarkers from the training and test sets were regressed against age, education and sex and the effects of these factors were corrected to compensate inter cohort demographic variability [109]; imaging biomarkers were additionally regressed and corrected against eTIV [110, 111] to compensate for head size. Correction of biomarkers was performed separately for training set and test set.

The comparison of the selected biomarkers in this study among the three clinical groups and the seven data cohorts considered in this study are shown in Supplementary Figure 3.SF1.

3.2.3 Mathematical modelling

Development of EBM and DEBM was based on the fundamental work of Fonteijn *et al.* [53]. According to these approaches, each biomarker is considered as either normal or abnormal and its probabilistic transition from the normal to the abnormal state is defined as event. The aim is to define in a data-driven manner the sequence of events that describe the most probable ordered cascade that characterizes the transition of a subject from the healthy state to the full-blown disease spectrum [61]. For this work, we employed python module pyebm (<https://github.com/EuroPOND/pyebm>), where both algorithms

are implemented.

In the EBM [53, 61] possible event sequences are sampled via a Markov Chain Monte Carlo (MCMC) process aimed at finding the sequence that best fits the biomarker observations from all subjects. At each Monte Carlo step a new sequence is sampled as a random swap between two biomarkers of the current benchmark sequence. If the new sequence is a better fit than the benchmark sequence, which is determined mathematically by the likelihood, then the new sequence is considered as the benchmark sequence for the following MCMC step.

The probability of an event for each biomarker is determined by a Gaussian mixture model (GMM) where the normal and abnormal components are modelled by Gaussian distributions. In EBM [61], distributions of normal and abnormal biomarkers are initialized as the distributions of biomarkers from the CN and AD subjects, respectively. The mixture model distribution for each biomarker is then found as the sum, weighted on the mixing parameters, of the two aforementioned distributions that best fits to biomarker values from all subjects. Optimization of the GMM function is performed along the Gaussian parameters and the mixing parameters and in order to avoid the possibility that biomarkers will not show a clear bimodal distribution, the standard deviations for normal and abnormal components in the GMM are constrained to be no greater than the standard deviations of CN and AD subjects, respectively.

The approach of DEBM model [70, 93] for the calculation of the central ordering, on the other hand, is a two-step process where first (i) a specific ordering is calculated for each subject by sorting the posterior probability that each biomarker has become abnormal and then (ii) the central ordering is calculated as the event sequence that minimizes the sum of probabilistic Kendall's tau distances between itself and all the subject-wise orderings. As the posterior probability is influenced by the physiological variability of biomarkers, DEBM assumes that single subject orderings are noisy estimates of the central ordering [93].

The original formulation of DEBM [93] also contains a specific mixture model, for which an initial estimate of the distributions of non-diseased and diseased subjects for each biomarker is performed using values from subjects at the opposite ends of the disease spectrum, as defined by a Bayesian classifier which is trained to remove outliers and wrongly labelled data. This allows efficient separation of the two Gaussian distributions of normal and abnormal values for each biomarker. The biased distributions are then refined including data from all subjects via a GMM that has constraints based on the aforementioned relationships between the expected and the biased distributions. The same objective GMM function as for EBM is optimized alternatively along the Gaussian parameters and the mixing parameters until the latter converge.

Optimal sequences were calculated as averages of orderings obtained from 50 bootstrapped iterations for both EBM and DEBM. Furthermore, in EBM the number of MCMC steps was set to 50,000 to ensure convergence of the likelihood. In practice convergence was typically observed before the 15,000-th MCMC step. See Supplementary Material SS1 for detailed mathematical modelling.

3.2.4 Model validation & Statistical Analysis

Validation of the models is performed by staging subjects from the training and test sets on the basis of the event sequences built on the basis of biomarkers from subjects from the training set. Specific methods

for staging subjects are available in the original works for both the EBM [61] and DEBM [93]. For the sake of simplicity, and in order to have a common staging system for both models, the method from [61] was employed in this work. This method assigns each subject a position of the central event sequence, resulting in a number of stages that is equal to the number of biomarkers considered for the sequence plus one, as it is necessary to add stage 0 where no biomarker is abnormal. The stage of each subject is calculated as the i -th step of the event sequence that maximizes the probability that all events up to have already occurred and events from $i+1$ to the end of the sequence are yet to occur. In case of missing biomarkers, the probability of the biomarkers to be abnormal was set to 0.5 [61]. Assuming that clinical diagnoses of all subjects are made through a biomarker-based assessment, it is expected that each subject, either from the training or test set, is staged at the earlier positions of the event sequences if CN and at the later positions if AD.

Measures of area under curve (AUC), sensitivity, specificity and balanced accuracy at optimal threshold were calculated for all pairwise comparisons among clinical groups, i.e. (i) AD vs. CN, (ii) AD vs. MCI, and (iii) MCI vs. CN. In order to assess significant differences between receiver operating characteristic (ROC) curves, the DeLong test [112] was performed.

To assess the validity of the EBM and DEBM central orderings we explored the linear correlation between subjects' model stages and MMSE scores. The MMSE is the most widely used screening tool to assess cognitive functions in both routine clinical practice and research settings and its score correlates with the different phases of AD progression [113]. In order to avoid circularity MMSE scores were excluded from the initial calculation of the event sequences. Moreover, in order to mitigate the ceiling effect typical of MMSE [114], the lower limit for the linear regression analysis was set as the model stage that provides the optimal threshold for separating CN and MCI subjects. To explore how much the missing biomarkers of test subjects (Tab. 3.3) affected the classification performances in both models, staging was also performed for a special subset of test subjects having at least one CSF measurement, MMSE score and imaging biomarkers. These restriction criteria reduced the original test subjects from 3068 to 562 (104 CN, 331 MCI, 127 AD) and the number of events considered in our original simulation from 13 to 12 as ADAS-Cog and RAVLT were excluded since they were available only for the PharmaCog data set, while MMSE was included.

Statistical analysis was performed with R version 3.5.1.

3.3 Results

3.3.1 Events ordering

Central event sequences and their variances were generated from biomarkers of training subjects for both EBM and DEBM and were plotted as positional variance diagrams (Fig. 3.1).

The event sequence obtained with the DEBM algorithm showed that amyloid related biomarkers became abnormal first. The abnormalities of $A\beta_{1,42}$ protein and $A\beta_{1,42}/p$ -Tau ratio are at the very first positions followed by cognitive scores, Tau protein-related biomarkers, and finally imaging markers of AD-relevant brain regions. Averaged volumes between left and right hemisphere of hippocampus and precuneus are respectively the first and the last brain areas to become abnormal while the medial temporal lobe is in between. The enlargement of the ventricles and the atrophy of the whole brain were in the last two

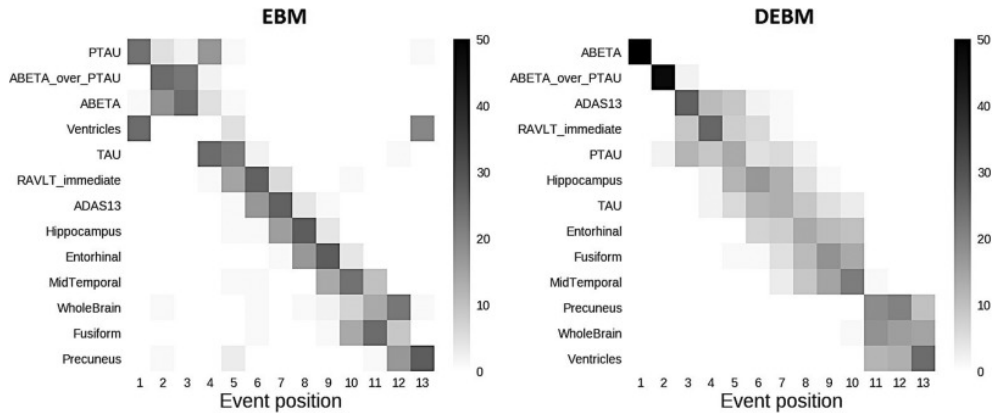


Figure 3.1: Positional variance diagrams of event orderings obtained with EBM and DEBM. Both diagrams show the number of times each biomarker occurred in a specific position from a batch of 50 independent bootstrapped sequences generated using biomarkers of training subjects with EBM (left) and DEBM (right) methods

positions.

In EBM, CSF biomarkers are the first to show abnormality, although with a different pattern with respect to DEBM. Tau related biomarkers became abnormal earlier and often before amyloid-related biomarkers. The sequence obtained with EBM followed a similar ordering for the cognitive scores although the specific order of RAVLT and ADAS scores is swapped. The enlargement of the ventricles is placed at the fourth position of the ordering although the positional variance showed that this event has nonzero probability of occurring in the first or last position of the sequence. Volumetric measures of the grey matter of the fusiform gyrus and precuneus are placed at the very last positions of the EBM benchmark sequence. Both EBM and DEBM showed good positional stability (see Fig. 3.1), and in the case of DEBM no event occurs far from the diagonal.

3.3.2 Staging of individuals across the AD spectrum

Subjects from both training and test set were staged on the basis of the event sequences derived from the training set. For the training set, in both EBM and DEBM cases, more than 60% of CN subjects were staged at position 0 where no abnormalities have occurred yet (Fig. 3.2 (a) & Fig. 3.2 (b)). Similarly, the majority of AD were staged at positions 12-13 (of 13 total) of both sequences. Most of the remaining CN subjects were spread across stages 1-6 in EBM and 1-4 in DEBM. The majority of the remaining AD individuals were staged across stages 7-11 for EBM and stages 5-12 for DEBM.

For the test set, staging of subjects obtained with EBM and DEBM is shown in panels (c) and (d) of Fig. 3.2 respectively. In this case more than 70% of AD subjects was staged at positions 12-13 and more than 60% of CN subjects were staged at position 0, but the strong separation between CN and AD observed in the training set was not reproducible in the test set for 30% of CN subjects were staged at positions 6-13. These test CN subjects belonged to two different phenotypic classes:

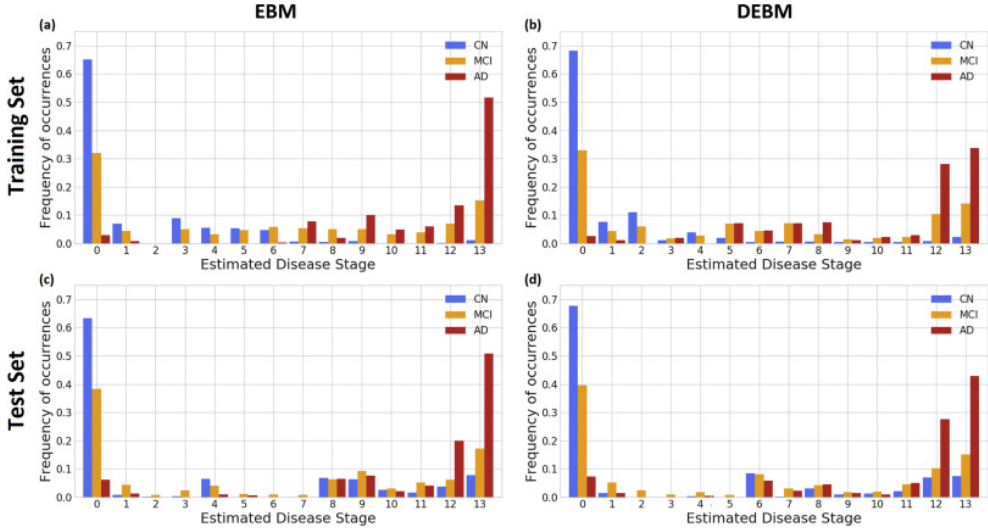


Figure 3.2: Subject staging based on the sequences obtained with EBM and DEBM methods. Staging of subjects from all diagnostic categories (Cognitively normal (CN) in blue, mild cognitive impairment (MCI) in orange, Alzheimer's disease (AD) in red) are shown for (a) training subjects on EBM sequence, (b) training subjects on DEBM sequence, (c) test subjects on EBM sequence and (d) test subjects on DEBM sequence. Histograms are normalized for each diagnostic category. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(1) subjects whose eTIV was very large or very small compared to the eTIV of the CN population. Indeed, the eTIV of these subjects showed a bimodal distribution with peaks at ± 1.1 standard deviations apart from the average of the test CN population;

(2) subjects aged 76.2 ± 8.7 on average, whose MMSE score was on average 29.11, but whose hippocampal normalized volume was significantly smaller compared to the hippocampal normalized volume for the test CN subjects ($(2.1 \pm 0.4) \times 10^{-3}$ vs. $(2.7 \pm 0.4) \times 10^{-3}$).

In each case, the distribution of MCI stages overlapped with the distribution of stages for CN and AD, but a considerable amount, always between 30% and 40%, was staged at position 0 in both EBM and DEBM models (Fig.2). MCI subjects staged at position 0 had an average MMSE score of 28.2 ± 2.1 for training set and 27.0 ± 2.1 for test set. Staging of the subjects from each data set on the basis of EBM and DEBM sequences shows a good separation between CN and AD subjects in each case, and generally few subjects are staged at positions 1-7 for EBM and 1-5 for DEBM as these stages correspond to CSF and cognitive biomarkers (see Supplementary Figure 3.SF2). Linear regression of DEBM stage vs EBM stage resulted in slopes < 1 for both the training and test set, meaning that on average EBM stage is always greater than DEBM stage (see Supplementary Figure 3.SF3).

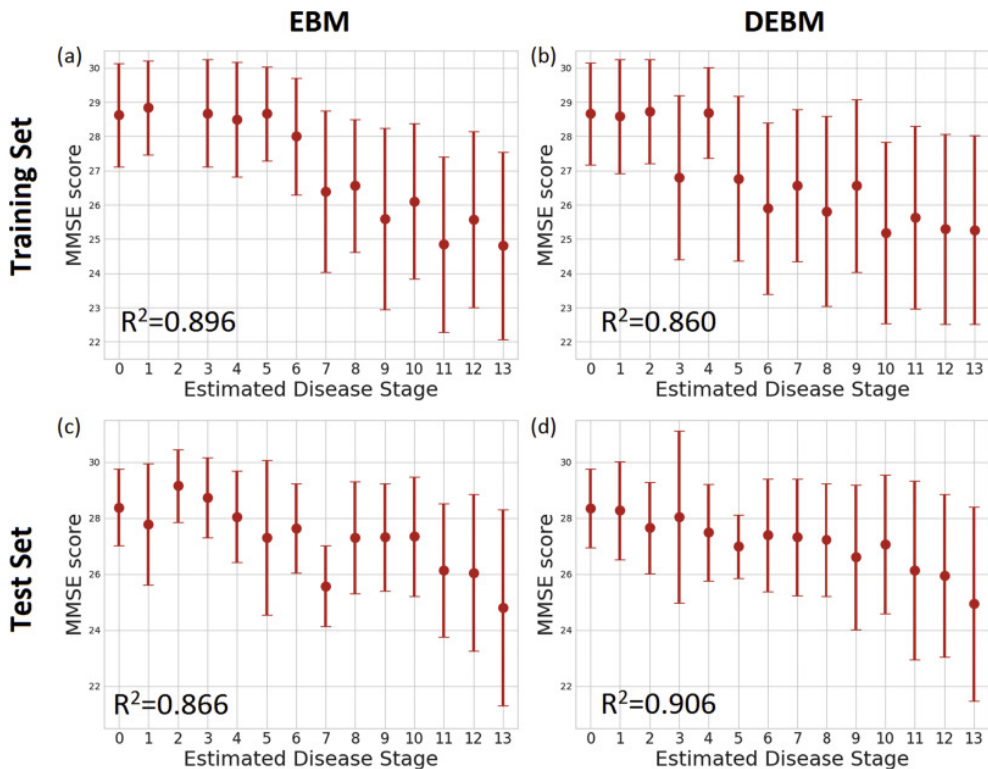


Figure 3.3: Correlation between MMSE score and subjects staging for (a) training set subjects on EBM sequence, (b) training set subjects on DEBM sequence, (c) test set subjects on EBM sequence, (d) test set subjects on DEBM sequence. Average and standard deviation of MMSE score of training and test subjects staged on the basis of EBM and DEBM sequences are shown. Coefficients of determination (R^2) of the linear regression of MMSE score vs disease stage are reported.

3.3.3 Staging vs MMSE correlation

Average and standard deviation of the MMSE scores of the training and test sets at each stage is shown in Fig. 3.3. The plot showed decreasing MMSE scores in the latter stages in both EBM and DEBM.

Linear regression of the MMSE scores of all subjects excluding the initial ceiling effect showed correlation between the decrease in MMSE score and patient staging of training subjects for both EBM (0.896) and DEBM (0.860). The limit of the initial ceiling was set as the model stage threshold that optimally separates CN and MCI subjects, that is stage 6 for EBM and stage 5 for DEBM in the case of the training set. Good linear correlation between MMSE scores and subject staging was observed for individuals from the test set (0.866 for EBM and 0.906 for DEBM), although the ceiling effect thresholds were different from the thresholds of the training set (stage 1 for both EBM and DEBM).

| Prediction of clinical diagnosis | | | | | | | | | | | |
|----------------------------------|-------|------|------|--------|-------|-------|------|------|--------|-------|-----------------------|
| | EBM | | | | | DEBM | | | | | p |
| | k_T | Sens | Spec | BalAcc | AUC | k_T | Sens | Spec | BalAcc | AUC | |
| Training | | | | | | | | | | | |
| AD vs CN | 7 | 0.97 | 0.96 | 0.96 | 0.97* | 5 | 0.92 | 0.94 | 0.93 | 0.95* | 1.88×10^{-3} |
| AD vs MCI | 9 | 0.59 | 0.96 | 0.77 | 0.81 | 5 | 0.48 | 0.94 | 0.71 | 0.76 | 5.30×10^{-5} |
| MCI vs CN | 6 | 0.88 | 0.52 | 0.70 | 0.73* | 5 | 0.92 | 0.52 | 0.72 | 0.73* | 0.537 |
| Test | | | | | | | | | | | |
| AD vs CN | 5 | 0.71 | 0.91 | 0.81 | 0.87 | 7 | 0.78 | 0.85 | 0.81 | 0.86 | 3.99×10^{-2} |
| AD vs MCI | 12 | 0.77 | 0.71 | 0.74 | 0.78 | 11 | 0.70 | 0.75 | 0.73 | 0.77 | 0.393 |
| MCI vs CN | 1 | 0.63 | 0.62 | 0.62 | 0.63 | 1 | 0.68 | 0.60 | 0.64 | 0.64 | 0.676 |

Table 3.5: Measurements of area under curve (AUC), sensitivity (Sens), specificity (Spec) and balanced accuracy (BalAcc) at a specific threshold (k_T) for the staging obtained with EBM and DEBM methods on training and test data sets not containing missing values. Thresholds are chosen to maximize the balanced accuracy in each classification task. P-values of DeLong test performed to compare AUCs of EBM and DEBM methods are reported in the last column. AUCs of training set denoted with * are significantly different from their corresponding values derived from the test subjects (p-value of DeLong test ≤ 0.05)

3.3.4 Prediction of clinical diagnosis

Clinical diagnosis classification of each individual from both training and test data sets was computed. All the possible combinations were assessed, i.e. AD vs. CN, AD vs. MCI and MCI vs. CN. The balanced accuracy and AUC values of the classification obtained on both training and test sets were comparable to other state-of-the-art classification approaches [61]. In the case of AD vs. CN, balanced accuracy and AUC of the ROC curve, alongside measures of sensitivity and specificity, are greater than 0.93 in the training set and greater than 0.81 for test set for both models (see Tab. 3.5). The comparison of the AUC showed significant differences (p-value ≤ 0.05) between EBM and DEBM in both training and test sets. For AD vs. MCI subjects, balanced accuracy and AUC in both training and test sets were always greater than 0.71. No significant differences were registered between the AUC of EBM and DEBM. In the case of MCI vs. CN subjects, balanced accuracy and AUC values were between 0.62 and 0.73 without significant differences between EBM and DEBM. In both models, a significant difference (p-value ≤ 0.05) between training and test sets was observed in two of the three classification tasks: (i) AD vs. CN; (ii) MCI vs CN. The maximum balanced accuracy threshold (k_T) used in the classification increases across the disease spectrum in both models with the exception of DEBM on ADNI subjects where the threshold is constant for all classifications. This is compatible with the idea that EBM and DEBM produce event sequences that track disease progression.

To fully explore the capabilities of the two models and to perform a fair head to head comparison we run similar analyses in the training and test sets considering all the 14 biomarkers (see Supplementary Figures 3.SF4, 3.SF5). On average, the general performance in discriminating subjects from the test set improved by 2 and 4 percentage points respectively for DEBM and EBM (see Supplementary Table 3.ST1). This improvement is achieved by the inclusion of the MMSE score, which is available for a large portion of test subjects.

Results of the case where all test subjects do not have missing biomarkers showed improvement in the performances for all the computed metrics. In the test set, on average, DEBM showed an increase of

| Prediction of clinical diagnosis with no missing biomarkers | | | | | | | | | | | |
|---|-------|------|------|--------|------|-------|------|------|--------|------|-----------------------|
| | EBM | | | | | DEBM | | | | | p |
| | k_T | Sens | Spec | BalAcc | AUC | k_T | Sens | Spec | BalAcc | AUC | |
| Training | | | | | | | | | | | |
| AD vs CN | 8 | 0.98 | 0.95 | 0.97 | 0.97 | 3 | 0.86 | 0.99 | 0.92 | 0.95 | 3.10×10^{-2} |
| AD vs MCI | 8 | 0.70 | 0.95 | 0.83 | 0.83 | 7 | 0.66 | 0.76 | 0.71 | 0.76 | 0.104 |
| MCI vs CN | 5 | 0.89 | 0.51 | 0.70 | 0.72 | 3 | 0.86 | 0.58 | 0.72 | 0.73 | 1.99×10^{-8} |
| Test | | | | | | | | | | | |
| AD vs CN | 4 | 0.88 | 0.94 | 0.91 | 0.95 | 3 | 0.91 | 0.91 | 0.91 | 0.94 | 0.332 |
| AD vs MCI | 4 | 0.57 | 0.94 | 0.76 | 0.80 | 5 | 0.63 | 0.87 | 0.75 | 0.79 | 1.65×10^{-2} |
| MCI vs CN | 4 | 0.88 | 0.43 | 0.66 | 0.66 | 3 | 0.91 | 0.52 | 0.71 | 0.70 | 0.296 |

Table 3.6: Measurements of area under curve (AUC), sensitivity (Sens), specificity (Spec) and balanced accuracy (BalAcc) at a specific threshold (k_T) for the staging obtained with EBM and DEBM methods on training and test data sets not containing missing values. P-values of DeLong test performed to compare AUCs of EBM and DEBM methods are reported in the last column. In DEBM and EBM AUCs of the training set were not significantly different to their corresponding AUCs in the test set (p-values of DeLong test always > 0.05).

4.3% in balanced accuracy and an increase of 3.0% in AUC compared with the metrics obtained from the complete 13 biomarker sequences. Similarly, EBM showed an increase of 7.2% in balanced accuracy and an increase of 5.5% in AUC. Generally, no statistically significant differences between staging of training and test subjects were observed (p-value > 0.05) for all groups in both models. Detailed results are reported in Tab. 3.6.

3.3.5 Sequence consistency

In order to ensure consistency of the benchmark sequence generated from the training set, a disease model was also built on the basis of the test set (i.e.: ADC, ARWiBo, EDS, OASIS, PharmaCog, ViTA) using both EBM and DEBM. ADAS-Cog and RAVLT cognitive scores were not included since these specific tests were available only for MCI subjects from the PharmaCog data set. MMSE was included so that all biomarker families (cognitive, CSF and imaging) were represented.

In both sequences obtained with the EBM, CSF biomarkers occupy the first positions of the sequences (Fig. 3.4(a)) but the second halves of the sequences differ considerably, especially in the position of ventricles and hippocampus. In total, 23 swaps between adjacent biomarkers are needed in order to turn the sequence obtained from the test set into the sequence obtained from the training set.

In DEBM, the event sequences obtained from training and test sets are similar. Only 11 swaps between adjacent events are needed to turn the test set sequence into the benchmarked training set sequence (Fig. 3.4(b)). With the exception of t-Tau and p-Tau both sequences obtained with DEBM can be divided in four partial rankings that contain the same biomarkers: $A\beta_{1,42}$ /p-Tau ratio, $A\beta_{1,42}$ and MMSE in the first partial ranking, hippocampus and entorhinal cortex in the second, middle temporal gyrus, fusiform gyrus and precuneus in the third and whole brain and ventricles in the last partial ranking.

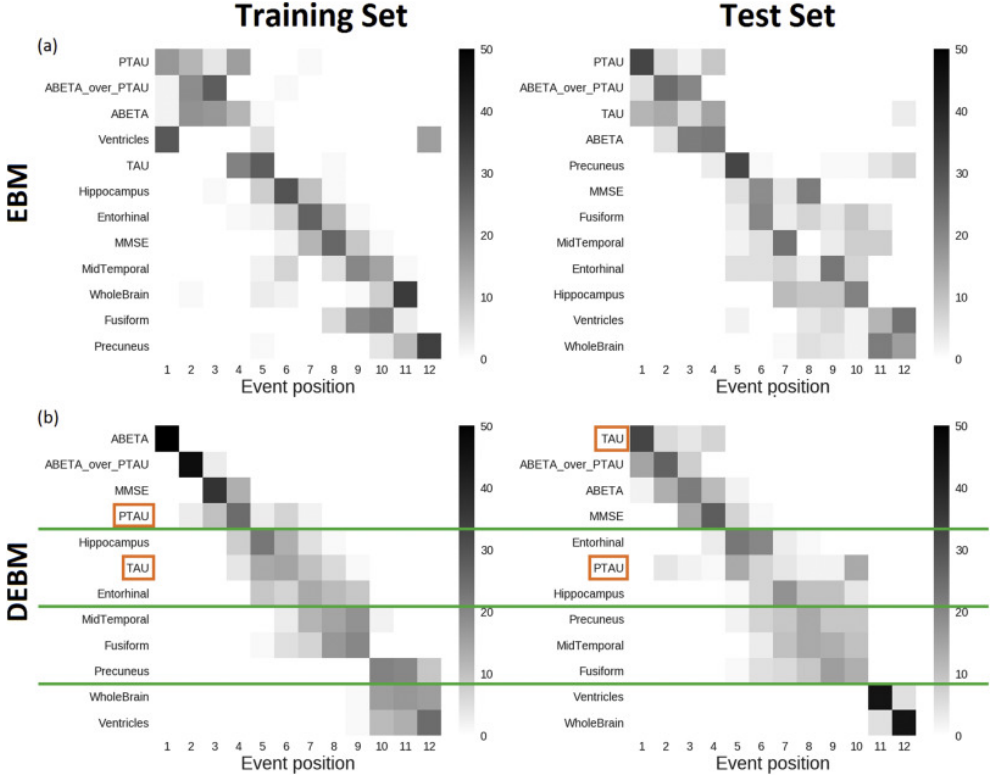


Figure 3.4: Positional variance diagrams of event sequences computed from training set (left) and test set (right) using EBM (a) and DEBM (b) algorithms. In the case of DEBM green lines divide the sequences into homogeneous blocks between the training and test sets. Orange boxes represent biomarker exceptions not conserved in the same block comparing the training vs. test positional variance diagrams. Clear event blocks cannot be identified for EBM sequences.

3.4 Discussion

To our knowledge, this is the first translational study showing viability of the EBM and DEBM, trained on research data, in a clinical setting. This is also the first cross-cohort assessment of the models' validity on cross-sectional multimodal biomarkers. Previous literature focused only on well characterized research datasets and synthetic data [58, 60, 61, 67, 70, 93, 115] but this kind of approach does not take into consideration the aspects of real clinical data. We investigated and compared the performance of EBM and DEBM when applied to the same training and test data sets which included subjects across the entire disease spectrum, accounting for missing data.

EBM and DEBM rely on different estimates of the Gaussian mixture models and in the definition of the optimal sequence of biomarkers. As highlighted in literature [93], the optimization technique adopted in DEBM, for which Gaussian parameters and mixing parameters are optimized alternatively, prevents the abrupt change of the mixing parameter for small changes in the Gaussian parameters that was observed in EBM.

We observed differences between EBM and DEBM optimal event sequences. The DEBM sequence is closer to Jack’s model [105] and also mirrors stages V and VI of cortical degeneration due to neurofibrillary tangles deposition as described in Braak’s Model [116]. The DEBM sequence starts with $A\beta_{1,42}$ and $A\beta_{1,42}/p\text{-Tau}$ ratio, while the EBM sequences suggests p-Tau as the first biomarker to become abnormal. Although in literature it is not completely understood which is triggering the other (if at all), much evidence suggests $A\beta_{1,42}$ deposition to be upstream of Tau deposition. The deposition of amyloid plaques presumably triggers the conversion of Tau protein to toxic state, while less evidence suggests that toxic Tau can enhance $A\beta_{1,42}$ toxicity via a feedback loop. Soluble toxic aggregates of $A\beta_{1,42}$ and p-Tau can self-propagate and spread throughout the entire brain, perhaps enhancing other destructive biochemical pathways [117] and triggering the abnormality cascade of the other biomarkers. It is important to consider, however, that the transition to abnormality of a biomarker may not correspond to its pathological change, since no a priori thresholds are set.

Coherently with Iturria-Medina’s model [60], where spatiotemporal abnormalities of multiple biomarkers are explored via a multi-factorial data-driven analysis, both EBM and DEBM orderings showed a drop in the performance of cognitive test scores after events related to CSF biomarkers. In particular, EBM ordering of cognitive results seems slightly more plausible, ordering the RAVLT before ADAS13, as RAVLT has been reported to be more sensitive to detect abnormal changes in pre-dementia condition [118] while ADAS is more specific to detect moderate AD conditions [119]. According to both methods, cognitive tests were positioned before group-level neurodegeneration events in the benchmark sequences. This fact might be in contrast with literature [5, 120] for which memory impairment occurs after volumetric decrease of brain regions. This difference can be explained by the fact that population-level volume changes may affect the event sequence [61]. The earlier position of cognitive scores with respect to imaging biomarkers could be explained partially by the different GMMs used in the two algorithms and partially because of specific inclusion criteria for the ADNI training subjects. In ADNI, no subjects with severe cognitive impairments were included since one of the inclusion criteria was to have MMSE score at least equal to 18. This may affect the position in which cognitive test scores were considered abnormal because the threshold that separates normal from abnormal values might be overestimated by the models, considering that no a priori assumptions are made in EBM and DEBM.

As far as the MRI biomarkers are concerned, DEBM showed an expected pattern of grey matter atrophy with AD progression. Abnormalities were ordered throughout the temporal lobes as follows: hippocampus, entorhinal cortex, fusiform and mid temporal regions. Precuneus was affected subsequently, in agreement with model of cortical atrophy progression proposed by ten Kate *et al.* [121], where atrophy of parietal regions is associated with progression from MCI to dementia. The DEBM sequence presented the whole brain and subcortical abnormalities as end-sequence events. EBM did not capture the expected atrophic evolution of the grey matter and the main anomaly was represented by ventricles. Their abnormality was reported in the fourth position of the optimal sequence and their variability is spanning from the first to the last position. Two different local likelihood maxima due to different subtypes of AD [86] in the EBM sequence space could be one possible reason. Also, this issue is not observable in DEBM, where normally the variance of an event is distributed continuously around its specific position, that means around the positional variance diagram bisector. The difference between the two models can be attributed to the smoothing effect intrinsic to the DEBM algorithm and, as highlighted in [93], to the specific mixture model used in EBM. The sequences generated by EBM and DEBM models, however, represent a general event ordering for the progression of the disease and individual trajectories may show variability with respect to the optimal sequences.

We demonstrated, using data from ADNI and 6 other independent clinical cohorts, the performances of EBM and DEBM across the entire Alzheimer's time course. Staging of subjects in both the training and test sets showed separation between AD and CN in the two methods. This meant that the algorithms were effective at distinguishing subjects having only a few abnormal biomarkers from those having only a few normal biomarkers. As expected, the majority of CN subjects from the training set were staged at position 0, where no abnormality manifested yet, and a large number of AD subjects was at end-sequence stages 11-13. Staging of the test subjects followed the same general trend as ADNI, although subjects with a lack of CSF values or cognitive assessments and with normal imaging biomarker values were staged in proximity of non-symptomatic stage 0. The large number of CN subjects in the test sets that were staged in the last positions for both models, can be partly explained considering that a significant portion of these individuals are CN elderlies with volumetric anomalies and no other biomarker available, thus contributing to subjects' misclassification although MMSE score showed no abnormalities. Another portion of misclassified CN subjects is formed by individuals with abnormal imaging biomarkers but here the misclassification is due to the linear regression correction since the average eTIV of test subjects is significantly lower than the average eTIV of training subjects, thus, the imaging biomarkers of test subjects are artificially considered as atrophic with respect to the imaging biomarkers from the training set subjects.

Some concerns may arise from the large number of MCI subjects staged at stage 0. The CSF and cognitive scores for the majority of these individuals were close but not yet over the probabilistic threshold values, therefore they were still in the normal ranges, and the models considered those subjects as normal. Despite this, staging evidences give comparable results to state-of-the-art classification techniques for prediction of conversion from MCI to dementia [81, 122].

EBM and DEBM showed good linear correlation with MMSE scores, fairly consistent with the clinical and regional biomarkers, thus producing an indirect validation of models with respect to the disease evolution. Both methods, after an initial plateau due to the ceiling effect typical for MMSE test [114], showed an expected linear decline [123]. Although it was a rather trivial approach, we tried to validate the EBM and DEBM event sequences even in absence of a validated pathological gold-standard across the data cohorts.

When all test subjects are considered, we detected a significant drop of performance in classifying AD vs CN as well as in MCI vs CN subjects from ADNI to the test cohorts. This is probably due to missing data (CSF biomarkers and cognitive scores), which is known to increase uncertainty in subject staging [61]. Indeed, when considering a reduced set of test subjects for which all biomarkers were available, the performances became much closer to those obtained from the training set and no more significant differences between training and test data sets were observable for both EBM and DEBM (p -values ≥ 0.05). This reinforces the importance to collect an adequate set of biomarkers for an accurate staging of single subjects into the correct diagnostic class.

As far as the test set is concerned, the classification of AD vs CN subjects was significantly better in EBM than in DEBM (p -values ≤ 0.05). In classifying AD vs MCI, EBM was slightly better with higher sensitivity, balanced accuracy and AUC. In MCI vs CN, DEBM reached higher sensitivity and balanced accuracy while EBM reached higher specificity. This evidence might represent specific hints to guide the usage of EBM and DEBM for physicians according to the initial diagnostic hypothesis they want to test in their clinical practice.

An interesting consideration for future works is the possibility to use such methods to follow MCI in specific sub-classes, namely: amnesic MCI, non-amnesic MCI and MCI due to AD. Additional studies with extended age range of subject, larger and additional groups and additional biomarkers such as other brain regions will be helpful to achieve a more accurate description of AD via event-based models. Clinically relevant information related to patients' staging, together with the models' robustness as well as progressive tracking capabilities along the CN-to-AD course, might be implemented into a clinical decision support tool, to aid diagnosis and prognostic assessment of AD at early stages.

Additional efforts will be needed to understand the capabilities of staging subjects during clinical routine by means of EBM and DEBM in: (I) reducing the number of patients needed for future clinical trials, (II) monitoring the efficacy of disease modifying drugs, (III) personalized medicine.

So far, EBM and DEBM have been validated against well-characterized research datasets, synthetic data and, in the present study, multicentric clinical cohorts, but none of them has been yet compared against different stages of the AD pathology. In the next future, we would have to focus on further validation of both models against databases of population of normal and abnormal post-mortem studies on subjects assessed with as many biomarkers as possible, such as those collected in the Religious Orders Study [124], Rush Memory and Aging Project [125], the Adult Changes in Thought study [126], and the National Alzheimer's Coordinating Center data set [127].

Some limitations of the current results should be considered in future validations of event-based models. First, the tools here described need to be further compared with other complementary techniques based on longitudinal data sets, such as: temporal continuous models and spatiotemporal models – see [80] for a recent review of the field. Second, as clinicians are the potential beneficiaries of the tools based on such models, independent evaluators should rate the diagnostic added value and accuracy of EBM and DEBM. Third, the greatest limitations in the methods applied is the assumption of a common or average disease trajectory across individuals, while AD is highly heterogeneous and clearly violates this assumption. In this perspective single subject orderings already available in DEBM, and data-driven subtype progression patterns estimated using SuStaIn (Subtype and Stage Inference) [86] could play a central role in the description of AD progression at the level of the single subject. Finally, computational time is worth considering: the extensive use of EBM or DEBM to analyse large volumes of data that must be pre-processed and that require large computational resources, such as: HPC, Grid, or Cloud [128, 129, 130], indeed the models can be trained a priori and then they should be used in the clinical practice only to evaluate new subjects on the basis of the preferred model within an acceptable time frame.

The state of the art of these data driven models is represented by research tools (<https://github.com/EuroPOND>), that should be implemented in more user-friendly interfaces compatible with the clinical routine. Efforts towards the opportunities for clinical adoption and perceived importance of such a tool in clinical setting has started to appear (<https://icometrix.com>) (see Supplementary Figure 3.SF6).

3.5 Conclusions

We have performed an inter-cohort model transferability study and model performance comparison via external validation approach for event-based models. In the field of healthcare, the importance of data driven models will grow in the coming years, and the results presented here represent the first

viability and generalizability proof of principle to train such models on research data and apply them clinically: on cross-sectional, less-well-characterized cohorts. We trained data-driven disease progression models with the ADNI data set and compared patients' ordering, staging and performance through ADC, ARWiBo, EDSD, OASIS, PharmaCog and ViTA data sets. Overall, we tested both models on 4556 subjects and 14 multimodal biomarkers. Both EBM and DEBM demonstrated similar and good classification performances especially when all biomarkers were available for test subjects. Orderings obtained from both models agreed with previous heuristic models. The event sequence generated through DEBM returned a more reasonable description of the course of AD, while EBM showed better classification performances, which are important considerations for future applications.

Supplementary Material

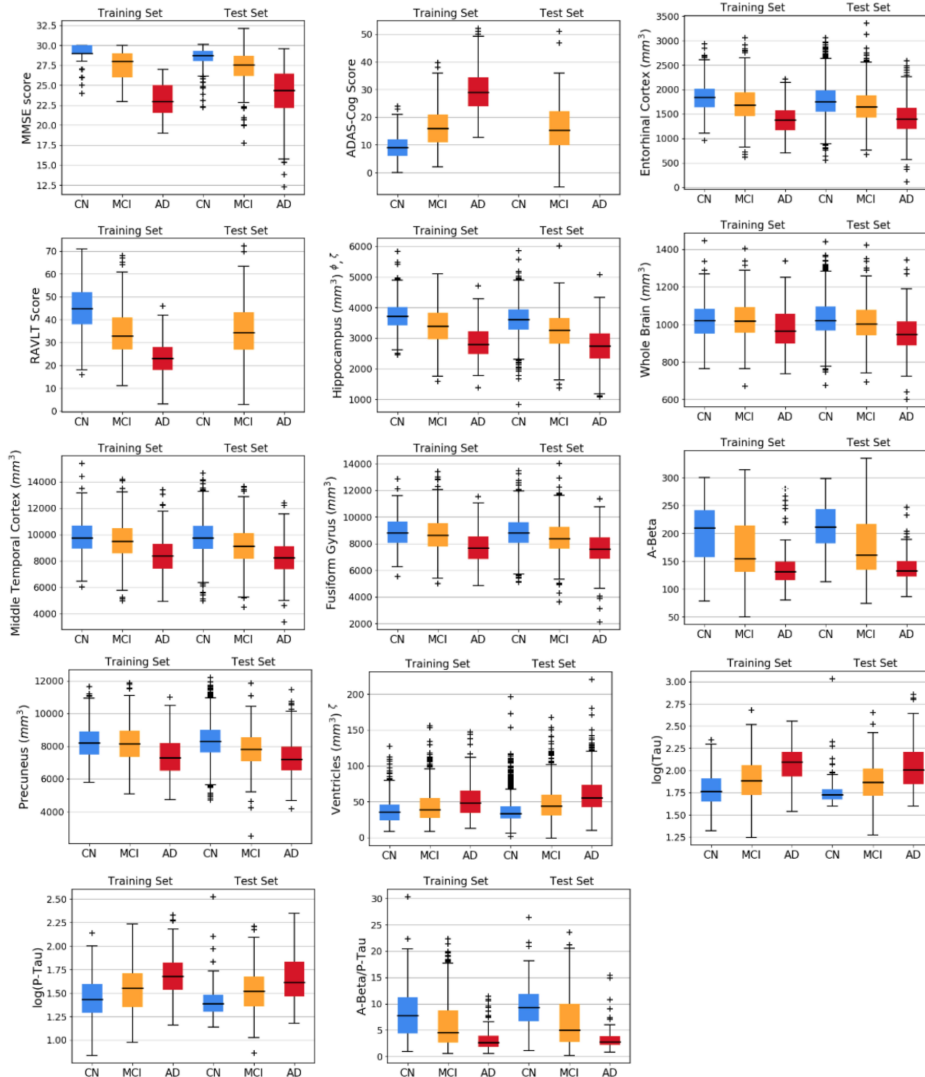


Figure 3.SF1: Box plots of biomarkers for each diagnostic category (CN, MCI and AD) for subjects from ADNI and Test datasets. Lower and upper hinges of each boxplot correspond to 25th and 75th percentiles of data. ϕ : no significant difference ($p\text{-value} \leq 0.05$) between biomarkers of MCI subjects from training and test datasets; ζ : no significant difference ($p\text{-value} \leq 0.05$) between biomarkers of AD subjects from training and test datasets

Mathematical modelling details

According to EBM and DEBM approaches, each biomarker is considered as either normal or abnormal and its probabilistic transition from the normal to the abnormal state is defined as event. The aim is to define in a data-driven manner the sequence of events that describe the most probable cascade of symptoms that characterize the transition from the healthy state to the full-blown disease spectrum.

Event Based Model (EBM)

In EBM various event sequences are sampled via a Markov Chain Monte Carlo (MCMC) process that aims at the maximization of the likelihood $P(X|S)$:

$$P(X|S) = \prod_{i=1}^N \left[\sum_{k=1}^{N_B} P(k) \prod_{j=1}^k P(x_{ij}|E_j) \prod_{j=k+1}^{N_B} P(x_{ij}|\neg E_j) \right]$$

where X denotes the entire data set, N_B is the number of biomarkers, $P(k)$ is the probability of being at stage k , x_{ij} is the j -th biomarker of subject i and $P(x_{ij}|E_j)$ and $P(x_{ij}|\neg E_j)$ are the likelihoods of measurement x_{ij} given that biomarker j has or has not become abnormal respectively, implying that events E_1, \dots, E_k already occurred and events E_{k+1}, \dots, E_{N_B} still have to occur. The central ordering is therefore the ordering for which $P(X|S)$ is maximum, or equivalently the ordering that best fits X .

Sequences are sampled via an MCMC process where at each Monte Carlo step a new sequence S' is sampled as a random swap between two biomarkers of the benchmark sequence S . If the likelihood of S' is greater than the likelihood of S , then S' is considered as the benchmark sequence for the following MCMC step. The transition to a new state can also happen if the likelihood of S' is less than the likelihood of the benchmark sequence, and in this case the transition occurs with probability:

$$p = e^{P(X|S') - P(X|S)}$$

so that event sequences can be chosen hierarchically as benchmark sequences based on their likelihood.

The normal and abnormal states for each biomarker are defined by a gaussian mixture model (GMM), where the populations of CN and AD subjects are described respectively by the normal distributions $N_j(\mu_{CN}, \sigma_{CN})$ and $N_j(\mu_{AD}, \sigma_{AD})$. To avoid the possibility that biomarkers will not show a clear bimodal distribution, the standard deviations of $P(x|E_j)$ and $P(x|\neg E_j)$ must be less or equal to the distributions of biomarkers from AD and CN subjects respectively. The mixture model distribution for each biomarker j is then found as the distribution that minimizes:

$$C_j = \sum_{i=1}^N \log (\theta_j P(x_{ij}|E_j, N_j(\mu_{AD}, \sigma_{AD})) + (1 - \theta_j) P(x_{ij}|\neg E_j, N_j(\mu_{CN}, \sigma_{CN})))$$

where N is the number of subjects in the dataset. The parameter θ_j is a mixing parameter between 0 and

1 weighs the CN and AD distributions for the j -th biomarker.

Discriminative Event Based Model (DEBM)

The approach of DEBM model for the calculation of the central ordering, on the other hand, is a two-step process where first (i) a specific ordering S_i is calculated for each subject by sorting the posterior probability that biomarker x_{ij} has become abnormal and then (ii) computing the central ordering S as the event sequence that minimizes the sum of modified Kendall's tau distances (see the following subsection) between itself and all the subject-wise orderings S_i .

As the posterior probability is influenced by the physiological variability of biomarkers, DEBM assumes that single subject orderings S_i are noisy estimates of the central ordering S . An initial estimate of the distributions of non-diseased and diseased subjects for each biomarker is performed using values from subjects at the very opposite sides of the disease spectrum, as defined by a Bayesian classifier which is trained to remove outliers and wrongly labelled data. It generates truncated Gaussian functions that neglect the tails of the distributions of the two populations, thus reducing the value of the standard deviations of the two distributions with respect to those of the whole population distribution. This allows to separate efficiently the two normal distributions for the j -th biomarker $N_j(\mu_{CN}, \sigma_{CN})$ and $N_j(\mu_{AD}, \sigma_{AD})$. With this method the resulting distributions are biased estimates of the expected distribution, characterized by smaller variance and a mean that is greater than the expected one for the distribution with the larger mean and minor for the distribution that has the smaller mean.

The biased distributions are then refined including data from all subjects via a GMM that has constraints based on the aforementioned relationships between the expected and the biased distributions, where the objective function for optimization of biomarker j is the same as that for EBM. The optimization of C_j is performed by alternatively optimizing the gaussian parameters $\mu_{CN}, \sigma_{AD}, \mu_{AD}, \sigma_{AD}$ and the mixing parameter θ_j until the latter converges. The mixing parameters of the Bayesian classifier are used as prior probabilities for the class they represent, i.e. pathological or non-pathological.

Modified Kendall's Tau

Traditional Kendall's Tau distance is often used in order to measure quantitative differences between sequences, and it can be defined as:

$$K(S, S') = \sum_{n=1}^{N-1} V_n(S, S')$$

Where S and S' are the two sequences, N is the total number of events and $V_n(S, S')$ is the number of adjacent swaps needed so that event n of sequence S is at the same position in both sequences. In a nutshell, Kendall's tau distance computes the total number of adjacent swaps that are needed to transform the sequence S' into sequence S . In DEBM model the estimates of the individual sequences are based on rankings of posterior probabilities of biomarkers being abnormal, therefore it is convenient to define a model version of Kendall's Tau distance that takes into account of posterior probabilities where swaps between events for which the difference of probability is large is penalized. In this case the number of

swaps $V_n(S, S')$ is replaced by $\widehat{V}_n(S, S')$ that is computed as:

$$\widehat{V}_n(S, S') = \sum_{l=n+1}^k p_n - p_l$$

Where n and k denote the positions of the same biomarker in S and S' respectively and p_i denotes the probability that biomarker at position i has become abnormal. After $\widehat{V}_n(S, S')$ has been calculated for a single biomarker S' is updated by swapping event at position k with event at position n .

Subject Staging

Specific methods for staging subjects on the basis of the event sequences are available in both EBM and DEBM original formulations. For the sake of simplicity, and in order to have a common staging system for both models, the method from EBM was employed in this work. This method stages each subject on the central event sequence, with the inclusion of stage 0 where no biomarker is abnormal, and assigns each individual the stage σ_i defined as:

$$\begin{aligned} \sigma_i &= \operatorname{argmax}_k P(X_i | S, k) \\ &= \operatorname{argmax}_k P(k) \prod_j^k P(x_{ij} | E_j) \prod_{j=k+1}^{N_B} P(x_{ij} | \neg E_j) \end{aligned}$$

The stage σ_i is the k -th step of the optimal sequence S that maximizes the probability that all events up to k already occurred for subject i and events from $(k+1)$ to N_B have not occurred yet given the biomarker set X_i . In case of a missing biomarker the probability of the biomarker to be in abnormal state was set to $1/2$.

| Performance metrics of EBM and DEBM including all biomarkers | | | | | | | | | | | |
|--|-------|------|------|--------|-------|-------|------|------|--------|-------|-----------------------|
| | EBM | | | | | DEBM | | | | | p |
| | k_T | Sens | Spec | BalAcc | AUC | k_T | Sens | Spec | BalAcc | AUC | |
| Training | | | | | | | | | | | |
| AD vs CN | 11 | 0.99 | 0.97 | 0.98 | 0.99* | 6 | 0.95 | 0.96 | 0.96 | 0.97* | 5.66×10^{-2} |
| AD vs MCI | 11 | 0.72 | 0.97 | 0.84 | 0.85* | 6 | 0.50 | 0.96 | 0.73 | 0.78 | 7.64×10^{-2} |
| MCI vs CN | 7 | 0.97 | 0.48 | 0.73 | 0.75* | 3 | 0.90 | 0.58 | 0.74 | 0.76* | 0.671 |
| Test | | | | | | | | | | | |
| AD vs CN | 11 | 0.90 | 0.84 | 0.87 | 0.91 | 3 | 0.70 | 0.97 | 0.84 | 0.89 | 1.24×10^{-3} |
| AD vs MCI | 11 | 0.70 | 0.84 | 0.77 | 0.80 | 9 | 0.64 | 0.82 | 0.73 | 0.78 | 0.222 |
| MCI vs CN | 1 | 0.64 | 0.69 | 0.66 | 0.68 | 1 | 0.69 | 0.66 | 0.67 | 0.67 | 0.251 |

Table 3.ST1: Measurements of area under curve (AUC), sensitivity (Sens), specificity (Spec) and balanced accuracy (BalAcc) at a specific threshold (k_T) for the subject staged with EBM and DEBM methods on train and test datasets when MMSE is included in the set of biomarkers. Thresholds are chosen to maximize the balanced accuracy in each classification task. P-values of Delong test performed to compare AUCs obtained with EBM and DEBM methods are reported in the last column. AUCs of ADNI subjects denoted with * are significantly different from their analogous of test subjects (p-value ≤ 0.05).



Figure 3.SF2: Staging based on the sequences obtained with EBM and DEBM for subjects of each test cohort. Staging of subjects from all diagnostic categories (Cognitively normal (CN) in blue, mild cognitive impairment (MCI) in orange, and Alzheimer's disease (AD) in red) are shown for the cases of (a) ADC subjects on EBM sequence; (b) ADC subjects on DEBM sequence; (c) ARWiBo subjects on EBM sequence; (d) ARWiBo subjects on DEBM sequence; (e) EDSD subjects on EBM sequence; (f) EDSD subjects on DEBM sequence; (g) OASIS subjects on EBM sequence; (h) OASIS subjects on DEBM sequence; (i) ViTA subjects on EBM sequence; (j) ViTA subjects on DEBM sequence.

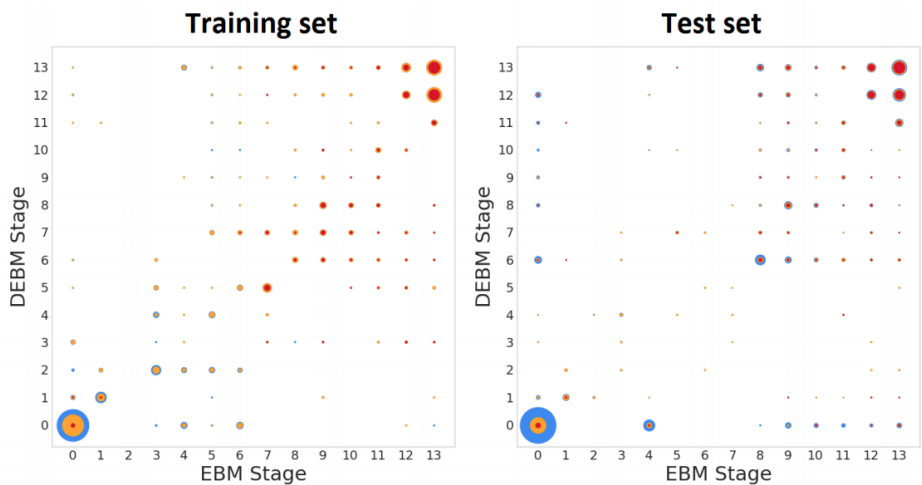


Figure 3.SF3: Scatter plot of DEBM stage vs. EBM stage for training (left) and test (right) subjects. Areas of annuli are proportional to the number of subjects. Linear regression resulted in slopes of 0.891 ($R^2=0.802$) for training subjects and 0.829 ($R^2=0.680$) for test subjects.

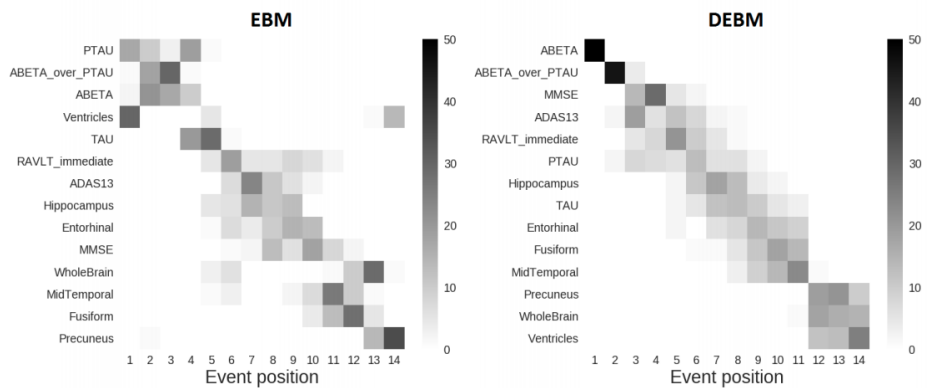


Figure 3.SF4: Positional variance diagrams of Event ordering obtained with EBM and DEBM when MMSE is included in the original set of biomarkers. Both diagrams show the number of times each biomarker occurred in a certain position from a batch of 50 independent bootstrapped sequences generated from biomarkers of subjects from the training set with EBM (left) and DEBM (right) methods.

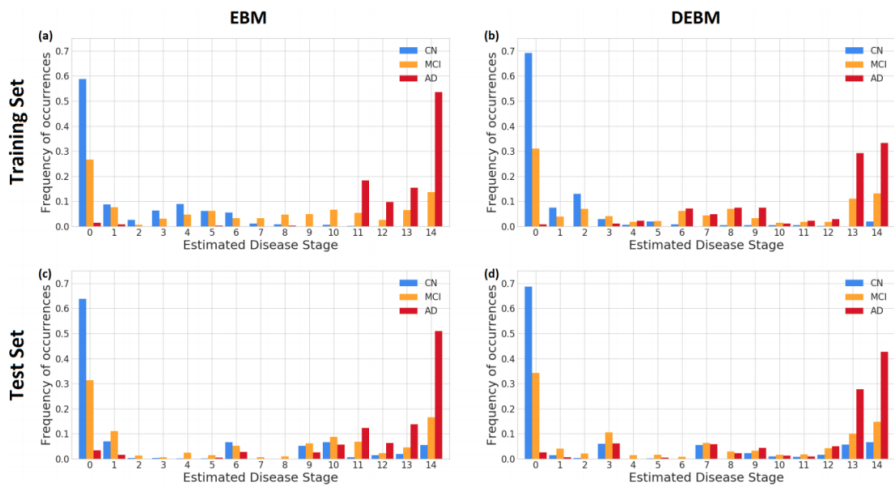


Figure 3.SF5: Subject staging based on the sequences obtained with EBM and DEBM methods when MMSE is included in the set of biomarkers. Staging of subjects from all diagnostic categories (Cognitively normal (CN) in blue, mild cognitive impairment (MCI) in orange, and Alzheimer’s disease (AD) in red) are shown for the cases of (a) training subjects on EBM sequence; (b) training subjects on DEBM sequence; (c) test subjects on EBM sequence and (d) test subjects on DEBM sequence. Histograms are normalized for each diagnostic category.

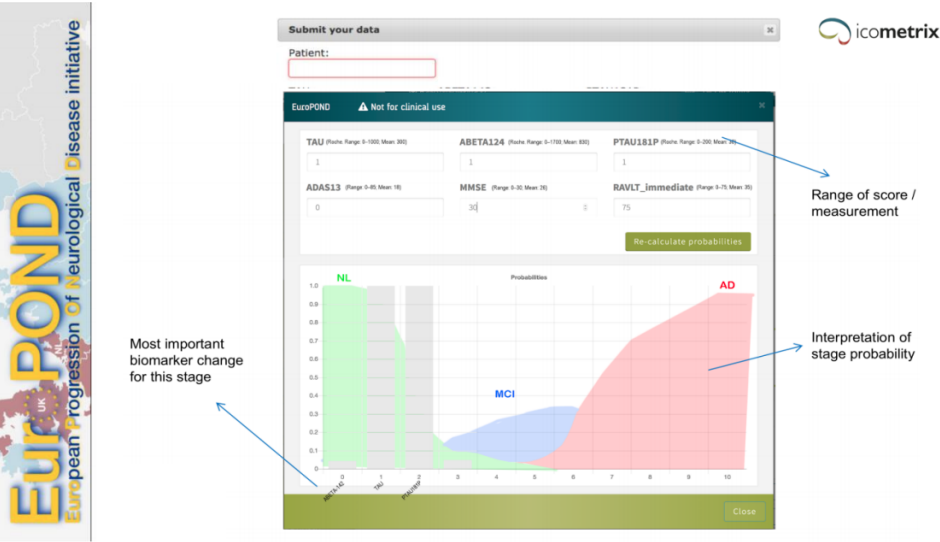


Figure 3.SF6: Single case interface (alpha release developed by Icometrix NV in the context of the EuroPOND H2020 initiative) to stage patient according to the biomarkers data availability. X-axis reports the patient’s stage, Y-axis reports the probability of biomarker abnormalities.

Chapter 4

Modelling the cascade of biomarker changes in progranulin related frontotemporal dementia

This chapter contains the content of the manuscript ‘Modelling the cascade of biomarker changes in *GRN* related frontotemporal dementia. Jessica L. Panman*, **Vikram Venkatraghavan***, Emma L. van der Ende, Rebecca Steketee, Lize C. Jiskoot, Jackie M. Poos, Elise G.P. Dopper, Lieke H.H. Meeter, Laura Donker Kaat, Serge A.R.B. Rombouts, Meike W. Vernooij, Anneke J.A. Kievit, Enrico Premi, Maura Cosseddu, Elise Bonomi, Jaume Olives, Jonathan D. Rohrer, Raquel Sanchez-Valle, Barbara Borroni, Esther E. Bron, John C. van Swieten, Janne M. Papma, Stefan Klein, GENFI consortium investigators. *Journal of Neurology, Neurosurgery & Psychiatry*, Accepted, 2020’. An earlier version of this chapter was presented (poster) at the Alzheimer’s Association International Conference (AAIC) 2020.

* Contributed equally in this work

Vikram Venkatraghavan’s contributions in this chapter are: Methodology, software, validation, and writing.

Abstract

Objective: Progranulin related frontotemporal dementia (FTD-GRN) is a fast progressive disease. Modelling the cascade of multimodal biomarker changes aids in understanding the etiology of this disease and enables monitoring of individual mutation carriers. In this cross-sectional study, we estimated the temporal cascade of biomarker changes for FTD-GRN, in a data-driven way.

Methods: We included 56 presymptomatic and 35 symptomatic *GRN* mutation carriers, and 35 healthy non-carriers. Selected biomarkers were neurofilament light chain (NfL), grey matter volume, white matter microstructure, and cognitive domains. We used discriminative event-based modelling to infer the cascade of biomarker changes in FTD-GRN and estimated individual disease severity through cross-validation. We derived the biomarker cascades in non-fluent variant primary progressive aphasia (nfvPPA) and behavioural variant FTD (bvFTD) to understand the differences between these phenotypes.

Results: Language functioning and NfL were the earliest abnormal biomarkers in FTD-GRN. White matter tracts were affected before grey matter volume, and the left hemisphere degenerated before the right. Based on individual disease severities, presymptomatic carriers could be delineated from symptomatic carriers with a sensitivity of 100% and specificity of 96.1%. The estimated disease severity strongly correlated with functional severity in nfvPPA, but not in bvFTD. In addition, the biomarker cascade in bvFTD showed more uncertainty than nfvPPA.

Conclusion: Degeneration of axons and language deficits are indicated to be the earliest biomarkers in FTD-GRN, with bvFTD being more heterogeneous in disease progression than nfvPPA. Our data-driven model could help identify presymptomatic *GRN* mutation carriers at risk of conversion to the clinical stage.

4.1 Background

Mutations in the progranulin (*GRN*) gene on chromosome 17q21 are a major cause of autosomal dominant inherited frontotemporal dementia (FTD) [131, 132]. The majority of mutation carriers develops a behavioural variant FTD (bvFTD) phenotype [133], and another significant proportion of patients present with non-fluent variant primary progressive aphasia (nfvPPA) [133, 134]. The age of symptom onset varies between 35 and 90 in *GRN* mutation carriers [131, 132], without clear associations with familial age of onset [134]. Brain changes in FTD-GRN patients can evolve symmetrically, or predominantly asymmetrically, in either the left or right hemisphere [86, 135].

Recent longitudinal studies have suggested that the time-window between emerging pathophysiological changes and the first clinical symptoms is short in *GRN* mutation carriers, and covers only two to four years [10, 136]. During this period, the serum neurofilament light chain (NfL) level – a marker of axonal degeneration – increases two to three-fold [30, 137], loss of grey and white matter emerges [10, 138], and cognitive functioning declines [136]. However, most of the biomarker studies in FTD-GRN have investigated one type of

biomarker, i.e. fluid, neuroimaging, or cognition, leaving the temporal relations and ordering of these biomarkers unknown. These temporal relations could potentially provide novel insights into disease progression mechanisms in *GRN* mutation carriers. Moreover, because of the fast progression of pathophysiological changes, determining the earliest abnormal biomarker is crucial, as the optimal window of opportunity for treatment might be small.

Recently, novel data-driven methods for disease progression modelling have emerged, focusing on the cascade of biomarker changes [52, 70]. Event-based models are a class of disease progression models that estimate the cascade of biomarker changes derived from cross-sectional data [53, 70, 86]. This is done without strong a priori assumptions regarding the relationship between different biomarkers. A promising novel method that estimates the cascade of biomarker change is Discriminative Event-Based Modelling (DEBM) [70, 93]. This model is robust to disease phenotypic heterogeneity in a cohort and can handle missing data.

In this study, we use DEBM to estimate the temporal cascade of biomarker changes in presymptomatic and symptomatic FTD-GRN mutation carriers, distinguishing between early and late biomarkers. Furthermore, we determine phenotypic differences in patterns of biomarker changes in *nfvPPA* and *bvFTD*, to gain more insights into their distinct disease progression mechanisms.

4.2 Methods

4.2.1 Sample and study procedures

Subjects were recruited prospectively from three European centres of the Genetic Frontotemporal dementia Initiative (GENFI): Rotterdam (the Netherlands), Brescia (Italy), and Barcelona (Spain). We collected cognitive and clinical data, MRI, and serum samples from 126 participants. We included 35 symptomatic *GRN* mutation carriers (Rotterdam: $n=11$, Brescia: $n=22$, Barcelona: $n=2$), 56 presymptomatic *GRN* mutation carriers (Rotterdam: $n=33$, Brescia: $n=17$, Barcelona: $n=6$), and 35 cognitively healthy non-carriers (Rotterdam: $n=34$, Brescia: $n=0$, Barcelona: $n=1$). Local clinical genetics departments performed DNA genotyping to confirm the presence of a *GRN* mutation. Non-carriers were first-degree family members of *GRN* patients without a mutation. Symptomatic mutation carriers were diagnosed based on the established clinical criteria for *bvFTD* [139] ($n=17$), *nfvPPA* [140] ($n=16$), or cortico-basal syndrome [141] ($n=2$). Mutation carriers were defined as presymptomatic when clinical criteria were not fulfilled, i.e., behavioural or cognitive symptoms were absent [142]. Clinical questionnaires were administered to the caregiver, spouse, or a family member, i.e. the Frontotemporal Lobar Degeneration Clinical Dementia Rating scale sum of boxes (FTD-CDR-SB) [143], the Neuropsychiatric Inventory (NPI [144], and the Frontotemporal Dementia Rating scale (FRS) [145]. The study was carried out

according to the declaration of Helsinki, approved by the local medical ethics board at each site, and all participants provided written informed consent.

4.2.2 Biomarker collection and processing

Biomarker selection

For biomarker selection, we performed a literature search using Pubmed. We included studies that (i) performed research in presymptomatic *GRN* mutation carriers, and (ii) biomarker studies that examined biomarkers in blood or CSF, neuroimaging biomarkers and cognition. We selected serum NfL [137], MMSE, cognitive domains of attention and processing speed, executive functioning, language, and social cognition [136, 146]; left and right grey matter volumes of the insula, frontal lobe, parietal lobe and temporal lobe [10, 138]; left and right white matter tracts of the anterior thalamic radiation, superior longitudinal fasciculus, uncinate fasciculus, and the forceps minor [10, 147]. For detailed information about the literature review and subsequent biomarker selection, please see Supplementary material 4A.

Neurofilament light chain

Serum samples were obtained through venepunctures and analysed with single molecular assay technology, as described previously [30]. Samples were measured in a single laboratory, in duplicate, with an intra-assay coefficient of variation below 5%. Inter-assay variation between batches was below 8%. NfL concentrations were expressed in pg/ml.

Magnetic Resonance Imaging

3D T1-weighted and diffusion tensor imaging were acquired with 3T MRI scanners across the three sites. MRI was missing in 25 participants due to unavailability (n=16) and insufficient quality due to motion artefacts (n=9). Availability of MRI and an overview of the scanning protocols are listed in Supplementary material, Table 4.ST1. Image processing was carried out in FMRIB Software Library [148], using default pipelines for grey matter volumes and white matter tracts. For grey matter volumetric regions of interest (ROI), we used the Montreal Neurological Institute (MNI) atlas [149], and for the fractional anisotropy of white matter tracts, we used the Johns Hopkins' University atlas [150]. Left and right regions and tracts were considered separately. Raw regional volumes and fractional anisotropy values were transformed to z-scores, based on the mean and standard deviation from the non-carriers. A detailed description of processing and ROI calculation is reported in supplementary material 4A.

Cognitive assessment

Cognitive data were collected from all participants in four cognitive domains, described in detail in supplementary material 4A. Raw cognitive test scores were transformed to z-scores based on the mean and standard deviation in non-carriers, and then combined into cognitive domain scores similar to previous studies [136].

Confounding factors correction

All selected biomarkers were tested for normality (see supplementary material 4A for details) and log-transformed in case of a skewed distribution. As most non-carriers originated from one centre, we used presymptomatic subjects for regressing out possible confounding effects using multiple linear regression, before continuing with event-based modelling. NfL levels were corrected for age and sex. Grey matter volumes and fractional anisotropy values were corrected for age, sex, total intracranial volume and MRI scanning protocol. Cognitive domain scores were corrected for confounding effects of age, sex and total years of education.

4.2.3 Temporal cascade of biomarker changes

The DEBM model introduced in Venkatraghavan et al. [70, 93] estimates the cascade of biomarker changes in a three-step process. For each biomarker, it first estimates the distributions of normal and pathological (or abnormal) values using Gaussian mixture modelling (GMM), and uses these to compute, for each subject, the probability that the biomarker is abnormal (explained in detail in supplementary material 4B). The method then estimates the biomarker cascade independently for each subject based on the biomarker values present for that subject. The mean cascade is estimated such that the sum of the probabilistic Kendall's Tau distances is minimized between the mean cascade and all the subject-specific cascades. For subjects with missing biomarker values, only the corresponding subset of the biomarker cascade present in the subject-specific cascade is used to compute the probabilistic Kendall's Tau distance. Lastly, the severity of disease as a summary measure for each subject is computed by estimating the subject's progression along the resulting disease progression timeline. In this section, we describe the experiments we performed for estimating the cascade of biomarker changes for non-imaging biomarkers, as well as for neuroimaging and non-imaging biomarkers together.

DEBM model for non-imaging biomarkers

As imaging was missing in a lot of subjects ($n=25$), we first estimated the cascade of biomarker changes procedure with solely NfL and cognitive biomarkers. Since the non-carriers are healthy in this cohort, the normal Gaussians were fixed at the mean and standard deviation of the biomarker values of the non-carriers. We used GMM only to estimate the

abnormal Gaussian and the mixing parameter for each biomarker. In order to estimate the positional variance in the estimated cascade, the entire dataset was randomly sampled using bootstrap sampling with 100 different random seeds, and the cascade of biomarker change was estimated for each of those randomly sampled datasets [70, 93].

DEBM model for neuroimaging and non-imaging biomarkers together

For the imaging biomarkers, we modified the GMM step in DEBM to make it better suited for the FTD-GRN population, known for its asymmetric pattern of atrophy [135]. Abnormal values of biomarkers that typically become abnormal late in the disease are usually under-represented in a specific patient population as compared to the early biomarkers. This could make the GMM of late biomarkers unstable, as previously reported [93]. Due to the asymmetrical atrophy patterns of FTD-GRN [86, 135], lateralized neuroimaging biomarkers that become abnormal early in the disease process may have a corresponding biomarker from the other hemisphere that remains stable until much later in the disease process. To exploit this, we assumed that the normal and abnormal Gaussians from the left and right hemispheric biomarkers (expressed as z-scores) are the same, and the biomarkers from both hemispheres only differ in their position along the disease progression timeline. With this assumption, we proposed a novel modification to the GMM optimization called Siamese GMM, in which the biomarkers of the same region from left and right hemispheres are jointly optimized. The abnormal and normal Gaussians are shared between the left and right hemispheres, but the mixing parameters are independently estimated (see supplementary material 4B for details). In this way, the numerical stability of GMM optimization in the late neuroimaging biomarkers improved.

For non-imaging biomarkers, GMM was performed as described in the previous section. After GMM, further steps of DEBM modelling were carried out as usual, to estimate the complete cascade of neuroimaging and non-imaging biomarker changes in presymptomatic and symptomatic GRN mutation carriers. The positional variance in the estimated cascade was again estimated using bootstrap sampling with 100 different random seeds. For brevity, in the remainder of the paper we refer to this model, which integrates neuroimaging and non-imaging biomarkers, as the multimodal DEBM.

Validation

To validate the DEBM models, we used 10-fold cross-validation. In each fold of the cross-validation, the DEBM model was built in the training set and the disease severity was estimated in the test set. We distinguished symptomatic mutation carriers from presymptomatic mutation carriers, and reported the corresponding sensitivity and specificity. Furthermore, in bvFTD and nfvPPA subjects, the estimated disease severity was correlated

with years since symptom onset and FTD-CDR-SB scores, using Pearson correlation. Symptomatic carriers without imaging biomarkers were excluded for the validation of the multimodal DEBM but were included in the non-imaging DEBM.

4.2.4 Differential phenotype analysis

In order to examine the differences between bvFTD and nfvPPA variants of FTD-GRN, we built separate DEBM models. Presymptomatic subjects were excluded from this analysis as no phenotype information is available. The numbers of symptomatic subjects in each group (17 with bvFTD, 16 with nfvPPA) are too small to build complete DEBM models reliably. As a solution, we assumed that the biomarkers for the two phenotypes shared the same normal and abnormal biomarker distributions, and that they only differ in their position along the disease progression timeline. We hence optimized the GMM such that the normal and abnormal Gaussians were estimated without considering the phenotypes, whereas the mixing parameters were estimated separately for each phenotype. As before, we estimated the cascade of biomarker changes in the two phenotypes for non-imaging and multimodal (neuroimaging and non-imaging together) biomarkers.

4.3 Results

4.3.1 Sample

A total of 126 subjects were included in this study. Availability and characteristics of the data are presented in Table 4.1. Details on biomarker availability and characteristics can be found in supplementary material 4A, Tables 4.ST2 and 4.ST3. Symptomatic mutation carriers were older, had fewer years of education, and had higher scores on the NPI and FTD-CDR-SB, and lower scores on the FRS than both presymptomatic mutation carriers and non-carriers. There were no differences in demographic or clinical characteristics between presymptomatic mutation carriers and non-carriers.

4.3.2 Cascade of biomarker changes

Non-imaging and multimodal DEBM models

In Figure 4.1a and 4.1b, we show the estimated mean cascade of biomarker changes and the uncertainty within the model for non-imaging and multimodal biomarkers. Language was the earliest biomarker to become abnormal followed by neurofilament light chain. It can be seen in Figure 4.1b that, left anterior thalamic radiation, left insula, and bilateral uncinate fasciculi were the earliest imaging biomarkers. It can also be observed that imaging biomarkers from the left-hemisphere became abnormal earlier than their right counterpart. GMM estimations with normal and abnormal Gaussian distributions are shown in Figure 4.2,

| | Symptomatic | | | Presymptomatic | | Non-carriers |
|--------------------------|----------------------------|------------------------------|----------------------------|----------------|---------------|--------------|
| | Total | bvFTD | ntvPPA | | | |
| N | | | | | | |
| Subjects (% female) | 35 ^A (60%) | 17 (47.1%) | 16 (75%) | 56 (69.6%) | 35 (54.4%) | |
| Rotterdam | 11 | 8 | 3 | 33 | 34 | |
| Brescia | 22 ^A | 9 | 11 | 17 | 0 | |
| Barcelona | 2 | 0 | 2 | 6 | 1 | |
| Data availability | | | | | | |
| Serum NFL | 91.7% | 88.9% | 93.8% | 98.21% | 91.67% | |
| Cognitive assessment | 91.7% | 88.9% | 93.8% | 69.64% | 91.67% | |
| T1-weighted MRI | 44.4% | 38.9% | 50.0% | 96.4% | 88.6% | |
| DTI | 50.0% | 44.4% | 56.3% | 92.9% | 91.4% | |
| Sample Characteristics | | | | | | |
| Age (years) | 62.57 ± 6.72 ^B | 62.93 ± 6.11 ^C | 61.78 ± 7.78 ^D | 51.52 ± 11.42 | 55.15 ± 12.55 | |
| Education (years) | 10.61 ± 4.59 ^B | 10.27 ± 4.91 ^C | 11.79 ± 4.02 | 13.79 ± 3.27 | 13.21 ± 2.84 | |
| TIV (litres) | 1.44 ± 0.17 | 1.50 ± 0.17 | 1.42 ± 0.14 | 1.39 ± 0.15 | 1.40 ± 0.14 | |
| NPI | 23.77 ± 28.38 ^B | 28.90 ± 30.64 ^{C,E} | 6.67 ± 6.03 ^E | 1.87 ± 3.37 | 2.24 ± 4.32 | |
| FRS | 56.50 ± 30.43 ^B | 48.86 ± 29.91 ^C | 67.20 ± 30.96 ^D | 97.27 ± 10.11 | 95.47 ± 7.45 | |
| FTD-CDR-SB | 7.64 ± 6.52 ^B | 9.68 ± 7.47 ^{C,E} | 5.25 ± 4.37 ^{D,E} | 0.04 ± 0.21 | 0.00 ± 0.00 | |
| Disease duration (years) | 2.45 ± 2.01 | 2.37 ± 1.92 | 2.48 ± 2.29 | N/A | N/A | |

Table 4.1: Data availability and characteristics. Abbreviations: bvFTD = behavioural variant frontotemporal dementia, ntvPPA = non-fluent variant primary progressive aphasia, NFL = neurofilament light chain, DTI = diffusion tensor imaging, TIV = total intracranial volume, GM = grey matter, NPI = Neuropsychiatric Inventory, FRS = Frontotemporal dementia rating scale, FTD-CDR-SB = Frontotemporal dementia Clinical Dementia Rating Scale Sum of Boxes. Sample characteristic variables are expressed as mean ± standard deviation. ^A The two remaining patients presented with cortico-basal syndrome. ^B Significant difference between symptomatic carriers and presymptomatic as well as non-carriers. ^C Significant difference between bvFTD patients and presymptomatic as well as non-carriers. ^D Significant difference between ntvPPA patients and presymptomatic as well as non-carriers. ^E Significant difference between bvFTD patients and ntvPPA patients.

where the estimated Gaussians are seen to fit the observed histograms well. Figure 4.1c shows the positional variance of the cascade of multimodal biomarker changes obtained when GMM of the imaging biomarkers was done without using Siamese GMM. Generally, the positional variance was smaller with Siamese GMM than without.

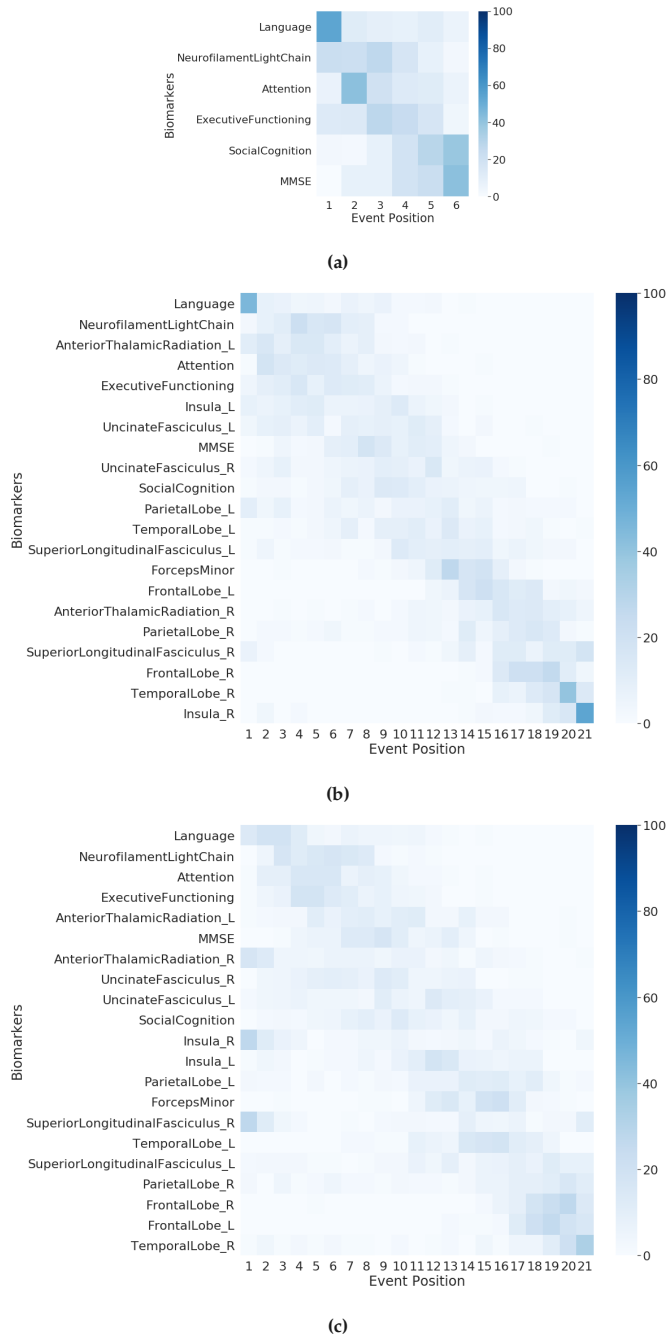


Figure 4.1: Cascade of biomarker changes in FTD-GRN along with the uncertainty associated with it. (a) Non-imaging biomarkers, (b) Multimodal biomarkers with Siamese GMM, (c) Multimodal biomarkers without Siamese GMM. The biomarkers are ordered based on the position in the estimated cascade. The color-map is based on the number of times a biomarker is at a position in 100 repetitions of bootstrapping.



Figure 4.2: Gaussian mixture modelling distributions. The histogram bins are divided in three colours, where the green part shows the proportion of non-carriers, the yellow part shows the proportion of presymptomatic carriers and the red part shows the proportion of symptomatic carriers. The Gaussians shown here are the ones that were estimated using Gaussian mixture modelling, where the green Gaussian is the normal one estimated using non-carriers and the red Gaussian is the abnormal one estimated using the carriers. The amplitudes of these Gaussians are based on the estimated mixing parameter. The grey curve shows the total estimated distribution, which is the summation of green and red Gaussians.

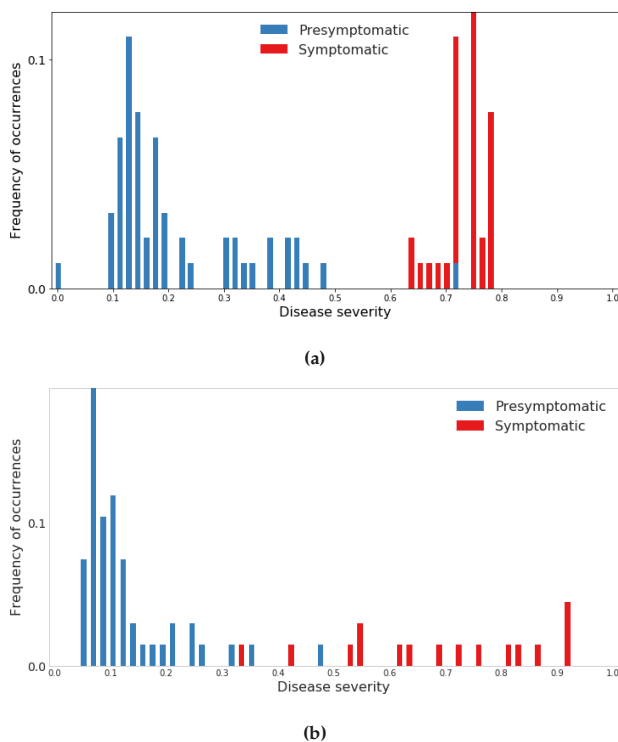


Figure 4.3: Frequency of occurrence of subjects with different disease severities, estimated using cross-validation. (a) results using non-imaging biomarkers in DEBM, (b) results using multimodal biomarkers in DEBM.

Validation

Figure 4.3a and 4.3b shows the estimated disease severity when using non-imaging and multimodal biomarkers respectively. It can be seen that estimated disease severity delineated the symptomatic subjects from the pre-symptomatic subjects. The sensitivity and specificity of this delineation were 1.0 and 0.982 respectively while using non-imaging biomarkers, and 1.0 and 0.961 respectively while using multimodal biomarkers.

Figure 4.4 shows the correlation of the estimated disease severity with years since symptom onset and FTD-CDR-SB for nfvPPA and bvFTD subjects, when using multimodal DEBM. It can be seen from Figure 4.4 that estimated disease severity strongly correlated with years since symptom onset ($R=0.95$, $p=0.0003$), and the FTD-CDR-SB ($R=0.84$, $p=0.0189$), in nfvPPA patients. However, estimated disease severity correlated poorly with years since symptom onset ($R=0.22$, $p=0.6331$) and the FTD-CDR-SB ($R=0.28$, $p=0.5866$) in bvFTD patients. Figure 4.SF2 in supplementary material 4B shows a similar plot when using non-imaging biomarkers, where estimated disease severity did not correlate with years since

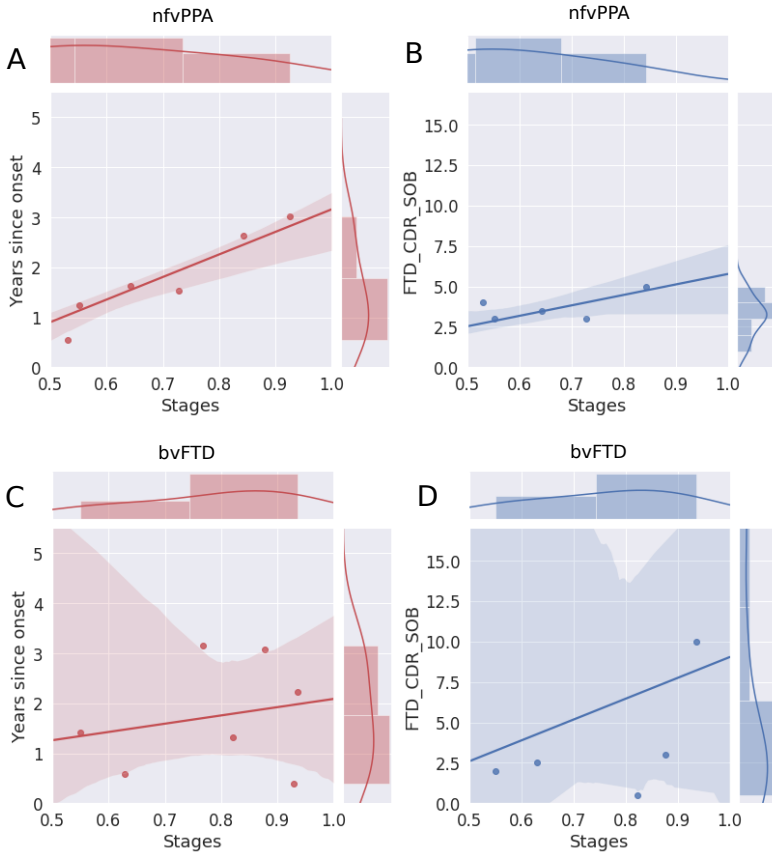


Figure 4.4: Correlation of disease severity (as estimated by multimodal DEBM using cross-validation) with years since onset and FTD-CDR-SOB. The 2D scatter plots in figures (a) and (c) show the correlations of disease severity with years since onset, for symptomatic nfvPPA and bvFTD subjects respectively. The 2D scatter plot in figures (b) and (d) show the correlations of disease severity with FTD-CDR-SOB. The plot on top of each subfigure shows the probability density function of the disease stages. The plots on the right of figures (a) and (c) show the probability density functions of years since symptom onset. The plots on the right of figures (b) and (d) show the probability density function of FTD-CDR-SOB.

symptom onset and FTD-CDR-SB, neither for nfvPPA nor for bvFTD subjects.

Differential phenotype analysis

Figure 4.5 shows the multimodal biomarker cascade for nfvPPA and bvFTD phenotypes. nfvPPA patients showed language and NfL as first abnormal biomarkers followed by other cognitive domains. Left-hemispheric imaging biomarkers became abnormal before right-hemispheric imaging biomarkers, starting with the uncinated fasciculus (white matter integrity), insula and temporal lobe (grey matter volume). Only the left superior longitudinal

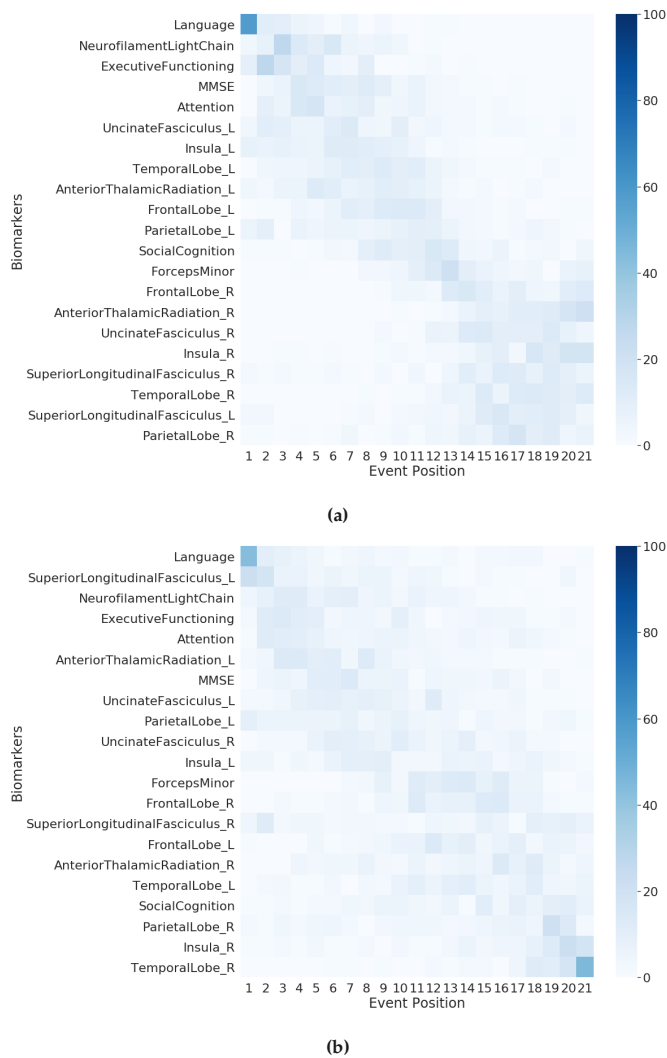


Figure 4.5: Cascade of multimodal biomarker changes in nfvPPA (a) and bvFTD (b) subjects along with the uncertainty associated with it. The biomarkers are ordered based on the position in the estimated cascade. The color-map is based on the number of times a biomarker is at a position in 100 repetitions of bootstrapping.

fasciculus was estimated as late biomarker, even later than its right-sided counterpart.

Interestingly, in bvFTD patients, the biomarker ordering also indicated that language and NfL were the earliest abnormal biomarkers. In contrast to the nfvPPA, the left superior longitudinal fasciculus (white matter integrity) was estimated as the first abnormal imaging biomarker in bvFTD. However, the biomarker orderings in bvFTD were predominantly

characterized by large uncertainty in the positioning of biomarkers in the disease timeline, with hardly any observable distinction between early and late biomarkers. Figure 4.SF3 in supplementary material 4B presents the non-imaging biomarker cascade for the two phenotypes, showing that the uncertainty in the mean cascade in bvFTD is more than in nfvPPA.

4.4 Discussion

In this study, we estimated the cascade of biomarker changes in FTD-GRN. We validated our model by delineating the symptomatic mutation carriers from the presymptomatic mutation carriers using the estimated disease severity. We demonstrated that language and NfL levels are the earliest biomarkers to become abnormal in the FTD-GRN spectrum. Other early biomarkers were the white matter microstructure of the thalamic radiation and the cognitive domain of attention and mental processing speed.

Our findings support other studies that proposed NfL as an early biomarker for disease onset in FTD-GRN [30, 137]. We demonstrated that the left anterior thalamic radiation also degenerated early. This is also supported by previous studies which suggested that white matter microstructure markers may correlate with changes in NfL [137, 151]. Cognitive changes in attention, mental processing speed, and executive functioning occurred relatively early in the estimated disease progression timeline. This corresponds well with the early white matter changes (i.e. NfL and fractional anisotropy changes), as attention and processing speed are cognitive functions that highly depend upon the integrity of axons and their myelin sheaths [152, 153]. The early involvement of these biomarkers point towards axonal degeneration as one of the first pathological processes in GRN mutation carriers. However, it must be noted that the estimated cascade shows the sequence of biomarker events when they are detectably abnormal. One of the important factors that affects the detectability of biomarker abnormality in a cross-sectional data set is the overlap between the normal and abnormal biomarker distributions. Therefore, the presented cross-sectional model cannot provide insight into the sequence of earliest (hardly detectable) changes in the carriers' biomarker levels. Figure 4.2 showed that the overlap in cognitive biomarkers is was relatively smaller than the overlap in neuroimaging biomarkers, which could explain the relative early positioning of the cognitive biomarker events.

With the differential phenotypic analysis, we estimated the biomarker cascade for nfvPPA and bvFTD patients. Strikingly, language functions deteriorated early in both nfvPPA and bvFTD. While not currently embedded in the clinical criteria for bvFTD [139], our results demonstrate the importance of decreased language functions in both phenotypes. This is in line with multiple previous studies [154, 155, 156]. In addition, multiple determinants of the complex language network were also affected early, for example the left insula, and

uncinate fasciculus [157]. While language deficits were estimated as the first detectable abnormal biomarker, the overlap with the second, the elevation in NfL levels, complicates distinguishing the timeline of these disease events. Furthermore, as depicted in Figure 4.2, (subtle) language deficits were less specific for disease onset than NfL levels. However, the high sensitivity of the language biomarker in our study, and the relative uncomplicated administration of language tests (compared to neuroimaging techniques, for example) offers potential for longitudinal research in the preclinical stage of FTD-GRN ideally in combination with NfL levels.

For nfvPPA, NfL levels and other cognitive domains became abnormal in early disease stages, consistent with findings from previous studies [30, 137, 158]. In addition, we showed that left hemispheric tracts and regions were affected in nfvPPA patients before right regions, accordant with the previously reported strong involvement of the left hemisphere in primary progressive aphasia [159, 160]. We showed that NfL levels and cognitive domains may be possible biomarkers for disease onset, while neuroimaging markers were highly correlated with clinical indicators of progression (years since onset, FTD-CDR-SB).

For bvFTD, however, the biomarker cascade was characterized by large uncertainty, and the estimated disease severities did not correlate with actual years since onset or FTD-CDR-SB. This uncertainty could indicate large neuroanatomical heterogeneities between bvFTD patients. Differences in neuroanatomical atrophy patterns have been associated with FTD-GRN patients before [86, 135]. Here, we demonstrated that this anatomical heterogeneity is predominantly associated with the bvFTD phenotype, while nfvPPA patients showed a clear pattern of left hemispheric degeneration before the right hemisphere was affected. Furthermore, bvFTD patients present with cognitive symptoms such as impaired social conduct and executive function but can also have severe memory problems. In summary, within the group of bvFTD, spatial and temporal brain degeneration and cognitive changes are more heterogeneous than in the nfvPPA group.

From a methodological point of view the strength of this paper lies in the introduction of the Siamese GMM approach in DEBM. We showed that Siamese GMM reduces the positional variance in neuroimaging biomarkers, most notably in the right insula, the right anterior thalamic radiation and the right superior longitudinal fasciculus. This is because GMM is known to be unstable in the presence of biomarkers with a large overlap between the normal and abnormal Gaussians [93]. This is often the case in biomarkers becoming abnormal late in the disease and having very few samples representative of the typical abnormal values expected in the disease. The joint GMM in the Siamese counterpart exploits the knowledge that FTD-GRN is generally an asymmetric brain disease, and uses the neuroimaging biomarkers that become abnormal early in the disease process to aid the GMM of its hemispheric counterpart that becomes abnormal far later in the disease process. Another

strong point about the DEBM model is that it infers disease progression from cross-sectional data, which is more readily available than longitudinal data, especially in a rare disease as FTD-GRN.

From the clinical point of view, a major strength of our study is the large, well-defined cohort of presymptomatic and symptomatic *GRN* mutation carriers, and availability of multimodal (i.e. fluid, imaging, and cognitive) biomarkers. Although we did not have FLAIR or T2 imaging data available for the current study, it would be interesting to incorporate white matter lesions in a future version of the model, as a number of studies have indicated the presence of white matter lesions in FTD-GRN carriers [161]. Additionally, including functional neuroimaging measures in future studies possibly provides new insights into the temporal biomarker sequence and underlying disease mechanism as well. Recent papers have addressed functional changes in FTD-GRN, showing thalamic-cortical hyperconnectivity in early preclinical stages [162] and presymptomatic abnormalities in neurophysiology [163].

A minor limitation in our study is the difference in mean age between the non-carrier, presymptomatic, and symptomatic mutation carrier groups. We adjusted for this in the analysis rather than matching the groups. It should be noted that the small sample size may have caused a large part of the uncertainty of our model, especially in the case of missing (neuroimaging) biomarkers. Our bvFTD and nfvPPA samples due to *GRN* mutations were relatively large compared to previous studies [164]. However, the DEBM model would improve substantially if the phenotypic samples were larger, as we could only include symptomatic subjects for the phenotypic analysis. Uncertainties in the estimation of the phenotypic biomarker cascades may be improved with upcoming longitudinal data, when some of the converted mutation carriers can be included in the phenotypic models.

In conclusion, with this DEBM study in the FTD-GRN spectrum, we were able to demonstrate that language functions and NfL levels are the earliest abnormal biomarkers, regardless of phenotype. However, bvFTD show more heterogeneity and uncertainty in disease progression, pointing towards more variability in biomarkers than nfvPPA. Our analyses suggest axonal degeneration and damage to the language network as the earliest biomarkers in *GRN* mutation carriers, which could potentially be used as endpoints in clinical trials for disease modifying treatments. Future efforts should be directed at confirmation and validation of these findings with longitudinal data. Future validation of these results in an external cohort such as the LEFFTDS [165] could further aid in confirming these results and elucidate any ethnic variations in the disease progression timeline. We expect that DEBM modelling will benefit individual prediction of symptom onset in the future, and may optimize selection of eligible mutation carriers for clinical trials.

Supplementary material 4A: Biomarkers

Biomarker selection

For biomarker selection, we extensively searched for relevant literature about presymptomatic FTD-GRN in Pubmed. We reviewed all empirical studies that included at least a presymptomatic GRN mutation carrier group. Next, we determined which biomarkers were frequently reported as abnormal in previous empirical studies and included these biomarkers accordingly, restricted to fluid biomarkers, grey matter brain regions, white matter tracts, and cognition. The selected biomarkers were: serum NfL [30, 137, 166], MMSE [136, 138, 163], cognitive domains of language, attention and processing speed, executive functioning, and social cognition [136, 146, 167, 168]; left and right volumes of the insula, frontal lobe, parietal lobe and the temporal lobe [138, 163, 169, 170, 171] [10, 86, 142, 172, 173, 174, 175, 176, 177]; white matter tracts: left and right fractional anisotropy of anterior thalamic radiation, superior longitudinal fasciculus, uncinate fasciculus, and forceps minor [10, 142, 147, 169, 175, 178]. Although the GRN mutation affects plasma progranulin protein levels, these levels were not selected as biomarker, as research has shown that these remain stable in both the presymptomatic and symptomatic stage [163, 179].

MRI processing and ROI calculation

An overview of MRI acquisition parameters is presented in supplementary material Table 4.ST1. The standard voxel-based morphometry pipeline from FSL [148, 180, 181] was used to process T1-weighted images. In brief, the brain was extracted from the images, and we carefully checked the brain extraction for missing brain tissue and areas of non-brain tissue, and adjusted the image accordingly. We corrected RF inhomogeneities by bias field correction with a Markov random field model and subsequently segmented the brain in grey matter, white matter, and cerebrospinal fluid images [182]. A study specific grey matter template was created in standard space using a balanced set of subjects, and all grey matter segmentations were registered to this template with non-linear registration, and then corrected for any local expansion or contraction by modulation of the Jacobian warp field [148]. Last, an isotropic Gaussian kernel with a sigma of 3mm was applied for smoothing of the grey matter images. Total intracranial volume (TIV) was calculated as the sum of the volumes from grey matter, white matter and cerebrospinal fluid in standard space. The structures from the MNI-atlas were used as grey matter ROIs. We extracted volumetric measurements from the ROIs by registering the structural MNI-atlas [149] to the grey matter images in standard space, and multiplying the grey matter density of the ROI with the total volume of the ROI, resulting in the grey matter volume within the ROI. Left and right regions were considered separately.

Diffusion tensor images were corrected for motion artefacts and eddy currents by alignment to the b=0 image, and subsequently, the tensor was fitted at each voxel to create fractional

anisotropy (FA) images. The FA images were processed with the tract-based spatial statistics (TBSS) pipeline as implemented in FSL [183]. Using non-linear registration, the images were aligned to the `FMRIB58_FA` template and then averaged into a mean FA image. The mean FA image was thresholded at 0.2 and thinned into a white matter skeleton. All individual FA images were projected onto this skeleton, resulting in skeletonized FA data for each participant. The probabilistic tracts from the Johns Hopkins University atlas [150] were applied as white matter ROIs to the skeleton mask, and the masked ROIs were used to extract FA values from the individual tracts. Left and right tracts were considered separately.

| | Rotterdam 1 | Rotterdam 2 | Brescia | Barcelona |
|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| N (s/p/nc) | 3/22/24 | 5/9/6 | 7/17/0 | 1/6/1 |
| Scanner | Philips Achieva 3T | Philips Achieva 3T | Siemens Skyra | Siemens Trio Tim |
| Head Coil | 8 channel SENSE | 32 channel SENSE | 32 channel | 64 channel |
| T1 weighted imaging | | | | |
| TR | 9.8 ms | 6.8 ms | 2000 ms | 2000 ms |
| TE | 4.6 ms | 3.1 ms | 2.9 ms | 2.9 ms |
| FOV | 224 × 168 mm | 256 × 256 mm | 282 × 282 mm | 282 × 282 mm |
| Voxel size | 0.88 × 0.88 × 1.2 mm | 1.1 mm ³ | 1.1 mm ³ | 1.1mm ³ |
| Flip angle | 8 | 8 | 8 | 8 |
| Slices | 140 | 207 | 208 | 208 |
| Diffusion tensor imaging | | | | |
| TR | 8250 ms | 7000 ms | 7300 ms | 7300 ms |
| TE | 80 ms | 69 ms | 90 ms | 90 ms |
| FOV | 256 × 256 mm | 240 × 240 mm | 240 × 240 mm | 240 × 240 mm |
| Voxel size | 2 × 2 × 2 mm | 2.5 × 2.5 × 2.5 mm | 2.5 × 2.5 × 2.5 mm | 2.5 × 2.5 × 2.5 mm |
| Slices | 70 | 59 | 59 | 59 |
| Directions | 60 | 68 | 68 | 68 |
| B-values | 0/1000 s/mm ² | 0/1000 s/mm ² | 0/1000 s/mm ² | 0/1000 s/mm ² |

Table 4.ST1: MRI acquisition protocols. Numbers are subjects included after quality check. Abbreviations: s = symptomatic, p = presymptomatic, nc = non-carrier, TR = repetition time, TE = echo time, FOV = field of view.

Cognitive assessment

The following cognitive tests were performed, depending on the protocol from the local site. For language, the Boston Naming Task [184] and semantic fluency (animals) [185] were used.

Tests concerning attention and processing speed were the Trail making test part A [186], Stroop part 1 and 2 [187], symbol substitution [188], letter digit substitution task [189], and forward digit span [188]. For executive functioning, we used Trail making test part B [186], Stroop task part 3 [187], phonological fluency [185] and digit span backwards [188]. Tests for social cognition were the Ekman faces test [190], emotion recognition from the mini social cognition and emotional assessment (MINI-SEA) [191], and Happé cartoon task [192]. Raw scores from tests in which a higher score indicates worse performance were reversed (i.e. Trail making test, Stroop). We transformed all raw test scores to z-scores, based on the mean and standard deviation of the non-carriers. Subsequently, cognitive domains were composed as the mean z-score of all available tests within that domain per individual, disregarding missing tests.

Biomarker statistics

Before modelling, we checked skewed distributions in the biomarkers with the following graphs and tests: histograms, q-q plots, skewness and kurtosis values (values between 2 and -2 indicate normality), Kolmogorov-Smirnov and Shapiro-Wilk's tests (values above 0.05 indicate normality). When three or more tests indicated skewness, the distributions were adjusted using log-transformations (\log_{10}), i.e. neurofilament light chain levels, MMSE, BNT, Trail Making Test, Stroop, facial emotion recognition. In the case of cognitive tests, log-transformation was performed before transforming raw scores to z-scores. Biomarker characteristics and statistical differences between groups are presented in Table 4A.3. Symptomatic mutation carriers had higher NfL levels, lower grey matter volumes, impaired white matter microstructure, and worse cognitive functions than both presymptomatic mutation carriers and non-carriers in all selected biomarkers. Post-hoc analysis revealed that these differences in biomarkers were specifically driven by the bvFTD patients. For nvfPPA patients, we found higher NfL levels and worse cognitive performance than both presymptomatic mutation carriers and non-carriers. NfvPPA patients showed smaller grey matter volumes than both presymptomatic mutation carriers and non-carriers, especially in left-sided ROIs, and lower fractional anisotropy levels in the left anterior thalamic radiation, left uncinate fasciculus, and the forceps minor. The volume of the right frontal lobe was smaller in nvfPPA patients compared with presymptomatic mutation carriers. Furthermore, bvFTD patients had smaller volumes of the right frontal and temporal lobe than nvfPPA patients, and lower fractional anisotropy values in the forceps minor, left superior longitudinal fasciculus and right uncinate fasciculus. There were no differences in any of the selected biomarkers between presymptomatic mutation carriers and non-carriers.

| | Total (n=35) ^A | | | | Symptomatic | | | | Presymptomatic | | | |
|------------------------------|--|-----------|--------------|--------------|--------------|--------------|--------------|---------------|----------------|-------------|---|-----------|
| | N | Mean ± SD | N | Mean ± SD | bvFTD (n=17) | N | Mean ± SD | nfvPPA (n=16) | N | Mean ± SD | N | Mean ± SD |
| Language | MMSE | 29 | -3.07 ± 1.50 | 15 | -3.26 ± 1.69 | 14 | -2.87 ± 1.28 | 55 | 0.22 ± 1.01 | | | |
| | Boston naming test | 32 | -2.86 ± 1.37 | 16 | -2.69 ± 1.58 | 14 | -3.00 ± 1.24 | 55 | 0.28 ± 1.09 | | | |
| | Semantic fluency | 25 | -1.97 ± 1.32 | 13 | -1.75 ± 1.44 | 12 | -2.21 ± 1.19 | 55 | 0.57 ± 1.37 | | | |
| | Attention, concentration and mental processing speed | 31 | -3.28 ± 1.38 | 15 | -3.14 ± 1.57 | 14 | -3.45 ± 1.30 | 55 | 0.00 ± 1.31 | | | |
| | | 33 | -2.35 ± 1.17 | 16 | -2.43 ± 1.32 | 15 | -2.26 ± 1.12 | 55 | -0.05 ± 0.75 | | | |
| Executive functioning | TMT-A | 32 | -2.65 ± 1.62 | 16 | -2.91 ± 1.63 | 14 | -2.23 ± 1.66 | 55 | -0.01 ± 0.92 | | | |
| | Stroop card 1&2 | 17 | -3.05 ± 2.18 | 10 | -2.96 ± 2.20 | 7 | -3.18 ± 2.32 | 55 | 0.14 ± 1.02 | | | |
| | LDST | 4 | -2.06 ± 1.59 | 2 | -2.09 ± 1.99 | 2 | -2.03 ± 1.90 | 17 | 0.22 ± 0.70 | | | |
| | Symbol substitution | 17 | -2.35 ± 1.41 | 7 | -3.01 ± 1.13 | 10 | -1.89 ± 1.45 | 22 | 0.00 ± 1.29 | | | |
| | Digit span forward | 31 | -1.66 ± 1.05 | 15 | -1.41 ± 1.29 | 14 | -1.95 ± 0.74 | 55 | -0.26 ± 0.95 | | | |
| Social cognition | Executive functioning | 32 | -2.33 ± 0.97 | 15 | -2.23 ± 1.19 | 15 | -2.37 ± 0.79 | 55 | -0.03 ± 0.75 | | | |
| | | TMT-B | 28 | -2.63 ± 0.97 | 13 | -2.50 ± 1.11 | 13 | -2.69 ± 0.89 | 55 | 0.03 ± 0.79 | | |
| | Stroop card 3 | 14 | -3.84 ± 2.29 | 8 | -3.73 ± 2.50 | 6 | -3.98 ± 2.20 | 55 | -0.36 ± 1.05 | | | |
| | Phonological fluency | 29 | -2.12 ± 0.95 | 14 | -1.92 ± 1.02 | 15 | -2.30 ± 0.87 | 55 | 0.30 ± 1.36 | | | |
| | Digit span backwards | 30 | -1.65 ± 1.14 | 15 | -1.61 ± 1.44 | 13 | -1.61 ± 0.75 | 55 | -0.08 ± 1.11 | | | |
| Mini-SEA Emotion Recognition | Social cognition | 15 | -1.87 ± 0.76 | 7 | -2.15 ± 0.92 | 8 | -1.62 ± 0.52 | 51 | -0.10 ± 1.02 | | | |
| | Ekman faces | 3 | -0.70 ± 0.60 | 2 | -0.37 ± 0.18 | 1 | -1.36 ± N/A | 26 | 0.14 ± 0.89 | | | |
| | HappéTOM | 10 | -1.98 ± 0.83 | 3 | -2.65 ± 1.13 | 7 | -1.69 ± 0.52 | 22 | -0.62 ± 0.98 | | | |
| | Happé non TOM | 5 | -2.07 ± 0.86 | 4 | -2.32 ± 0.75 | 1 | -1.05 ± N/A | 28 | 0.42 ± 0.77 | | | |
| | | 5 | -1.65 ± 0.81 | 4 | -1.81 ± 0.84 | 1 | -1.03 ± N/A | 28 | 0.34 ± 1.20 | | | |

Table 4.5T2: Availability and characteristics of cognitive data. Abbreviations: bvFTD = behavioural variant frontotemporal dementia, nfvPPA = non-fluent variant primary progressive aphasia, MMSE = mini mental state examination, TMT = trailmaking test, LDST = letter digit substitution task, mini-SEA = mini social cognition and emotional assessment, TOM = theory of mind. Values are mean z-scores ± standard deviation based on non-carriers, uncorrected for confounding factors. ^A The two remaining participants presented with cortico-basal degeneration.

| | Symptomatic | | | Presymptomatic |
|---|---------------------------|-----------------------------|---------------------------|----------------|
| | Total | bvFTD | nvPPA | |
| Neurofilament light chain | 1.90 ± 0.25 ^A | 1.89 ± 0.23 ^A | 1.91 ± 0.28 ^B | 1.10 ± 0.22 |
| Left frontal lobe (GM) | -2.75 ± 1.80 ^A | -3.42 ± 2.06 ^A | -2.46 ± 1.40 ^B | 0.30 ± 0.65 |
| Right frontal lobe (GM) | -1.72 ± 1.79 ^A | -2.76 ± 1.43 ^{A,C} | -0.93 ± 1.79 ^D | 0.30 ± 0.65 |
| Left insula (GM) | -2.32 ± 1.56 ^A | -2.45 ± 1.79 ^A | -2.35 ± 1.51 ^B | -0.32 ± 0.95 |
| Right insula (GM) | -1.02 ± 1.13 ^A | -1.47 ± 1.26 ^A | -0.74 ± 0.98 | -0.08 ± 0.84 |
| Left parietal lobe (GM) | -1.87 ± 1.11 ^A | -2.18 ± 1.39 ^A | -1.74 ± 0.84 ^B | -0.03 ± 1.02 |
| Right parietal lobe (GM) | -1.19 ± 2.00 ^A | -1.42 ± 2.08 ^A | -0.89 ± 2.15 | -0.06 ± 0.96 |
| Left temporal lobe (GM) | -2.97 ± 2.42 ^A | -3.21 ± 2.59 ^A | -2.98 ± 2.51 ^B | -0.19 ± 0.96 |
| Right temporal lobe (GM) | -1.14 ± 2.66 ^A | -2.22 ± 3.40 ^{A,C} | -0.12 ± 1.69 | -0.08 ± 0.94 |
| Left anterior thalamic radiation (FA) | -2.28 ± 1.34 ^A | -2.73 ± 1.60 ^A | -1.77 ± 0.98 ^B | -0.33 ± 0.95 |
| Right anterior thalamic radiation (FA) | -1.24 ± 1.23 ^A | -1.78 ± 1.51 ^A | -0.66 ± 0.66 | -0.27 ± 0.77 |
| Forceps Minor (FA) | -3.00 ± 1.52 ^A | -4.01 ± 1.52 ^{A,C} | -2.08 ± 0.96 ^B | 0.46 ± 0.93 |
| Left superior longitudinal fasciculus (FA) | -1.50 ± 1.39 ^A | -2.42 ± 1.28 ^{A,C} | -0.61 ± 0.96 | 0.02 ± 0.88 |
| Right superior longitudinal fasciculus (FA) | -1.14 ± 1.12 ^A | -1.47 ± 1.14 ^A | -0.74 ± 1.06 | -0.11 ± 0.60 |
| Left uncinate fasciculus (FA) | -2.63 ± 1.15 ^A | -3.00 ± 1.43 ^A | -2.29 ± 0.88 ^B | -0.35 ± 0.86 |
| Right uncinate fasciculus (FA) | -1.92 ± 2.16 ^A | -3.19 ± 2.07 ^{A,C} | -0.77 ± 1.74 | -0.51 ± 1.12 |
| MMSE | -2.71 ± 1.19 ^A | -2.71 ± 1.28 ^A | -2.71 ± 1.14 ^B | 0.06 ± 0.91 |
| Attention and processing speed | -2.06 ± 1.09 ^A | -2.11 ± 1.15 ^A | -2.05 ± 1.12 ^B | -0.22 ± 0.65 |
| Executive functioning | -2.12 ± 0.88 ^A | -2.00 ± 0.99 ^A | -2.24 ± 0.82 ^B | -0.14 ± 0.72 |
| Social cognition | -1.89 ± 0.64 ^A | -2.13 ± 0.74 ^A | -1.52 ± 0.42 ^B | -0.19 ± 0.96 |

Table 4.ST3: Biomarker characteristics. Abbreviations: bvFTD = behavioural variant frontotemporal dementia, nvPPA = non-fluent variant primary progressive aphasia, GM volume = grey matter volume, FA = fractional anisotropy, MMSE = Mini Mental State Examination. Values are mean z-score (based on non-carriers) ± standard deviation, after correction for confounding factors of age, gender, and years of education. ^A Both the entire group of symptomatic mutation carriers and only bvFTD patients significantly differed from presymptomatic mutation carriers as well as non-carriers ($p < 0.05$, Bonferroni corrected). ^B Significant difference between nvPPA patients and presymptomatic mutation carriers as well as non-carriers ($p < 0.05$, Bonferroni corrected). ^C Significant difference between bvFTD patients and nvPPA patients ($p < 0.05$, Bonferroni corrected). ^D Significant difference between nvPPA patients and presymptomatic mutation carriers ($p < 0.05$, Bonferroni corrected).

Supplementary material 4B: DEBM

Gaussian mixture modelling

DEBM uses Gaussian mixture modelling to transform biomarker values to posterior probabilities of them being abnormal. This is done by assuming the probability density functions of normal and abnormal values are represented by Gaussians $\mathbb{N}(\mu_{\neg E}, \sigma_{\neg E})$ and $\mathbb{N}(\mu_E, \sigma_E)$ respectively, where the occurrence of the biomarker abnormality event is denoted by E and the absence of such an event is denoted by $\neg E$. Gaussian mixture modelling is an optimisation task to estimate these normal and abnormal Gaussians as well as the mixing parameter based on maximum log-likelihood, where the log-likelihood for biomarker B is computed as the summation over all *GRN* mutation carriers in the dataset as follows:

$$L_B = \sum_{\forall j \in \text{Carriers}} \log f(B_j)$$

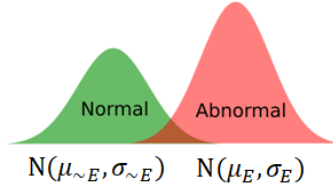


Figure 4.SF1: Illustrations of the Gaussian probability density functions for normal and abnormal values of biomarker B .

Here, the likelihood $f(B)$ is computed as follows:

$$f(B) = \theta_{\neg E} \times p(B|\mu_{\neg E}, \sigma_{\neg E}) + \theta_E \times p(B|\mu_E, \sigma_E)$$

Where $\theta_{\neg E} + \theta_E = 1$, and the mixing parameters $\theta_{\neg E}$ and θ_E show the relative proportions of the two Gaussians in the dataset. The abnormal Gaussian is initialized using the mean and standard deviation of the symptomatic subjects, while the normal Gaussian is initialized using the non-carriers. Since non-carriers are healthy controls, we fix $\mu_{\neg E}$ and $\sigma_{\neg E}$ to their initialized values and only optimize the remaining parameters in the Gaussian mixture model. The mixing parameter and the Gaussian parameters are optimized alternately until convergence as detailed previously [93].

For imaging-biomarkers with left and right counter parts, we propose a novel modification to the Gaussian mixture model optimization called Siamese Gaussian mixture model (Siamese GMM). We propose to jointly optimize the parameters of these biomarkers, by taking

advantage of symmetry in the brain. The log-likelihood for the joint optimization for the imaging biomarkers I^L and I^R is given below:

$$L_I = \sum_{\forall j \in \text{Carriers}} \log f(I_j^L) + \log f(I_j^R)$$

where $f(I_j^L)$ and $f(I_j^R)$ are expressed mathematically as:

$$f(I_j^L) = \theta_{\neg E}^L \times p(I_j^L | \mu_{\neg E}, \sigma_{\neg E}) + \theta_E^L \times p(I_j^L | \mu_E, \sigma_E)$$

$$f(I_j^R) = \theta_{\neg E}^R \times p(I_j^R | \mu_{\neg E}, \sigma_{\neg E}) + \theta_E^R \times p(I_j^R | \mu_E, \sigma_E)$$

$\theta_{\neg E}^L + \theta_E^L = 1$ and $\theta_{\neg E}^R + \theta_E^R = 1$. The mixing parameters ($\theta_{\neg E}^L, \theta_E^L, \theta_{\neg E}^R, \theta_E^R$) and the abnormal Gaussian parameters (μ_E, σ_E) are again optimized alternately until convergence [93]. This joint optimization of the left and right counter parts by sharing the normal and abnormal Gaussians reduces the number of parameters to be optimized, and thus improves the robustness. In case of asymmetrical atrophy patterns, where one of the biomarkers is stronger than the other, the joint optimization also helps in making the GMM more stable for the weaker biomarker.

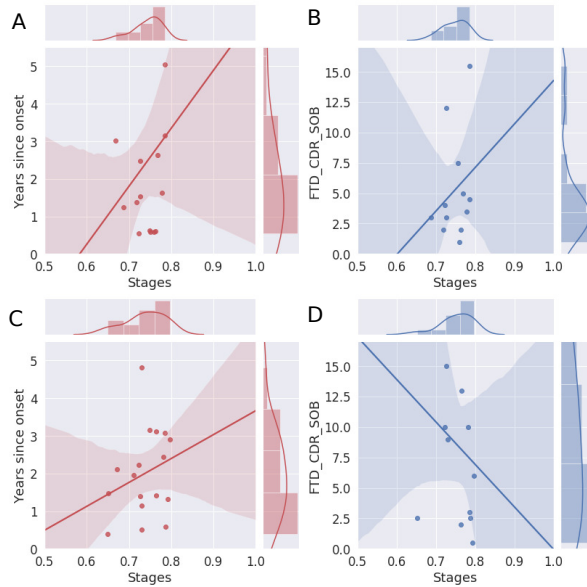


Figure 4.SF2: Correlation of disease severity (as estimated by non-imaging DEBM using cross-validation) with years since onset and FTD-CDR-SOB. The 2D scatter plots in figures A and C show the correlations of disease severity with years since onset, for symptomatic nfvPPA and bvFTD subjects respectively. The 2D scatter plot in figures B and D show the correlations of disease severity with FTD-CDR-SOB. The plot on top of each subfigure shows the probability density function of the disease stages. The plots on the right of figures A and C show the probability density functions of years since symptom onset. The plots on the right of figures B and D show the probability density function of FTD-CDR-SOB.

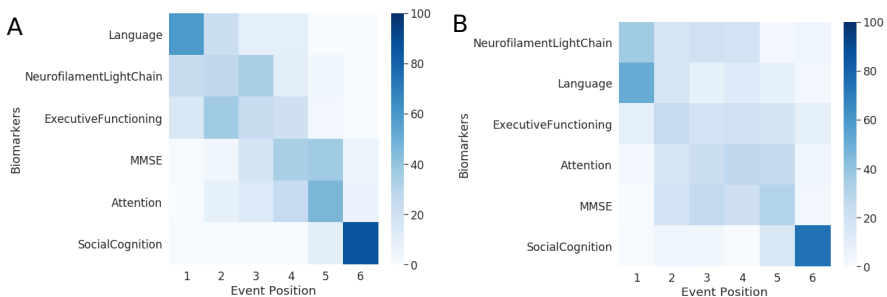


Figure 4.SF3: Cascade of non-imaging biomarker changes in nfvPPA (A) and bvFTD (B) subjects along with the uncertainty associated with it. The biomarkers are ordered based on the position in the estimated cascade. The color-map is based on the number of times a biomarker is at a position in 100 repetitions of bootstrapping.

Chapter 5

The sequence of structural, functional and cognitive changes in multiple sclerosis

This chapter contains the content of the manuscript ‘The sequence of structural, functional and cognitive changes in multiple sclerosis. Iris Dekker, Menno M. Schoonheim, **Vikram Venkatraghavan**, Anand J.C. Eijlers, Iman Brouwer, Esther E. Bron, Stefan Klein, Mike P. Wattjes, Alle M. Wink, Jeroen J.G. Geurts, Bernard M.J. Uitdehaag, Neil P. Oxtoby, Daniel C. Alexander, Hugo Vrenken, Joep Killestein, Frederik Barkhof, Viktor Wottschel. *NeuroImage Clinical*, Accepted, 2020’. An earlier version of this chapter was also presented (oral) at the Computational approaches for ageing and age-related diseases (CompAge) workshop 2020, and presented (poster) at the European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS) conference 2019.

Vikram Venkatraghavan’s contributions in this chapter are: Software, writing: review & editing.

Abstract

Background: As disease progression remains poorly understood in multiple sclerosis (MS), we aim to investigate the sequence in which different disease milestones occur using a novel data-driven approach.

Methods: We analysed a cohort of 295 relapse-onset MS patients and 96 healthy controls, and considered 28 features, capturing information on T2-lesion load, regional brain and spinal cord volumes, resting-state functional centrality (“hubness”), microstructural tissue integrity of major white matter (WM) tracts and performance on multiple cognitive tests. We used a discriminative event-based model to estimate the sequence of biomarker abnormality in MS progression in general, as well as specific models for worsening physical disability and cognitive impairment.

Results: We demonstrated that grey matter (GM) atrophy of the cerebellum, thalamus, and changes in corticospinal tracts are early events in MS pathology, whereas other WM tracts as well as the cognitive domains of working memory, attention, and executive function are consistently late events. The models for disability and cognition show early functional changes of the default-mode network and earlier changes in spinal cord volume compared to the general MS population. Overall, GM atrophy seems crucial due to its early involvement in the disease course, whereas WM tract integrity appears to be affected relatively late despite the early onset of WM lesions.

Conclusion: Data-driven modelling revealed the relative occurrence of both imaging and non-imaging events as MS progresses, providing insights into disease propagation mechanisms, and allowing fine-grained staging of patients for monitoring purposes.

5.1 Introduction

Multiple sclerosis (MS) is a chronic inflammatory, demyelinating and neurodegenerative disease of the central nervous system (CNS) [16] frequently leading to physical disability and cognitive decline [193]. The underlying pathological processes result in tissue damage, leaving behind demyelinating lesions and white (WM) and grey matter (GM) atrophy that can be visualised and quantified by brain and spinal cord imaging [18]. Alterations in structural and functional networks of the brain also have clear clinical relevance [194]. Usually considered in isolation, various studies have considered these features of MS. However, the sequence in which these changes occur remains unclear, in part due to scarcity of longitudinal data.

Event-based modelling (EBM) is a probabilistic data-driven approach to study disease progression that uses cross-sectional data to estimate the temporal sequence of events and subsequently stage patients within this sequence [53, 61]. This type of model has been applied in Alzheimer’s disease [53, 57, 61], Huntington’s disease [53, 95] and a recent EBM study in MS patients provided insights into the sequence of GM atrophy, but did not include features derived from other modalities [94].

In the present study we go beyond the aspect of atrophy in MS and consider a broader set of structural, functional, and cognitive outcomes. We explored measures quantifying demyelination (focal WM lesions) [193], neurodegeneration (GM atrophy) [195], microstructural changes of WM tracts (fractional anisotropy) [196], and functional centrality of key brain networks [194, 197] using a discriminative EBM (dEBM), which is more accurate and computationally efficient than the original EBM implementation [93]. The imaging biomarkers were supplemented with measures of cognitive performance [198]. Our multimodal dEBM could improve the interpretation of studies using single biomarkers, provide useful insights into disease propagation mechanisms, and aid in fine-grained staging and precise monitoring of patients. Therefore, the primary aim was to build a model that reflects a sequence of events in disease evolution in MS patients with a relapse onset. The secondary aim was to explore the event sequence for patients in relation to worsening physical and cognitive burden separately, because underlying disease processes could be different.

5.2 Methods

5.2.1 Participants

In this retrospective analysis study, we included data from the Amsterdam MS cohort based on the availability of multimodal data, resulting in the inclusion of 96 healthy controls (HC) and 295 patients with relapse-onset MS (ROMS) according to the 2011 revisions of the McDonald criteria [199]. Patients with a primary progressive disease onset have been excluded.

The institutional ethics review board of the VU University Medical Center approved the protocol and written informed consent was obtained from all participants prior to inclusion.

5.2.2 Clinical assessments

The Expanded Disability Status Scale (EDSS) score [200] was assessed in all patients and was used to classify patients into three groups according to having minimal (EDSS 0.0 – 2.5), moderate (EDSS 3.0 – 3.5) or severe disability (EDSS ≥ 4.0) as defined in [200]. Cognitive performance was assessed in all patients and HCs using an expanded Brief Repeatable Battery of Neuropsychological tests [201] with different cognitive domains tested, as described previously [198]. Raw test scores were corrected for the confounding effects of sex, age and education trends seen in the HCs [202]. Cognitive domain-specific z-scores were calculated using the mean and standard deviation (SD) of the HCs. Patients were sub-divided into three cognitive-performance groups according to the z-scores obtained from the neuropsychological tests. Patients with $z \leq -2$ on at least 2 out of 7 cognitive domains of the neuropsychological tests were labelled as cognitively impaired (CI), patients with $z \leq -1.5$ on at least 2 cognitive

domains but not fulfilling CI criteria were classified as mildly cognitively impaired (MCI) and the remaining patients were classified as cognitively preserved (CP) [203]. Level of education was measured using a scale ranging from 1 (unfinished primary school) to 7 (a university degree or higher) [204].

5.2.3 Magnetic Resonance Imaging

A 3 Tesla whole-body MR system was used to scan all participants (GE Signa HDxt, Milwaukee, WI) using an 8-channel phased-array head coil. The scan protocol included a 3D T1-weighted fast spoiled gradient-echo sequence for volume measures (TR: 7.8 ms, TE: 3 ms, $240 \times 240 \text{ mm}^2$ field of view (FOV), 176 sagittal slices of 1 mm thickness, $0.94 \times 0.94 \text{ mm}^2$ in-plane resolution), a 3D fluid-attenuated inversion recovery (FLAIR) sequence for lesion detection (TR: 8000 ms, TE: 125 ms, TI: 2350 ms, $250 \times 250 \text{ mm}^2$ FOV, 132 sagittal slices of 1.2 mm thickness, $0.98 \times 0.98 \text{ mm}^2$ in-plane resolution), a diffusion-weighted imaging (DWI) sequence to detect microstructural changes in WM tracts (TR: 13 s, TE: 86 ms, 2.4 mm contiguous axial slices, $2.0 \times 2.0 \text{ mm}^2$ in-plane resolution, 30 volumes with b-value = 900 s/mm^2 , 5 volumes with b-value = 0 s/mm^2), and a whole-brain resting-state fMRI sequence to measure eigenvector centrality (functional centrality; (200 volumes, TR: 2200 ms, TE: 35 ms, 3 mm contiguous axial slices covering the entire brain, $3.3 \times 3.3 \text{ mm}^2$ in-plane resolution)). FLAIR images were generally only acquired for patient, not HCs. More details on the protocol can be found in a previous report on this cohort [205].

FLAIR images were used to segment WM lesions in MS patients using a k-Nearest-Neighbours approach with tissue type priors (kNN-TTP) [206]. Lesion maps were registered to 3D T1-weighted images and filled using a validated patch-based approach [207].

Brain parcellation of cortical and subcortical regions was obtained using geodesic information flows (GIF) [208] on the 3D T1-weighted MRI scans, a method that has been used previously in applications of MS [209, 210], including a predecessor study on EBM-based atrophy progression [94], and other neurological disorders [211], as well as a pre-processing tool for segmenting WM hyperintensities [212]. GIF is an atlas-propagation-based method that registers T1 scans of 160 subjects with manually delineated brain structures to each target scan, then identifies the closest local matches and uses those matches for segmentation. The atlas segmentations are based on the Desikan-Killiany-Tourville protocol, which was designed to improve accuracy and consistency of brain labels compared to the classic Desikan-Killiany atlas database [213]. To further quantify regional lesion loads, the white matter was initially divided into 10 concentric bands between the ependyma of the ventricles and the pial surface based on normalized subject-specific distance maps derived from Laplace equation isolines [210, 214]. The bands were then grouped as inner (band 1-2), intermediate deep (band 3-8), and outer bands (band 9-10) in order to obtain a data-driven approximation of the

stratification used in (pre-)clinic. Infratentorial lesions were subsequently discarded because they were only present in a small subset of patients.

Spinal cord atrophy was quantified as mean upper cervical cord area (MUCCA) using SCT-PropSeg [215]. Analyses were performed on the 3D T1-weighted images of the brain, which cover a sufficient length of the cervical spinal cord. We measured over a length of 30mm along the central canal, starting at the top of the second cervical vertebra, C2. MUCCA measurements on brain images have been shown to be as reproducible as those performed on dedicated spinal cord MRI [216, 217].

Functional MRI processing steps for obtaining eigenvector centrality maps (ECM) have been published previously [218]. The MELODIC pipeline (part of FSL [181], using standard settings) was used to process resting-state fMRI images, followed by nonlinear registration to Montreal Neurological Institute standard space, and resampling to a resolution of 4 mm isotropic. The MELODIC outcomes were further processed using fastECM [219] to estimate voxel-wise eigenvector centrality as a network measure of functional hubness (brain function) in the default-mode network (DMN), basal ganglia and sensorimotor network.

DWI scans were pre-processed using FSL5, including motion- and eddy current correction on images and gradient vectors, followed by diffusion tensor fitting for diffusion tensor imaging (DTI). The resulting fractional anisotropy (FA) maps were then fed into the tract-based spatial statistics (TBSS) pipeline [220], after which the skeleton was masked using the JHU white-matter tractography atlas from FSL to define WM tracts.

There was only a minor amount of motion artefacts present in the advances imaging sequences, and we did not observe any difference in artefact severity between groups.

5.2.4 Discriminative event-based model

The EBM uses cross-sectional data to estimate the ordered sequence of cumulative abnormality in a disease, together with uncertainty in the ordering. Here, we used the discriminative EBM (dEBM; <https://github.com/EuroPOND/pyebm>) described previously as it has been shown to be more accurate and computationally efficient compared to other EBM implementations [70, 93]. The dEBM estimates the probability for each biomarker being normal or abnormal using a Gaussian mixture model (GMM) based on data from a disease and a reference population. Based on the probability distributions of the biomarkers in the two groups, an individual sequence of biomarker abnormality is calculated for each patient. Finally, these individual sequences are combined statistically to give an ordering for the whole population [93]. The uncertainty of this ordering is estimated by bootstrapping, i.e. repeating the experiment with random subsets of subjects. Subjects can

be staged within the event sequence by identifying the events that have already become abnormal for each individual subject [61].

5.2.5 Selected biomarkers

We included multimodal biomarkers with relevance in MS whilst limiting the overall number of features in the model to allow for better interpretability of results and faster computation. The following 28 MS-related biomarkers were considered (before statistical post-selection as described below):

GM volumes of the thalamus, hippocampus, basal ganglia (without thalamus and hippocampus), cerebellar GM, cingulate, frontal lobe, insula, occipital lobe, parietal lobe and temporal lobe. These regions cover the entire brain to allow for a rough estimate of the general atrophy sequence.

MUCCA was included for all subjects as an indicator of spinal cord volume.

T2-hyperintense lesion loads on FLAIR images were considered only for patients and split according to the inner (i.e., periventricular), outer (i.e., juxtacortical) and intermediate deep WM bands in order to mimic (pre-)clinical use.

Functional centrality in the default mode network (DMN), sensorimotor cortex network and basal ganglia network; the voxelwise ECM-measures were averaged within the respective anatomical regions. The selected networks are linked to MS progression in the domains cognition [218], fatigue [221] and clinical recovery [222].

Microstructural changes of WM tracts measured by fractional anisotropy (FA) in 3 major WM tracts related to cognition (anterior thalamic radiation and cingulum [223, 224]) and motor function (corticospinal tract [225]) and all other WM tracts combined (forceps minor, inferior fronto-occipital fasciculus, inferior longitudinal fasciculus, superior longitudinal fasciculus, uncinate fasciculus and superior longitudinal fasciculus); the voxelwise FA-measures were averaged within the respective anatomical regions.

Cognitive function by cognitive domain: executive function, verbal memory, information processing, verbal fluency, visuospatial, working memory and attention.

5.2.6 Statistics

Normality of data was checked by visual inspection of histograms combined with Kolmogorov-Smirnov testing. Parametric (independent-samples t-test) and non-parametric (Mann-Whitney U test and chi-square test) tests were used to compare groups (Tables 5.1, 5.2

and 5.3) for demographic, clinical and imaging characteristics. All measures, except lesions, which could only be obtained in patients, were corrected for the confounding effects of age, sex and education seen in HCs using one linear regression model per biomarker. The residuals of these fits were then transformed into z-scores using the mean and SD from HCs.

We used SPSS 22.0 and 24.0 (IBM Corp., Armonk, NY, USA) and the scipy python package (version 1.2.1) for statistical analyses. The level of significance for demographic and clinical data (Tables 5.1, 5.2 and 5.3) was set at $p < 0.05$.

5.2.7 Model fitting

The dEBM relies on a Gaussian Mixture of the biomarker distributions, and requires a sufficient separation of the respective distributions from the control and disease groups. Therefore, we performed a biomarker post-selection and included only those biomarkers that passed a two-sided independent samples t-test at a significance level of $p/leq 0.1$. We used 1000 bootstraps sampled from the same cohort in order to estimate the positional variance of the event sequence. Individual subjects were finally staged within the model between stage 0 (no abnormal biomarkers) and stage N (all N biomarkers are abnormal).

Three dEBMs were built to characterize the structural, functional, and cognitive changes in ROMS progression generally (Model 1), and specifically for disability worsening (Model 2) and cognitive decline (Model 3).

- Model 1: Event sequence in all ROMS patients as a progression from HC.
- Model 2: Event sequence in ROMS patients progressing from low (EDSS 0.0 – 2.5) to high disability level (EDSS ≥ 4.0). Intermediate patients with an EDSS of 3.0 or 3.5 were excluded from the GMM initialisation but used to estimate the event sequence. HCs were excluded for this analysis.
- Model 3: Event sequence in MS patients progressing from cognitively preserved (CP) to cognitively impaired (CI). Patients with mild cognitive impairment (MCI) were excluded from the GMM initialisation but used to estimate the event sequence. HCs were excluded for this analysis.

5.3 Results

At the time of data acquisition, 243 of ROMS patients were diagnosed with relapsing-remitting MS (RRMS) and 52 patients with secondary progressive MS (SPMS). The average age was 46.7 (standard deviation 11.0) years and patients had their symptom onset 12.6 ± 1.6 years prior to assessment. The proportion of women was higher in the patient group (71.5%) than

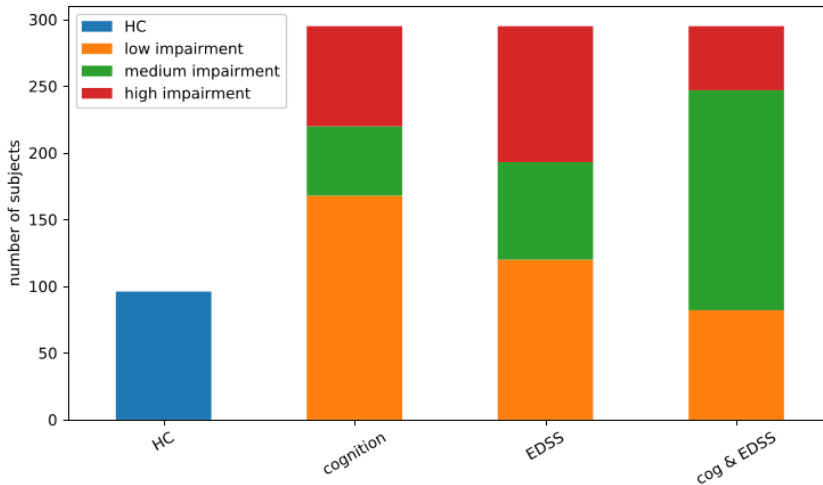


Figure 5.1: Overview of diagnostic groups, and separation of ROMS subgroups.

in HCs (58.3%, $p=0.016$) and HCs had a higher educational level ($p=0.017$). The median EDSS was 3.0 (IQR 2 – 4) with 120 patients having low disability (EDSS 0.0 – 2.5) and 102 patients having high disability (EDSS ≥ 4.0). Seventy-five patients were cognitively impaired (CI), 52 patients were classified as MCI and 168 patients as CP (Figure 5.1). Demographics and MRI metrics of patients and HCs are shown in Table 5.1. For the MRI measures, only two functional networks (DMN and basal ganglia) were not significantly different between patients and healthy controls after correction for confounding variables.

Biomarker post-selection resulted in 21, 25, and 17 biomarkers included in the final models 1 (general MS progression), 2 (disability in MS) and 3 (cognitive decline in MS), respectively.

We visualize the models using positional variance diagrams (PVD; see Figure 5.2, 5.4, 5.6). The positional variance diagram shows the most likely sequence of events on the y-axis, while the x-axis represents the event position within the sequence ranging from one to the number of events. The intensity of each field represents the number of bootstraps where an event appeared at that respective position. This indicates uncertainty in the sequence, such that a strong confidence in the ordering results in a dark diagonal in the positional variance diagram.

| Clinical / Imaging measures | Patients (n=295) | Healthy controls (n=96) | p-value |
|--|--------------------|-------------------------|---------------------|
| Clinical measures | | | |
| Age [years] * | 47.0 ± 10.7 | 45.9 ± 10.4 | 0.37 ^a |
| Sex [female, %] ** | 211 (71.5) | 56 (58.3) | 0.016 ^b |
| Education level *** | 5 (4 - 6) | 6 (4 - 7) | 0.017 ^c |
| Symptom duration [years] * | 12.6 ± 1.6 | N/A | - |
| DMT ever used [%]** | 173 (58.6) | N/A | - |
| EDSS *** | 3.0 (2.0 - 4.0) | N/A | - |
| CP/MCI/CI [%] ** | 25.4 / 17.6 / 56.9 | N/A | - |
| RRMS/SPMS | 243/52 | N/A | - |
| Information processing speed [z-score] * | -1.12 (1.4) | 0.0 (1.0) | <0.001 ^a |
| Executive functioning [z-score] * | -0.95 (1.7) | 0.0 (0.8) | <0.001 ^a |
| Working memory [z-score] * | -1.02 (1.5) | 0.0 (0.9) | <0.001 ^a |
| Verbal memory [z-score] * | -0.48 (1.2) | 0.0 (0.9) | <0.001 ^a |
| Verbal fluency [z-score] * | -0.44 (1.1) | 0.0 (1.0) | <0.001 ^a |
| Visuospatial memory [z-score] * | -0.61 (1.2) | 0.0 (0.9) | <0.001 ^a |
| Attention [z-score] * | -0.65 (1.1) | 0.0 (0.7) | <0.001 ^a |
| MRI measures | | | |
| <i>T2-hyperintense lesion loads [mL]*</i> | | | |
| Total T2-hyperintense lesion load | 14.2 (12.7) | N/A | - |
| Inner lesions | 4.3 (3.3) | N/A | - |
| Deep lesions | 6.5 (6.9) | N/A | - |
| Outer lesions | 3.4 (3.5) | N/A | - |
| Infratentorial | 0.02 (0.04) | N/A | - |
| <i>Brain and spinal cord volumes [mL] *</i> | | | |
| Total brain volume | 1135.3 (110.3) | 1181.8 (128.8) | <0.001 ^a |
| Basal Ganglia | 34.3 (3.5) | 36.7 (4.1) | <0.001 ^a |
| Hippocampus | 7.6 (0.7) | 8.0 (0.8) | <0.001 ^a |
| Thalamus | 10.1 (1.0) | 11.7 (1.4) | <0.001 ^a |
| Cingulate | 27.7 (3.3) | 29.0 (3.6) | 0.002 ^a |
| Frontal lobe | 179.92 (18.8) | 185.1 (20.8) | 0.224 ^a |
| Insula | 10.7 (1.2) | 11.3 (1.4) | <0.001 ^a |
| Occipital lobe | 66.6 (8.0) | 70.3 (8.2) | <0.001 ^a |
| Parietal lobe | 91.1 (10.2) | 95.0 (9.6) | 0.002 ^a |
| Temporal lobe | 128.0 (13.4) | 132.8 (15.0) | 0.006 ^a |
| Cerebellar grey matter | 94.6 (9.5) | 99.0 (10.1) | 0.003 ^a |
| MUCCA | 64.7 (7.8) | 68.6 (5.7) | <0.001 ^a |
| <i>Functional hubness (EC [z-scores])</i> | | | |
| Basal ganglia network | -0.128 (0.23) | 0 (0.26) | 0.371 ^a |
| Default mode network | 0.009 (0.21) | 0 (0.22) | 0.755 ^a |
| Sensorimotor cortex network | -0.046 (0.22) | 0 (0.22) | 0.234 ^a |
| <i>White matter tract integrity (FA [0-1])</i> | | | |
| Anterior thalamic radiation | 0.453 (0.035) | 0.479 (0.027) | <0.001 ^a |
| Corticospinal tract | 0.653 (0.028) | 0.668 (0.028) | <0.001 ^a |
| Cingulum | 0.564 (0.047) | 0.598 (0.041) | <0.001 ^a |
| Other WM tracts | 0.525 (0.035) | 0.561 (0.026) | <0.001 ^a |

Table 5.1: Clinical and imaging measures of patients and healthy controls. The p-values of the MRI measures are based on the z-scored comparisons. Biomarkers with a $p < 0.1$ were included in the model. * Mean (standard deviation), ** number (percentage), *** median (IQR). ^a Independent-samples t-test, ^b chi-square test, ^c Mann-Whitney U test.

5.3.1 Model 1: Sequence of events in relapse-onset multiple sclerosis progression.

The PVD of Model 1 is shown in Figure 5.2. Despite considerable uncertainty, ROMS tends to start with decreases in the corticospinal tract FA as well as cerebellar and thalamic atrophy. Neurodegeneration continues to involve the occipital and parietal lobes (position 5 and 7 of 21), through the temporal lobe, spinal cord (MUCCA) and basal ganglia (position 10, 12 and 13 of 21), with the cingulate and insula being affected later (position 16-17 of 21). Deficiency in visuospatial cognition is the earliest cognitive abnormality at position 4, shortly after thalamic atrophy, followed by verbal fluency, verbal memory and information processing (position 6, 9 and 11 of 21). Other cognitive domains are estimated to be affected later. FA changes of the cingulum and the non-specific WM-tracts appear in the last third of the event sequence between the basal ganglia and the cingulate volume events. Anterior thalamic radiation FA becomes abnormal late (position 18 of 21).

The staging reveals that healthy controls are mostly placed at earlier stages and no HC being staged higher than stage 11 of 21 (median stage 2, mean stage 3) while ROMS patients are spread across all stages with a median stage of 8 (mean 9.4) as shown in Figure 5.3.

The effect on leaving out individual biomarkers or groups of biomarkers from a certain modality is very small as shown qualitatively in the Supplementary Materials. The PVD for the main tracts of the JHU WM tractography atlas is shown in Figure 5.SF4.

5.3.2 Model 2: Sequence of events in the progression of low-to-high disability in ROMS.

Table 5.2 shows the comparison between patients with high disability (EDSS of 4.0 or higher, $n = 102$) and patients with low disability (EDSS of 2.5 or lower, $n = 120$). Patients with high disability were older (average 53.1 versus 41.0 years, $p < 0.001$), had longer symptom duration (average 18.8 versus 10.6 years, $p < 0.001$), had a lower level of education (5 versus 4, $p=0.001$), and a higher percentage of cognitive impairment (47.1% versus 13.3%, $p < 0.001$) than patients with low disability. Not all MRI measures showed significant differences ($p < 0.1$ was accepted in the biomarker post-selection) between patients with high versus low disability. The included markers are listed in Figure 5.4.

The sequence for progression from low to high disability is shown in Figure 5.4. Insular and cerebellar GM atrophy occur early in the event sequence together with changes in centrality of the default-mode and basal-ganglia networks, and visuospatial perception (position 1-5 of 25). Atrophy continues to occur in the thalamus, temporal lobe, MUCCA, parietal lobe, basal ganglia, while occipital and frontal lobe atrophy occur relatively late (position 19 and 21 of 25 respectively). Lesion load becomes abnormal first in the inner (periventricular) regions,

| Clinical / Imaging measures | Total (n=222) | EDSS ≤ 2.5 (n=120) | EDSS ≥ 4.0 (n=102) | p-value |
|--|--------------------|--------------------|--------------------|---------------------|
| Clinical measures | | | | |
| Age [years]* | 46.5 (10.6) | 41.0 (8.5) | 53.1 (8.9) | <0.001 ^a |
| Sex [female, %]** | 160 (72.1) | 87 (72.5) | 73 (71.6) | 0.88 ^b |
| Education level [median, IQR]*** | 5 (4-6) | 5 (4-6) | 4 (3-6) | 0.001 ^c |
| Symptom duration [years]* | 14.4 (8.4) | 10.6 (5.7) | 18.8 (8.8) | <0.001 ^a |
| DMT used** | 126 (56.8) | 67 (55.8) | 59 (57.8) | 0.76 ^b |
| EDSS*** | 2.5 (2.0 - 4.5) | 2.0 (1.5 - 2.5) | 5.0 (4.0 6.0) | <0.001 ^c |
| CP/MCI/CI [%]*** | 53.6 / 17.6 / 28.8 | 68.3 / 18.3 / 13.3 | 36.3 / 16.7 / 47.1 | <0.001 ^b |
| RRMS/SPMS | 176 / 46 | 119 / 1 | 57 / 45 | <0.001 ^b |
| Information processing speed [z-score] * | -1.18 (1.4) | -0.67 (1.2) | -1.80 (1.4) | <0.001 ^a |
| Executive functioning [z-score] * | -1.04 (1.8) | -0.50 (1.0) | -1.71 (2.4) | <0.001 ^a |
| Working memory [z-score] * | -1.03 (1.4) | -0.51 (1.0) | -1.69 (1.6) | <0.001 ^a |
| Verbal memory [z-score] * | -0.45 (1.1) | -0.25 (1.0) | -0.70 (1.2) | 0.002 ^a |
| Verbal fluency [z-score] * | -0.52 (1.1) | -0.22 (1.0) | -0.86 (1.0) | <0.001 ^a |
| Visuospatial memory [z-score] * | -0.63 (1.2) | -0.28 (1.1) | -1.04 (1.2) | <0.001 ^a |
| Attention [z-score] | * -0.66 (1.1) | -0.39 (0.9) | -0.98 (1.3) | <0.001 ^a |
| MRI measures | | | | |
| <i>T2-hyperintense lesion loads [mL]*</i> | | | | |
| Total T2-hyperintense lesion load | 14.7 (13.2) | 11.2 (9.0) | 18.9 (15.9) | <0.001 ^a |
| Inner lesions | 4.3 (3.3) | 3.6 (2.5) | 5.2 (3.8) | <0.001 ^a |
| Deep lesions | 6.9 (7.4) | 5.1 (4.9) | 9.1 (9.2) | <0.001 ^a |
| Outer lesions | 3.5 (3.4) | 2.6 (2.3) | 4.6 (4.1) | <0.001 ^a |
| Infratentorial | 0.01 (0.04) | 0.01 (0.03) | 0.02 (0.04) | 0.31 ^a |
| <i>Brain and spinal cord volumes [mL] *</i> | | | | |
| Total brain volume | 1133.0 (111.3) | 1153.1 (109.4) | 1109.4 (109.4) | 0.003 ^a |
| Basal Ganglia | 34.2 (3.6) | 34.9 (3.6) | 33.3 (3.5) | <0.001 ^a |
| Hippocampus | 7.6 (0.7) | 7.7 (0.8) | 7.4 (0.7) | 0.011 ^a |
| Thalamus | 10.1 (1.5) | 10.6 (1.3) | 9.4 (1.3) | <0.001 ^a |
| Cingulate | 27.7 (3.3) | 28.1 (3.2) | 27.2 (3.4) | 0.04 ^a |
| Frontal lobe | 180.1 (19.0) | 185.4 (19.6) | 173.9 (16.2) | <0.001 ^a |
| Insula | 10.7 (1.3) | 11.0 (1.3) | 10.3 (1.2) | <0.001 ^a |
| Occipital lobe | 66.3 (8.0) | 68.3 (7.8) | 64.0 (7.7) | <0.001 ^a |
| Parietal lobe | 91.1 (10.2) | 94.0 (10.2) | 87.8 (9.3) | <0.001 ^a |
| Temporal lobe | 127.7 (13.5) | 130.7 (13.9) | 124.1 (12.1) | <0.001 ^a |
| Cerebellar grey matter | 94.3 (9.7) | 96.6 (9.3) | 91.6 (9.5) | <0.001 ^a |
| MUCCA | 64.5 (8.3) | 65.6 (8.3) | 63.2 (8.2) | 0.07 ^a |
| <i>Functional hubness (EC [z-scores])</i> | | | | |
| Basal ganglia network | -0.016 (0.24) | -0.067 (0.235) | 0.044 (0.224) | <0.001 ^a |
| Default mode network | -0.004 (0.21) | -0.035 (0.221) | 0.033 (0.200) | 0.017 ^a |
| Sensorimotor cortex network | -0.039 (0.22) | -0.021 (0.216) | -0.058 (0.216) | 0.206 ^a |
| <i>White matter tract integrity (FA [0-1])</i> | | | | |
| Anterior thalamic radiation | 0.45 (0.03) | 0.47 (0.02) | 0.44 (0.04) | <0.001 ^a |
| Corticospinal tract | 0.65 (0.03) | 0.66 (0.02) | 0.64 (0.03) | <0.001 ^a |
| Cingulum | 0.56 (0.05) | 0.58 (0.04) | 0.55 (0.05) | <0.001 ^a |
| Other WM tracts | 0.53 (0.04) | 0.54 (0.03) | 0.51 (0.04) | <0.001 ^a |

Table 5.2: Demographics of high vs low EDSS. Biomarkers with a $p < 0.1$ were included in the model.

* Mean (standard deviation), ** number (percentage), *** median (IQR). ^a Independent-samples t-test, ^b chi-square test, ^c Mann-Whitney U test.

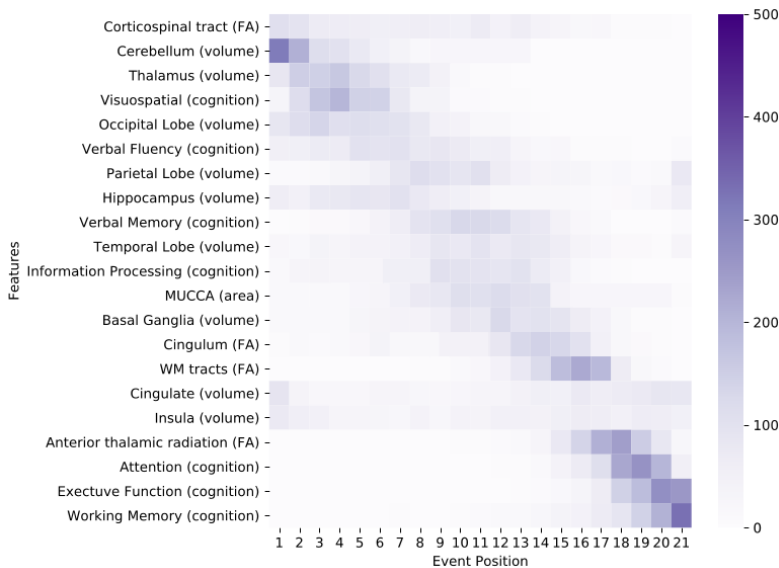


Figure 5.2: Positional variance diagram for the general ROMS population (Model 1).

The maximum-likelihood sequence of abnormality is shown on the y-axis (top to bottom). Colour intensity in each row indicates positional variance: the darker the colour, the higher the confidence of the event position across 1000 bootstraps (capped at 500 for visualisation). The biomarker ordering reflects the sequence obtained from fitting all subjects.

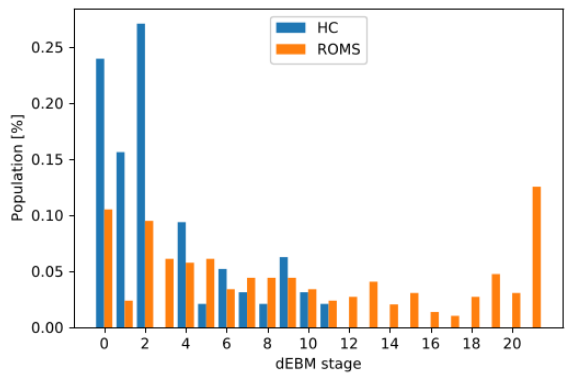
EC = eigenvector centrality; EDSS: expanded disability status scale; FA: fractional anisotropy as a measure for microstructural WM tract changes; MUCCA: mean upper cervical cord area.

then in the deep WM and the outer regions (i.e. juxtacortical). Changes in the FA biomarkers appear in the last third of the sequence, and cognitive changes in attention, working memory and executive function are last to become abnormal.

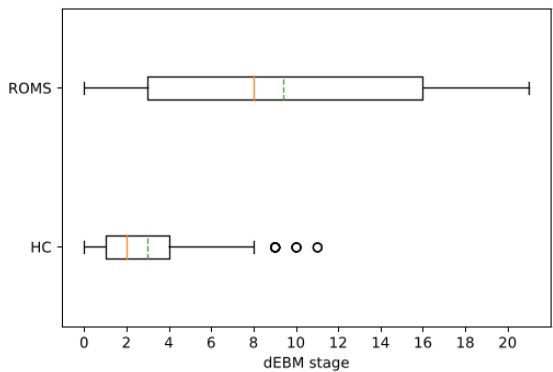
The patient staging shows that ROMS patients with all levels of disability can be found in all 25 stages (see Figure 5.5). However, there is a clear trend such that patients with low EDSS have a median stage of 4 (mean 6.9), patients with medium disability have a median stage of 6 (mean 9.3), and patients with a high level of disability have a median stage of 16 (mean 14.6).

5.3.3 Model 3: Sequence of events in ROMS as cognition declines.

All 295 patients had complete cognitive tests: 75 patients were classified as CI, 52 as MCI, and 168 patients as CP. Patients with CI were older (average 50.4 versus 45.7 years; $p = 0.001$), had a longer symptom duration (average 17.6 versus 13.3 years, $p < 0.001$), had a lower educational level (4 versus 6, $p < 0.001$) and a higher EDSS score (median 4.0 versus 3.0; $p < 0.001$), see Table 5.3 for the comparisons between CI and CP patients.



(a)



(b)

Figure 5.3: Patient staging for Model 1 (ROMS). (a) Staging of HC and ROMS subjects within the 21 disease stages. (b) Boxplot of staging indicating median (solid orange line) and mean (dashed green line) of the groups.

The ordering of events in the dEBM of cognitive impairment is shown in Figure 5.6. Similar to Model 2, the progression in cognitive decline is accompanied by early insular atrophy and increased functional DMN centrality. The event sequence continues with atrophy of the hippocampus, cervical cord, frontal, parietal, occipital and temporal lobes, and the thalamus (position 3-9 of 17) and finally the basal ganglia (position 14 of 17). Lesion events occur in close succession after most atrophy measures (position 10-12 of 17). Changes in WM tract FA occur at the end with the corticospinal tract being affected earlier than the rest (position 13 of 17).

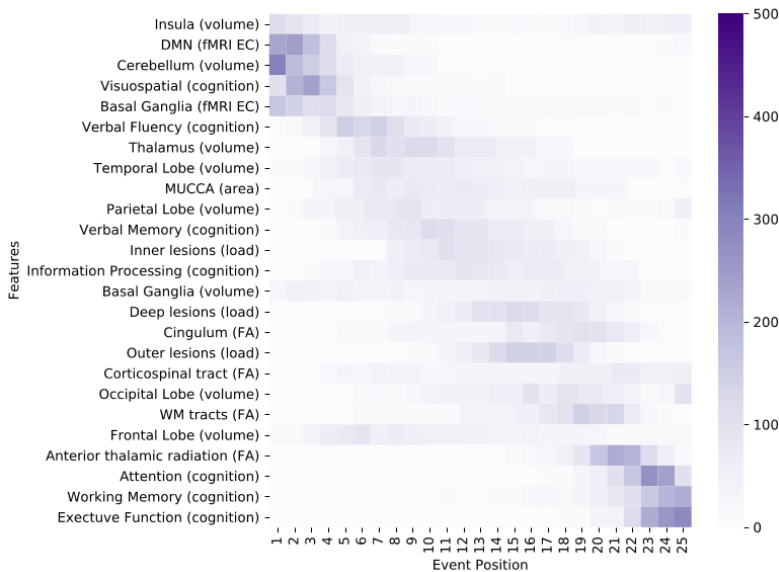
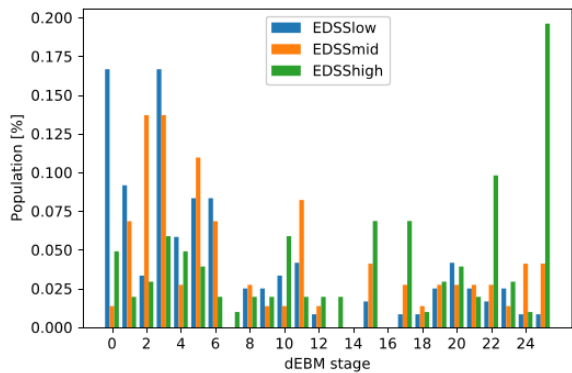


Figure 5.4: Positional variance diagram for the progression from low to high disability in ROMS patients (Model 2). The maximum-likelihood sequence of abnormality is shown on the y-axis (top to bottom). Colour intensity in each row indicates positional variance: the darker the colour, the higher the confidence of the event position across 1000 bootstraps (capped at 500 for visualisation). The biomarker ordering reflects the sequence obtained from fitting all subjects. EC = eigenvector centrality; EDSS: expanded disability status scale; FA: fractional anisotropy as a measure for microstructural WM tract changes; MUCCA: mean upper cervical cord area.

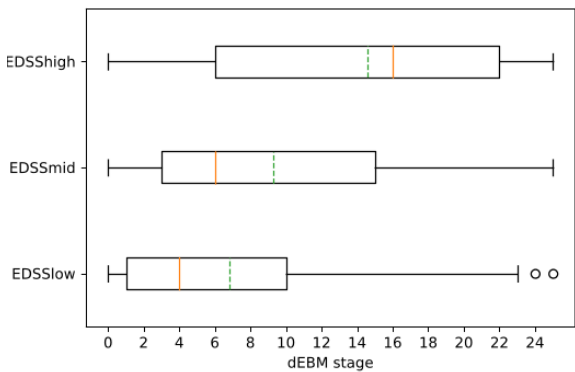
As in Model 2, all three groups are spread across all stages (see Figure 5.7). Cognitively preserved ROMS patients have a median stage of 5 (mean 5.3), patients with MCI have a median stage of 7.5 (mean 7.8), and cognitively impaired patients have a median stage of 12 (mean 11.4).

5.3.4 Discussion

Current understanding of disease progression in MS is largely based on studies that each considered a small number of MS pathology features in isolation. This body of work has identified lesion number and location, regional atrophy, changes in functional centrality of brain networks, or alterations in WM tract microstructure as features of interest. Until now, the sequence of accumulated abnormality in these biomarkers relative to each other remained largely undetermined. Our data-driven dEBM analysis suggests that changes of the corticospinal tract, and GM volume changes of cerebellum, thalamus and occipital lobe are early events; whereas microstructural changes in other WM tracts and changes in cognitive domains attention, executive function and working memory are relatively late events in MS progression. We also estimated sequences specific to disability worsening and cognitive



(a)



(b)

Figure 5.5: Patient staging for Model 2 (disability). (a) Staging of subjects with different levels of disability within the 25 disease stages. (b) Boxplot of staging indicating median (solid orange line) and mean (dashed green line) of the groups.

impairment motivated to reveal new insight into the underlying mechanisms of each, and to provide a quantitative template for patient assessment. The results of this secondary analysis suggest that functional network centrality of the default mode network is involved early in both, with DTI-related WM tract abnormality occurring later.

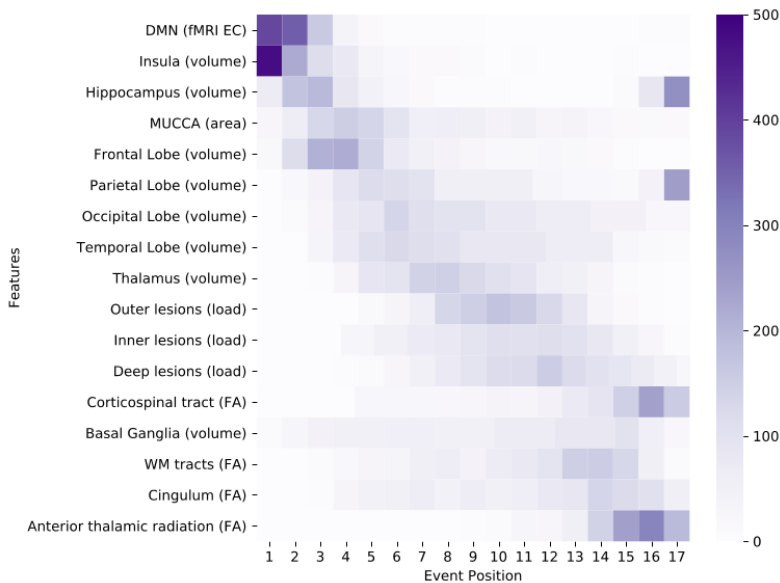


Figure 5.6: Positional variance diagram for the progression in ROMS patients as cognition declines (Model 3). The maximum-likelihood sequence of abnormality is shown on the y-axis (top to bottom). Colour intensity in each row indicates positional variance: the darker the colour, the higher the confidence of the event position across 1000 bootstraps (capped at 500 for visualisation). The biomarker ordering reflects the sequence obtained from fitting all subjects.

EC = eigenvector centrality; EDSS: expanded disability status scale; FA: fractional anisotropy as a measure for microstructural WM tract changes; MUCCA: mean upper cervical cord area.

5.3.5 Model 1: Sequence of events in relapse-onset multiple sclerosis progression.

The general ROMS model suggests that cerebellar atrophy is an early event. Although studies in patients with a clinically isolated syndrome (CIS) have not been conclusive on the presence of early cerebellar volume loss [226], a recent EBM study in MS patients also showed cerebellar atrophy as part of the atrophy sequence [94]. Early thalamic and hippocampal atrophy in our study are in accordance with findings in previous studies reporting atrophy in these areas already in CIS patients [227, 228]. Insular and cingulate atrophy occur relatively late in our study but the bootstrap analysis shows a bimodal distribution for these biomarkers with clusters at the beginning and the end of the sequence (see Figure 5.2), which might indicate heterogeneity in the population such that some patients have the event early whereas others experience this later. Among the volumetric measurements, MUCCA abnormality occurs at an intermediate position in the event ordering, while previous literature indicates that spinal cord atrophy can be seen already in CIS patients on a group level and with high clinical relevance [229]. This could be due to differences in measurement sensitivity or cohort size and requires further study.

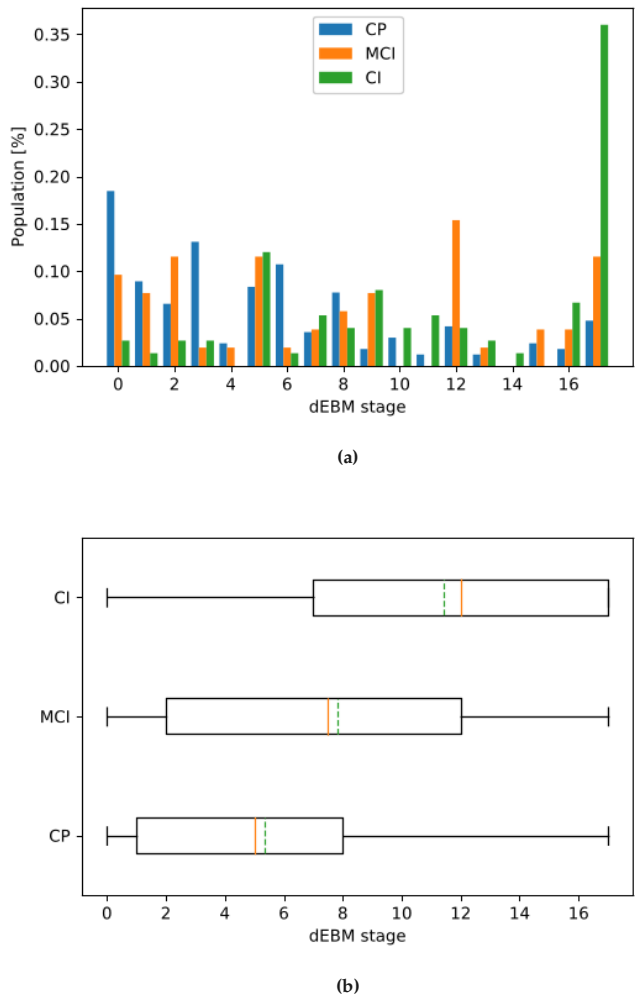


Figure 5.7: Patient staging for Model 3 (cognition). Top: Staging of subjects with different levels of cognitive abilities within the 17 disease stages. Bottom: Boxplot of staging indicating median (solid orange line) and mean (dashed green line) of the groups.

A second marker of white matter abnormality, the FA of the corticospinal tract, appears as the first event, which agrees with previous work [229]. In addition, research found that WM changes occur early in the disease [196], whereas most FA markers other than corticospinal tract damage included in this study were late events in the model. A probable explanation could be that the WM damage from lesions within the tracts is relatively small compared to overall sizes of the tract ROIs, so that the tract features mainly represent normal-appearing

WM and hence have little disease signal. Additionally, there is substantial inter-patient heterogeneity in the anatomical distribution of WM damage, which creates an unclear relation between microstructural changes in specific WM tracts and progression along the disease course.

The seven included cognitive domains are spread across the progression timeline but previous literature does not provide many concrete indications regarding the true positioning of those biomarkers given the lack of longitudinal data. However, the domains attention, executive function, and working memory were consistently late events in our analyses, which is supported by previous research [230]. Overall, we showed that the obtained event sequence is well in line with previous work on individual features but provides additional insight in the relative positioning of the multimodal features. The obtained sequence can potentially be used to stage patients within the disease course and help with clinical monitoring of disease progression beyond relapses and physical disability. However, the relatively high uncertainty limits use for individual patients at this stage.

5.3.6 Model 2: Sequence of events in the progression of low-to-high disability in relapse-onset multiple sclerosis.

The model for progression from low disability to high disability has many similarities to the general MS event sequence (Model 1) such as the early occurrence of cerebellar atrophy or visuospatial memory impairment, and the late events for white matter tract FA and the cognitive domains of attention, working memory and executive function. This is somewhat expected as minor impairment starts early in the disease course when brain structure is most similar to healthy controls. The most notable difference with Model 1 is the early increase of eigenvector centrality of the DMN and basal ganglia functional network, which supports findings on functional centrality as a correlate of physical disability [197]. Similarly, basal ganglia atrophy appears early in Model 2, supporting recent findings of deep GM atrophy being a driving factor in disability worsening [209]. The insula appears to be the earliest event but the considerable uncertainty suggests variability between individuals.

Changes of the MUCCA measurement appear earlier and FA changes of the corticospinal tract appear later with respect to Model 1. This ostensibly contradictory finding could be interpreted such that initial damage of the corticospinal tract already occurred in patients with low disability (i.e. first event in the progression from HC to MS) and more severe damage (i.e. spinal cord atrophy) will become apparent later. At the same time the cord area is not strongly affected initially but changes become more detectable after MS onset has occurred as indicated by previous studies that have shown the relevance of spinal cord atrophy in explaining long-term disability [231, 232].

The thalamus is broadly involved in cognitive and sensorimotor functions [233], which could explain the very early position in Model 1 and an early position in Model 2, and can be interpreted as a further increase in abnormality alongside the increase in disability.

MS lesions appear to become significant towards the cortex as disability progresses, i.e. first in the periventricular white matter, then in the deep WM and finally closer to the cortex, which is in line with other studies showing a larger lesion load around the ventricles with fewer lesions juxtacortically [234, 235]. It should be noted, however, that this study does not include measurements of cortical lesions, which needs to be addressed in subsequent studies.

5.3.7 Model 3: Sequence of events in relapsing-onset multiple sclerosis as cognition declines.

In the dEBM sequence from CP to CI, early events were atrophy of the insula, hippocampus and spinal cord, as well as the increased functional centrality of the DMN. The early appearance of insular atrophy in this model is interesting in the light of previous studies showing the fastest volume loss in these areas in patients with SPMS [209, 236]. We infer that these volume changes are an early event in the general MS population, confirmed by their respective positioning in a previous EBM study sequence [94].

A meaningful comparison of the functional centrality of networks is impeded by the exclusion of the basal ganglia and sensorimotor network biomarkers from the model due to statistically indistinguishable biomarker distributions between CP and CI groups, indicating that these have limited relevance to cognitive decline in MS. However, the increased functional centrality of the DMN was an early event in both model 2 and 3 suggesting that abnormality of DMN functional centrality could be an early indication of future cognitive and physical decline, as has been suggested extensively in MS literature [203, 218].

The interpretation and relevance of the early positioning of MUCCA in the cognitive model is difficult to understand but might reflect the overlap between patients with CI and patients with increased physical disability (64% of patients with CI in this cohort also have more severe physical disability; see also Table 5.2 and Figure 5.1). Lesion events appear in direct succession and the positional variance diagram (Figure 5.6) indicates that abnormal lesion volumes occur in all three locations roughly at the same time, indicating that other measures such as atrophy and brain function are more important for cognition.

Though thalamic atrophy has been associated with cognitive decline and disease progression [198], it appears relatively later (mid-sequence) than expected in the dEBM sequence. This could be the result of a floor-effect as there is already thalamic atrophy present in CP patients [228] and further changes arise late in the progression from CP to CI.

Microstructural WM changes appear late in model 3, which is consistent with model 2 and could imply that these measures reflect advanced stages of disease progression. A previous study showed that only CI patients with atrophy had microstructural WM changes and CP patients without atrophy did not have WM tract abnormalities [205]. Alternatively, the order in which different tracts become abnormal varies and more tracts are affected with advanced disease [196].

5.3.8 Considerations regarding features in the models

White matter lesions are a sensitive indicator for MS diagnosis [237] and are used extensively in daily clinical practice. We analysed lesion locations at three depths, with the inner band including the lesions close to the ventricles, the outer band including those close to the cortex, and the intermediate deep WM lesions in between [210]. While this definition is not as stringent as the clinically used stratification into periventricular, juxtacortical and deep lesions, it is a useful approximation that can be derived in a consistent and data-driven fashion. Infratentorial lesions were only present in a small subset of patients and were therefore discarded from further analysis despite their involvement in clinical disability. Although minor (vascular) WM lesions could be present in controls, these lesions could not be included due to the lack of FLAIR imaging in controls. As such, in the analysis of general ROMS progression, Model 1, we did not include lesions. However, lesions would be expected to occur very early in the MS sequence. MUCCA measurement was performed using SCT-PropSeg on 3DT1 head images, which may have reduced sensitivity to change compared to dedicated cervical cord imaging although several studies have shown good agreement between MUCCA derived from head and cervical images [216, 217]. We note that the considerable positional variance in the estimated ordering means that the exact positions of events should be interpreted with caution. Additionally, the ordering does not imply causation.

While we took care to include biomarkers of relevance to MS pathology, many more candidate biomarkers could be included in the future. Features such as spinal cord lesions [238], cerebrospinal fluid alterations [239], or (semi)quantitative MR measures of myelination such as magnetization transfer ratio (MTR) have been shown to be sensitive to the MS pathogenesis but were unavailable in this cohort.

5.3.9 Study limitations

MS is a heterogeneous disease with multiple concurrent disease processes, which are difficult to model, especially with limited data. As a consequence, some biomarkers show clear bimodal behaviour in the positional variance diagram (e.g., cingulate and insular atrophy in Model 1), which suggests different orderings for subgroups of our cohort. While this impedes

interpretation of some results, we believe that it is an important finding. An alternative way to model heterogeneous trajectories is to use advanced data-driven subtyping models such as SuStaIn [86], which could potentially identify clusters of subjects that share a differential sequence of events and hence model the disease progression in MS more reliably. However, this typically requires a larger dataset than is available here.

The effects of disease modifying treatment is very challenging to model due to the heterogeneity in the disease progression and the resulting treatment options. In general, we would expect a reduction of EDSS or lesion occurrence as these are the main outcome measures for clinical trials. In this study, this would lead to a change in group assignments, especially for Model 2, but we would not expect a strong effect on other biomarkers or their event sequence. A comparison of sequences obtained from treated and untreated patients, as well as the effect of a complex statistical correction for treatment effects, should be performed in an independent and sufficiently large cohort.

EBM provides a temporal ordering of biomarker abnormality, but no actual information about time as the intervals between subsequent events are not linear; this means that the division into late and early events can only be interpreted relative to other markers within the overall disease course. A combination of EBM-type models with longitudinal data and survival models, however, could give an estimate of the timescales of disease progression [61, 93].

5.3.10 Conclusion

This study has revealed the sequence of observable (biomarker) changes in brain structure, function, and cognition in the progression of ROMS, including specific sequences associated with disability worsening and cognitive decline. In general, changes in GM volume, especially of the thalamus, insula, hippocampus and cerebellum were the earliest events in MS and MS-related physical disability and cognitive decline, which also showed strong involvement of default-mode dysfunction. Microstructural changes in WM tracts were predominantly late events, which deserves further investigation as it appears to contradict the early occurrence of focal white matter lesions in many tracts, possibly indicating that overall tract integrity is maintained for a longer period of time. The relatively high uncertainty could be reduced using advanced models taking into account multiple concurrent disease trajectories within one cohort. Future research should also include patients soon after first symptoms arise (i.e. CIS) to determine the earliest disease pathologies in MS with high certainty.

| Clinical / Imaging measures | Total (n=243) | Cognitively preserved (n=168) | Cognitively impaired (n=75) | p-value |
|--|-----------------|-------------------------------|-----------------------------|---------------------|
| Clinical measures | | | | |
| Age [years]* | 47.1 (10.6) | 45.7 (10.3) | 50.4 (10.5) | 0.001 ^a |
| Sex [female, %]** | 174 (71.6) | 125 (74.4) | 49 (65.3) | 0.15 ^b |
| Education level [median, IQR]*** | 5 (4 - 6) | 6 (4 - 6) | 4 (3 6) | <0.001 ^c |
| Symptom duration [years]* | 14.6 (8.6) | 13.3 (7.7) | 17.6 (9.6) | <0.001 ^a |
| DMT used** | 141 (58.0) | 96 (57.1) | 45 (60.0) | 0.68 ^b |
| EDSS*** | 3.0 (2.0 - 4.0) | 3.0 (2.0 - 3.5) | 4.0 (3.0 6.0) | <0.001 ^c |
| RRMS/SPMS | 197 / 46 | 148 / 20 | 49 / 26 | <0.001 ^b |
| Information processing speed [z-score] * | -1.01 (1.4) | -0.35 (1.0) | -2.52 (1.2) | <0.001 ^a |
| Executive functioning [z-score] * | -0.95 (1.8) | -0.19 (0.8) | -2.73 (2.3) | <0.001 ^a |
| Working memory [z-score] * | -0.99 (1.6) | -0.39 (0.8) | -2.40 (2.0) | <0.001 ^a |
| Verbal memory [z-score] * | -0.44 (1.2) | 0.02 (0.9) | -0.15 (1.1) | <0.001 ^a |
| Verbal fluency [z-score] * | -0.38 (1.1) | -0.04 (1.0) | -1.2 (1.0) | <0.001 ^a |
| Visuospatial memory [z-score] * | -0.56 (1.2) | -0.14 (1.0) | -1.49 (1.1) | <0.001 ^a |
| Attention [z-score] * | -0.60 (1.2) | -0.22 (0.7) | -1.48 (1.6) | <0.001 ^a |
| MRI measures | | | | |
| <i>T2-hyperintense lesion loads [mL]*</i> | | | | |
| Total T2-hyperintense lesion load | 13.7 (12.8) | 10.3 (8.4) | 21.4 (17.0) | <0.001 ^a |
| Inner lesions | 4.1 (3.2) | 3.3 (2.3) | 5.9 (4.0) | <0.001 ^a |
| Deep lesions | 6.3 (7.2) | 4.6 (4.5) | 10.3 (10.0) | <0.001 ^a |
| Outer lesions | 3.3 (3.3) | 2.4 (2.1) | 5.3 (4.4) | <0.001 ^a |
| Infratentorial | 0.01 (0.04) | 0.01 (0.04) | 0.01 (0.03) | 0.95 ^a |
| <i>Brain and spinal cord volumes [mL] *</i> | | | | |
| Total brain volume | 1136.4 (110.9) | 1147.5 (110.5) | 1111.4 (108.3) | 0.019 ^a |
| Basal Ganglia | 34.3 (3.5) | 34.9 (3.3) | 33.0 (3.7) | <0.001 ^a |
| Hippocampus | 7.6 (0.7) | 7.7 (0.7) | 7.5 (0.8) | 0.02 ^a |
| Thalamus | 10.1 (1.5) | 10.5 (1.3) | 9.2 (1.5) | <0.001 ^a |
| Cingulate | 27.6 (3.3) | 27.8 (3.3) | 27.3 (3.3) | 0.256 ^a |
| Frontal lobe | 180.4 (19.2) | 182.9 (19.1) | 174.8 (18.4) | 0.002 ^a |
| Insula | 10.7 (1.3) | 10.9 (1.3) | 10.2 (1.1) | <0.001 ^a |
| Occipital lobe | 66.9 (8.3) | 68.3 (8.1) | 63.9 (8.0) | <0.001 ^a |
| Parietal lobe | 91.3 (10.6) | 92.5 (10.6) | 88.7 (10.1) | 0.01 ^a |
| Temporal lobe | 128.2 (13.5) | 129.5 (13.6) | 125.0 (12.9) | 0.02 ^a |
| Cerebellar grey matter | 94.5 (9.5) | 95.4 (9.5) | 92.6 (9.1) | 0.031 ^a |
| MUCCA | 64.5 (7.9) | 65.7 (7.6) | 61.6 (8.0) | 0.001 ^a |
| <i>Functional hubness (EC [z-scores])</i> | | | | |
| Basal ganglia network | -0.020 (0.242) | -0.026 (0.252) | -0.006 (0.216) | 0.549 ^a |
| Default mode network | 0.007 (0.213) | -0.013 (0.223) | 0.052 (0.184) | 0.027 ^a |
| Sensorimotor cortex network | -0.042 (0.222) | -0.037 (0.227) | -0.053 (0.209) | 0.600 ^a |
| <i>White matter tract integrity (FA [0-1])</i> | | | | |
| Anterior thalamic radiation | 0.45 (0.04) | 0.46 (0.03) | 0.43 (0.04) | <0.001 ^a |
| Corticospinal tract | 0.65 (0.03) | 0.66 (0.03) | 0.64 (0.03) | 0.002 ^a |
| Cingulum | 0.56 (0.05) | 0.58 (0.04) | 0.54 (0.06) | <0.001 ^a |
| Other WM tracts | 0.53 (0.04) | 0.54 (0.03) | 0.50 (0.05) | <0.001 ^a |

Table 5.3: Demographics of cognitive preserved vs cognitive impaired. Biomarkers with a $p < 0.1$ were included in the model. * Mean (standard deviation), ** number (percentage), *** median (IQR). ^a Independent-samples t-test, ^b chi-square test, ^c Mann-Whitney U test.

Supplementary material

The combination of features from different domains, such as volumetric, FA or cognition, has no effect on the general event ordering, except for small variations within the uncertainty estimated by the bootstrap experiments. In the supplementary materials we present additional data from experiments with subsets of the biomarkers used in Model 1 to support this statement.

Exclusion of cognitive biomarkers

The event sequence obtained when using no cognitive biomarkers is almost identical (Figure 5.SF1) to the original ordering of Model 1 as shown in Figure 5.2, except for the excluded biomarkers, of course. The only difference is in the last three biomarkers Cingulate, Insula and Anterior thalamic radiation, where Cingulate and Anterior thalamic radiation have swapped position. This can be explained by the very high positional uncertainty of these biomarkers.

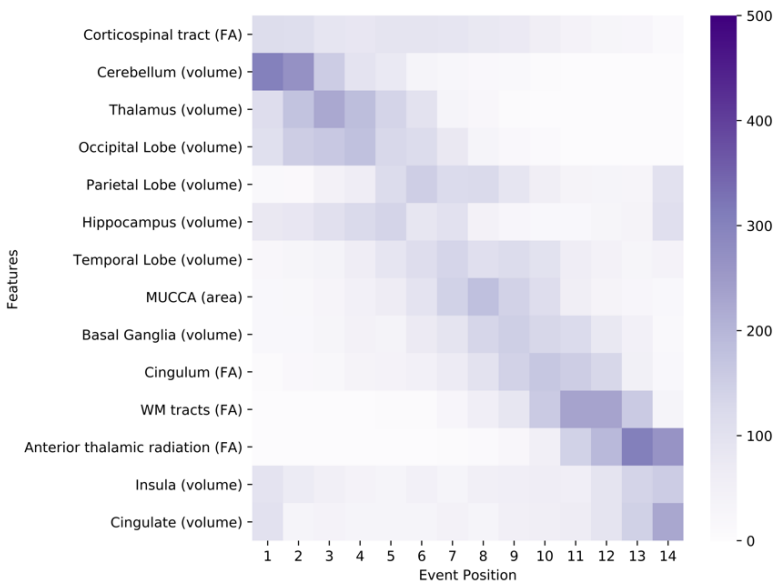


Figure 5.SF1: Positional variance diagram for subjects and biomarkers from Model 1 but excluding cognitive biomarkers. Please note that the colourbar is capped at 500 to improve visualisation.

Similarly, when creating an EBM using only cognitive biomarkers (Figure 5.SF2) we obtain an ordering that is alike to the ordering of cognitive features within the more comprehensive Model 1. Only ‘Verbal Memory’ and ‘Information Processing’, and ‘Executive Function’ and ‘Working Memory’ have swapped positions. These pairs of biomarkers are very close in event position, so this change is within the uncertainty estimated from the bootstraps.

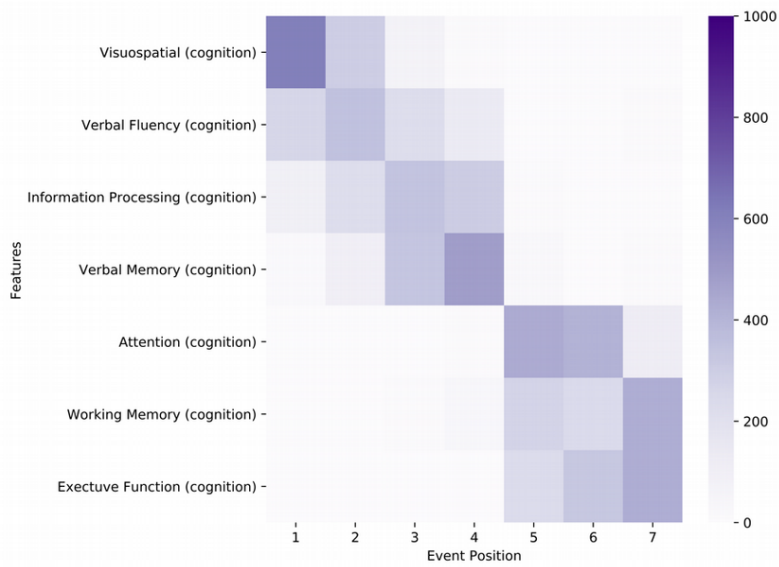


Figure 5.SF2: Positional variance diagram for subjects from Model 1 using only cognitive biomarkers.

Variation of included tracts for FA biomarkers

The relative ordering of the four tract-related biomarkers, when modelled independently of all other biomarkers as shown in Figure 5.SF3, is identical to the one obtained in Model 1 as shown in Figure 5.2.

The biomarker ‘WM tracts’ includes several large tracts, which raises the question of their relative positioning to the other three separately included features. In Figure 5.SF4 we show the ordering and positional variance of all major tracts in the JHU WM tractography atlas. It can be seen that the relative ordering of the original three separately included tracts from Model 1 does not change when fitting an EBM using all main tracts from the JHU WM tractography atlas (see Figure 5.SF4).

The FA of the uncinate fasciculus is positioned after the anterior thalamic radiation at the end of the sequence, and only the inferior longitudinal fasciculus is positioned between the corticospinal tract and the cingulum, albeit with some uncertainty. The other 3 tracts that were previously bundles together as ‘WM tracts’ remain in between the cingulum and the anterior thalamic radiation. Overall, this ordering is well in line with the sequence obtained in Model 1.

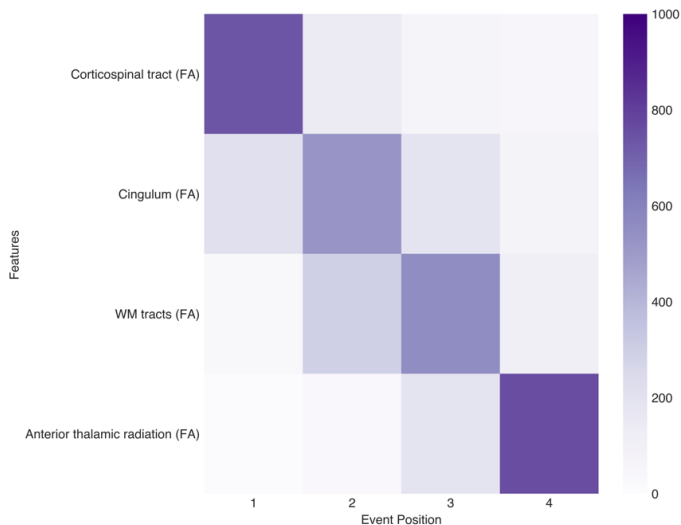


Figure 5.SF3: Positional variance diagram for subjects from Model 1 using only the four tract-based biomarkers.

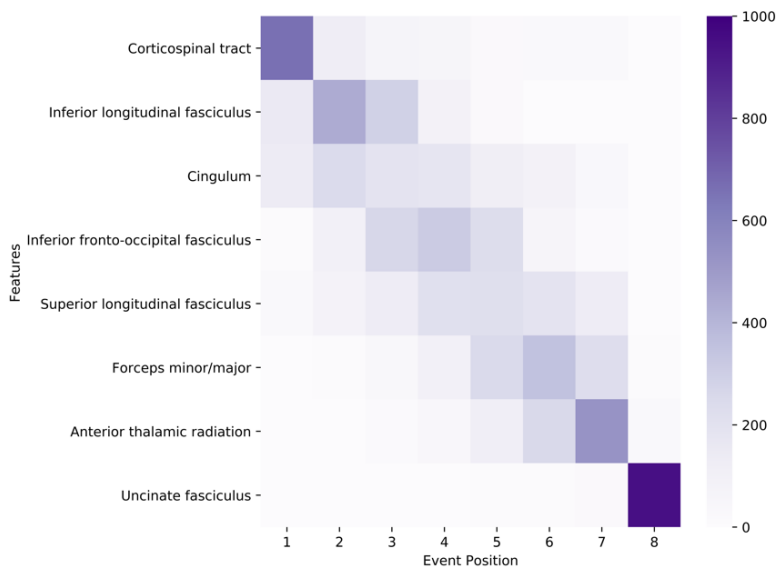


Figure 5.SF4: Positional variance diagram for subjects from Model 1 using individual from the JHU atlas instead of merging some of them.

Chapter 6

The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge: Results after 1 Year Follow-up

This chapter contains the content of the manuscript 'The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge: Results after 1 Year Follow-up'. Razvan V. Marinescu, Neil P. Oxtoby, Alexandra L. Young, Esther E. Bron, Arthur W. Toga, Michael W. Weiner, Frederik Barkhof, Nick C. Fox, Arman Eshaghi, Tina Toni, Marcin Salaterski, Veronika Lunina, Manon Ansart, Stanley Durrleman, Pascal Lu, Samuel Iddi, Dan Li, Wesley K. Thompson, Michael C. Donohue, Aviv Nahon, Yarden Levy, Dan Halbersberg, Mariya Cohen, Huiling Liao, Tengfei Li, Kaixian Yu, Hongtu Zhu, Jose G. Tamez-Pena, Aya Ismail, Timothy Wood, Hector Corrada Bravo, Minh Nguyen, Nanbo Sun, Jiashi Feng, B. T. Thomas Yeo, Gang Chen, Ke Qi, Shiyang Chen, Deqiang Qiu, Ionut Buciuman, Alex Kelner, Raluca Pop, Denisa Rimoccea, Mostafa M. Ghazi, Mads Nielsen, Sebastien Ourselin, Lauge Sorensen, **Vikram Venkatraghavan**, Keli Liu, Christina Rabe, Paul Manser, Steven M. Hill, James Howlett, Zhiyue Huang, Steven Kiddle, Sach Mukherjee, Anais Rouanet, Bernd Taschler, Brian D. M. Tom, Simon R. White, Noel Faux, Suman Sedai, Javier de Velasco Oriol, Edgar E. V. Clemente, Karol Estrada, Leon Aksman, Andre Altmann, Cynthia M. Stonnington, Yalin Wang, Jianfeng Wu, Vivek Devadas, Clementine Fourrier, Lars Lau Raket, Aristeidis Sotiras, Guray Erus, Jimit Doshi, Christos Davatzikos, Jacob Vogel, Andrew Doyle, Angela Tam, Alex Diaz-Papkovich, Emmanuel Jammeh, Igor Koval, Paul Moore, Terry J. Lyons, John Gallacher, Jussi Tohka, Robert Cizek, Bruno Jedynak, Kruti Pandya, Murat Bilgel, William Engels, Joseph Cole, Polina Golland, Stefan Klein, Daniel C. Alexander. *In revision*'.

Vikram Venkatraghavan's contributions in this chapter are: participation in the challenge using a newly developed method (Team EMC1), which ended up as one of the winning methods; drafting of the method's description, and critical review of the entire manuscript.

Abstract

Accurate prediction of progression in subjects at risk of Alzheimer’s disease is crucial for enrolling the right subjects in clinical trials. However, a prospective comparison of state-of-the-art algorithms for predicting disease onset and progression is currently lacking. We present the findings of *The Alzheimer’s Disease Prediction Of Longitudinal Evolution* (TADPOLE) Challenge, which compared the performance of 92 algorithms from 33 international teams at predicting the future trajectory of 219 individuals at risk of Alzheimer’s disease. Challenge participants were required to make a prediction, for each month of a 5-year future time period, of three key outcomes: clinical diagnosis, Alzheimer’s Disease Assessment Scale Cognitive Subdomain (ADAS-Cog13), and total volume of the ventricles. No single submission was best at predicting all three outcomes. For clinical diagnosis and ventricle volume prediction, the best algorithms strongly outperform simple baselines in predictive ability. However, for ADAS-Cog13 no single submitted prediction method was significantly better than random guessing. On a limited, cross-sectional subset of the data emulating clinical trials, performance of best algorithms at predicting clinical diagnosis decreased only slightly (3% error increase) compared to the full longitudinal dataset. Two ensemble methods based on taking the mean and median over all predictions, obtained top scores on almost all tasks. Better than average performance at diagnosis prediction was generally associated with the additional inclusion of features from cerebrospinal fluid (CSF) samples and diffusion tensor imaging (DTI). On the other hand, better performance at ventricle volume prediction was associated with inclusion of summary statistics, such as patient-specific biomarker trends. The submission system remains open via the website <https://tadpole.grand-challenge.org>, while code for submissions is being collated by TADPOLE SHARE: <https://tadpole-share.github.io/>. Our work suggests that current prediction algorithms are accurate for biomarkers related to clinical diagnosis and ventricle volume, opening up the possibility of cohort refinement in clinical trials for Alzheimer’s disease.

6.1 Introduction

Accurate prediction of the onset of Alzheimer’s disease (AD) and its longitudinal progression is important for care planning and for patient selection in clinical trials. Current opinion holds that early detection will be critical for the successful administration of disease modifying treatments during presymptomatic phases of the disease prior to widespread brain damage, e.g. when pathological amyloid and tau start to accumulate [240]. Moreover, accurate prediction of the progression of at-risk subjects will help select homogenous patient groups for clinical trials, thus reducing variability in outcome measures that can obscure positive effects on patients at the right stage to benefit.

Several mathematical and computational methods have been developed to predict the onset and progression of AD. Traditional approaches leverage statistical regression to model relationships between target variables (e.g. clinical diagnosis or cognitive/imaging markers) with other known markers [241, 242] or measures derived from these markers such as the rate of cognitive decline [243]. More recent approaches involve supervised machine learning techniques such as support vector machines, random forests and artificial neural networks. These approaches have been used to discriminate AD patients from cognitively normal

individuals [244, 245], and for discriminating at-risk individuals who convert to AD in a certain time frame from those who do not [246, 247]. The emerging approach of disease progression modelling aims to reconstruct biomarker trajectories or other disease signatures across the disease progression timeline, without relying on clinical diagnoses or estimates of time to symptom onset. Examples include models built on a set of scalar biomarkers to produce discrete [53, 61] or continuous [51, 52, 57, 68, 248] biomarker trajectories; spatio-temporal models that focus on evolving image structure [249, 250], potentially conditioned by non-imaging variables [251]; and models that emulate putative disease mechanisms to estimate trajectories of change [60, 252, 253]. All these models show promise for predicting AD biomarker progression at group and individual levels. However, previous evaluations within individual publications provide limited information because: (1) they use different data sets or subsets of the same dataset, different processing pipelines, and different evaluation metrics and (2) over-training can occur due to heavy use of popular training datasets. Currently, the field lacks a comprehensive comparison of the capabilities of these methods on standardised tasks relevant to real-world applications.

Community challenges have consistently proved effective in moving forward the state of the art in technology to address specific data-analysis problems by providing platforms for unbiased comparative evaluation and incentives to maximise performance on key tasks [254]. In medical image analysis, for example, such challenges have provided important benchmarks in tasks such as registration [255] and segmentation [256], and revealed fundamental insights about the problem studied, for example in structural brain-connectivity mapping [257]. Previous challenges in AD include the CADDementia challenge [258], which aimed to identify clinical diagnosis from MRI scans. A similar challenge, the *International challenge for automated prediction of MCI from MRI data* [259], asked participants to predict diagnosis and conversion status from extracted MRI features of subjects from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) study [260]. Yet another challenge, *The Alzheimer’s Disease Big Data DREAM Challenge* [261], asked participants to predict cognitive decline from genetic and MRI data. These challenges have however several limitations: (i) they did not evaluate the ability of algorithms to predict biomarkers at future timepoints (with the exception of one sub-challenge of DREAM), which is important for patient stratification in clinical trials; (ii) the test data was available to organisers when the competitions were launched, leaving room for potential biases in the design of the challenges; (iii) the training data was drawn from a limited set of modalities.

The Alzheimer’s Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge (<https://tadpole.grand-challenge.org>) aims to identify the data, features and approaches that are the most predictive of future progression of subjects at risk of AD. In contrast to previous challenges, our challenge is designed to inform clinical trials through identification of patients most likely to benefit from an effective treatment, i.e., those at early stages of

disease who are likely to progress over the short-to-medium term (1-5 years). The challenge focuses on forecasting the trajectories of three key features: clinical status, cognitive decline, and neurodegeneration (brain atrophy), over a five-year timescale. It uses “rollover” subjects from the ADNI study [260] for whom a history of measurements (imaging, psychology, demographics, genetics) is available, and who are expected to continue in the study, providing future measurements for testing. TADPOLE participants were required to predict future measurements from these individuals and submit their predictions before a given submission deadline. Since the test data *did not exist* at the time of forecast submissions, the challenge provides a performance comparison substantially less susceptible to many forms of potential bias than previous studies and challenges. The design choices were published [69] before the test set was acquired and analysed. TADPOLE also goes beyond previous challenges by drawing on a vast set of multimodal measurements from ADNI which might support prediction of AD progression.

This article presents the results of the TADPOLE Challenge and documents its key findings. We summarise the challenge design and present the results of the 92 prediction algorithms contributed by 33 participating teams worldwide, evaluated after an 18-month follow-up period. We discuss the results obtained by TADPOLE participants, which represent the current state-of-the-art in Alzheimer’s disease prediction. To understand what key characteristics of algorithms were important for good predictions, we also report results on which input data features were most informative, and which feature selection strategies, data imputation methods and classes of algorithms were most effective.

6.2 Results

6.2.1 Predictions

TADPOLE Challenge asked participants to forecast three key biomarkers: (1) clinical diagnosis, which can be either cognitively normal (CN), mild cognitive impairment (MCI), or probable AD; (2) Alzheimer’s Disease Assessment Scale Cognitive Subdomain (ADAS-Cog13) score; and (3) ventricle volume (divided by intra-cranial volume) from MRI. The exact time of future data acquisitions for any given individual was unknown at forecast time, so participants submitted month-by-month predictions for every individual. Predictions of clinical status comprise relative likelihoods of each option (CN, MCI, and AD) for each individual at each month. Predictions of ADAS-Cog13 and ventricle volume comprise a best-guess estimate as well as a 50% confidence interval for each individual at each month. Full details on challenge design are given in the TADPOLE white paper [69].

| Submission | Feature selection | Number of features | Missing data imputation | Diagnosis prediction model | ADAS/Vent. prediction model | Training time | Prediction time (one subject) |
|------------------------------|-------------------|--------------------|----------------------------------|--------------------------------|--|---------------|-------------------------------|
| AlgosForGood | manual | 16+5* | forward-filling | Aalen model | linear regression | 1 min. | 1 sec. |
| Apocalypse | manual | 16 | population average | SVM | linear regression | 40 min. | 3 min. |
| ARAMIS-Pascal | manual | 20 | population average | Aalen model | - | 16 sec. | 0.02 sec. |
| ATRI-Biostat-JMM | automatic | 15 | random forest | random forest | linear mixed effects model | 2 days | 1 sec. |
| ATRI-Biostat-LTJMM | automatic | 15 | random forest | random forest | DPM | 2 days | 1 sec. |
| ATRI-Biostat-MA | automatic | 15 | random forest | random forest | DPM + linear mixed effects model | 2 days | 1 sec. |
| BGU-LSTM | automatic | 67 | none | feed-forward NN | LSTM | 1 day | millisec. |
| BGU-RF/ BGU-RFFIX | automatic | ≈ 67+1340* | none | semi-temporal RF | semi-temporal RF | a few min. | millisec. |
| BIGS2 | automatic | all | Iterative Thresholded SVD | RF | linear regression | 2.2 sec. | 0.001 sec. |
| Billabong (all) | manual | 15-16 | linear regression | linear scale | non-parametric SM | 7 hours | 0.13 sec. |
| BORREGOSTECTMY | automatic | ≈100 + 400* | nearest-neighbour | regression ensemble | ensemble of regression + hazard models | 18 hours | 0.001 sec. |
| BravoLab | automatic | 25 | hot deck | LSTM | LSTM | 1 hour | a few sec. |
| CBIL | manual | 21 | linear interpolation | LSTM | LSTM | 1 hour | one min. |
| Chen-MCW | manual | 9 | none | linear regression | DPM | 4 hours | < 1 hour |
| CN2L-NeuralNetwork | automatic | all | forward-filling | RNN | RNN | 24 hours | a few sec. |
| CN2L-RandomForest | manual | >200 | forward-filling | RF | RF | 15 min. | < 1 min. |
| CN2L-Average | automatic | all | forward-filling | RNN/RF | RNN/RF | 24 hours | < 1 min. |
| CyberBrains | manual | 5 | population average | linear regression | linear regression | 20 sec. | 20 sec. |
| DIKU (all) | semi-automatic | 18 | none | Bayesian classifier /LDA + DPM | DPM | 290 sec. | 0.025 sec. |
| DIVE | manual | 13 | none | KDE+DPM | DPM | 20 min. | 0.06 sec. |
| EMC1 | automatic | 250 | nearest neighbour | DPM + 2D spline + SVM | DPM + 2D spline | 80 min. | a few sec. |
| EMC-EB | automatic | 200-338 | nearest-neighbour | SVM classifier | SVM regressor | 20 sec. | a few sec. |
| FortuneTellerFish-Control | manual | 19 | nearest neighbour | multiclass ECOC SVM | linear mixed effects model | 1 min. | < 1 sec. |
| FortuneTellerFish-SuStaln | manual | 19 | nearest neighbour | multiclass ECOC SVM + DPM | linear mixed effects model + DPM | 5 hours | < 1 sec. |
| Frog | automatic | ≈ 70+420* | none | gradient boosting | gradient boosting | 1 hour | - |
| GlassFrog-LCMEM-HDR | semi-automatic | all | forward-fill/nearest-neigh. | multi-state model | DPM + regression | 15 min. | 2 min. |
| GlassFrog-SM | manual | 7 | linear model | multi-state model | parametric SM | 93 sec. | 0.1 sec. |
| GlassFrog-Average | semi-automatic | all | forward-fill/linear | multi-state model | DPM + SM + regression | 15 min. | 2 min. |
| IBM-OZ-Res | manual | Oct-15 | filled with zero | stochastic gradient boosting | stochastic gradient boosting | 20 min. | 0.1 sec. |
| ITEMCEM | manual | 48 | mean of previous values | RF | LASSO + Bayesian ridge regression | 20 min. | 0.3 sec. |
| ImaUCL (all) | manual | 5 | regression | multi-task learning | multi-task learning | 2 hours | millisec. |
| Mayo-BAL-ASU | manual | 15 | population average | linear mixed effects model | linear mixed effects model | 20 min. | 1.3 sec. |
| Orange | manual | 17 | none | clinician's decision tree | clinician's decision tree | none | 0.2 sec. |
| Rocket | manual | 6 | median of diagnostic group | linear mixed effects model | DPM | 5 min. | 0.3 sec. |
| SBIA | manual | 30-70 | dropped visits with missing data | SVM + density estimator | linear mixed effects model | 1 min. | a few sec. |
| SPMC-Plymouth (all) | automatic | 20 | none | unknown | - | unknown | 1 min. |
| SmallHeads-NeuralNetwork | automatic | 376 | nearest neighbour | deep fully-connected NN | deep fully-connected NN | 40 min. | 0.06 sec. |
| SmallHeads-LinMixedEffects | automatic | unknown | nearest neighbour | - | linear mixed effects model | 25 min. | 0.13 sec. |
| Sunshine (all) | semi-automatic | 6 | population average | SVM | linear model | 30 min. | < 1 min. |
| Threedays | manual | 16 | none | RF | - | 1 min. | 3 sec. |
| Tohka-Ciszek-SMNSR | manual | ≈ 32 | nearest neighbour | - | SMNSR | several hours | a few sec. |
| Tohka-Ciszek-RandomForestLin | manual | 32 | mean patient value | RF | linear model | a few min. | a few sec. |
| VikingAI (all) | manual | 10 | none | DPM + ordered logit model | DPM | 10 hours | 8 sec. |
| BenchmarkLastVisit | None | 3 | none | constant model | constant model | 7 sec. | millisec. |
| BenchmarkMixedEffects | None | 3 | none | Gaussian model | linear mixed effects model | 30 sec. | 0.003 sec. |
| BenchmarkMixedEffects-APOE | None | 4 | none | Gaussian model | linear mixed effects model | 30 sec. | 0.003 sec. |
| BenchmarkSVM | manual | 6 | mean of previous values | SVM | support vector regressor (SVR) | 20 sec. | 0.001 sec. |

Table 6.1: Summary of prediction methods used in the TADPOLE submissions. Keywords: SVM – Support Vector Machine, RF – random forest, LSTM – long short-term memory network, NN – neural network, RNN – recurrent neural network, SMNSR – Sparse Multimodal Neighbourhood Search Regression, DPM – disease progression model, KDE – kernel density estimation, LDA – linear discriminant analysis, SM – slope model, ECOC – error-correcting output codes, SVD – singular value decomposition (*) Augmented features, or summary statistics, such as trends, slope, min/max, moments, generally derived patient-wise using longitudinal data. Color tags denote prediction method category: ■ regression/proportional hazards model, ■ random forest, ■ neural networks, ■ disease progression model, ■ machine learning (other), ■ benchmark, ■ other. The left-side box denotes the category for diagnosis prediction method, while the right-side box denotes the category for ADAS/Ventricle prediction method.

6.2.2 Algorithms

We had a total of 33 participating teams, who submitted a total of 58 predictions from the longitudinal prediction set (D2), 34 predictions from the cross-sectional prediction set (D3), and 6 predictions from custom prediction sets (see Online Methods section 6.5.1 for description of D2/D3 datasets). A total of 8 D2/D3 submissions from 6 teams did not have predictions for all three target variables, so we computed the performance metrics for only the submitted target variables. Another 3 submissions lacked confidence intervals for either ADAS-Cog13 or ventricle volume, which we imputed using default low-width confidence ranges of 2 for ADAS-Cog13 and 0.002 for Ventricles normalised by intracranial volume (ICV).

Table 6.1 summarises the methods used in the submissions in terms of feature selection, handling of missing data, predictive models for clinical diagnosis and ADAS/Ventricles biomarkers, as well as training and prediction times. A detailed description of each method is in Online Methods Section 6.5.4. In particular, some entries constructed augmented features (i.e. summary statistics), which are extra features such as slope, min/max or moments that are derived from existing features.

In addition to the forecasts submitted by participants, we also evaluated four benchmark methods, which were made available to participants during the submission phase of the challenge: (i) *BenchmarkLastVisit* uses the measurement of each target from the last available clinical visit as the forecast, (ii) *BenchmarkMixedEffects* uses a mixed effects model with age as predictor variable for ADAS and Ventricle predictions, and Gaussian likelihood model for diagnosis prediction, (iii) *BenchmarkMixedEffectsAPOE* is as (ii) but adds APOE status as a covariate and (iv) *BenchmarkSVM* uses an out-of-the-box support vector machine (SVM) classifier and regressor (SVR) to provide forecasts. More details on these methods can be found in Online Methods section 6.5.4. We also evaluated two ensemble methods based on taking the mean (*ConsensusMean*) and median (*ConsensusMedian*) of the forecasted variables over all submissions.

To control for potentially spurious strong performance arising from multiple comparisons, we also evaluated 100 random predictions by adding Gaussian noise to the forecasts of the simplest benchmark model (*BenchmarkLastVisit*). In the subsequent results tables we will show, for each performance metric, only the best score obtained by any of these 100 random predictions (*RandomisedBest*) – See end of Online Methods section 6.5.4 for more information on *RandomisedBest*.

6.2.3 Forecasts from the longitudinal prediction set (D2)

Table 6.2 compiles all metrics for all TADPOLE submitted forecasts, as well as benchmarks and ensemble forecasts, from the longitudinal D2 prediction set. For details on datasets

| | Submission | Overall Rank | Diagnosis | | ADAS-Cog13 | | | Ventricles (% ICV) | | | | | |
|--|------------------------------|--------------|-----------|-------|------------|-------|-------|--------------------|------|-------|------|------|------|
| | | | Rank | MAUC | BCA | Rank | MAE | WES | CPA | Rank | MAE | WES | CPA |
| | ConsensusMedian | - | - | 0.925 | 0.857 | - | 5.12 | 5.01 | 0.28 | - | 0.38 | 0.33 | 0.09 |
| | Frog | 1 | 1 | 0.931 | 0.849 | 4 | 4.85 | 4.74 | 0.44 | 10 | 0.45 | 0.33 | 0.47 |
| | ConsensusMean | - | - | 0.920 | 0.835 | - | 3.75 | 3.54 | 0.00 | - | 0.48 | 0.45 | 0.13 |
| | EMC1-Std | 2 | 8 | 0.898 | 0.811 | 23-24 | 6.05 | 5.40 | 0.45 | 1-2 | 0.41 | 0.29 | 0.43 |
| | VikingAI-Sigmoid | 3 | 16 | 0.875 | 0.760 | 7 | 5.20 | 5.11 | 0.02 | 11-12 | 0.45 | 0.35 | 0.20 |
| | EMC1-Custom | 4 | 11 | 0.892 | 0.798 | 23-24 | 6.05 | 5.40 | 0.45 | 1-2 | 0.41 | 0.29 | 0.43 |
| | CBIL | 5 | 9 | 0.897 | 0.803 | 15 | 5.66 | 5.65 | 0.37 | 13 | 0.46 | 0.46 | 0.09 |
| | Apocalypse | 6 | 7 | 0.902 | 0.827 | 14 | 5.57 | 5.57 | 0.50 | 20 | 0.52 | 0.52 | 0.50 |
| | GlassFrog-Average | 7 | 4-6 | 0.902 | 0.825 | 8 | 5.26 | 5.27 | 0.26 | 29 | 0.68 | 0.60 | 0.33 |
| | GlassFrog-SM | 8 | 4-6 | 0.902 | 0.825 | 17 | 5.77 | 5.92 | 0.20 | 21 | 0.52 | 0.33 | 0.20 |
| | BORREGOTECMTY | 9 | 19 | 0.866 | 0.808 | 20 | 5.90 | 5.82 | 0.39 | 5 | 0.43 | 0.37 | 0.40 |
| | BenchmarkMixedEffects | - | - | 0.846 | 0.706 | - | 4.19 | 4.19 | 0.31 | - | 0.56 | 0.56 | 0.50 |
| | EMC-EB | 10 | 3 | 0.907 | 0.805 | 39 | 6.75 | 6.66 | 0.50 | 9 | 0.45 | 0.40 | 0.48 |
| | lmaUCL-Covariates | 11-12 | 22 | 0.852 | 0.760 | 27 | 6.28 | 6.29 | 0.28 | 3 | 0.42 | 0.41 | 0.11 |
| | CN2L-Average | 11-12 | 27 | 0.843 | 0.792 | 9 | 5.31 | 5.31 | 0.35 | 16 | 0.49 | 0.49 | 0.33 |
| | VikingAI-Logistic | 13 | 20 | 0.865 | 0.754 | 21 | 6.02 | 5.91 | 0.26 | 11-12 | 0.45 | 0.35 | 0.20 |
| | lmaUCL-Std | 14 | 21 | 0.859 | 0.781 | 28 | 6.30 | 6.33 | 0.26 | 4 | 0.42 | 0.41 | 0.09 |
| | RandomisedBest | - | - | 0.800 | 0.803 | - | 4.52 | 4.52 | 0.27 | - | 0.46 | 0.45 | 0.33 |
| | CN2L-RandomForest | 15-16 | 10 | 0.896 | 0.792 | 16 | 5.73 | 5.73 | 0.42 | 31 | 0.71 | 0.71 | 0.41 |
| | FortuneTellerFish-SuStaln | 15-16 | 40 | 0.806 | 0.685 | 3 | 4.81 | 4.81 | 0.21 | 14 | 0.49 | 0.49 | 0.18 |
| | CN2L-NeuralNetwork | 17 | 41 | 0.783 | 0.717 | 10 | 5.36 | 5.36 | 0.34 | 7 | 0.44 | 0.44 | 0.27 |
| | BenchmarkMixedEffectsAPOE | 18 | 35 | 0.822 | 0.749 | 2 | 4.75 | 4.75 | 0.36 | 23 | 0.57 | 0.57 | 0.40 |
| | Tohka-Ciszek-RandomForestLin | 19 | 17 | 0.875 | 0.796 | 22 | 6.03 | 6.03 | 0.15 | 22 | 0.56 | 0.56 | 0.37 |
| | BGU-LSTM | 20 | 12 | 0.883 | 0.779 | 25 | 6.09 | 6.12 | 0.39 | 25 | 0.60 | 0.60 | 0.23 |
| | DIKU-GeneralisedLog-Custom | 21 | 13 | 0.878 | 0.790 | 11-12 | 5.40 | 5.40 | 0.26 | 38-39 | 1.05 | 1.05 | 0.05 |
| | DIKU-GeneralisedLog-Std | 22 | 14 | 0.877 | 0.790 | 11-12 | 5.40 | 5.40 | 0.26 | 38-39 | 1.05 | 1.05 | 0.05 |
| | CyberBrains | 23 | 34 | 0.823 | 0.747 | 6 | 5.16 | 5.16 | 0.24 | 26 | 0.62 | 0.62 | 0.12 |
| | AlgosForGood | 24 | 24 | 0.847 | 0.810 | 13 | 5.46 | 5.11 | 0.13 | 30 | 0.69 | 3.31 | 0.19 |
| | lmaUCL-halfD1 | 25 | 26 | 0.845 | 0.753 | 38 | 6.53 | 6.51 | 0.31 | 6 | 0.44 | 0.42 | 0.13 |
| | BGU-RF | 26 | 28 | 0.838 | 0.673 | 29-30 | 6.33 | 6.10 | 0.35 | 17-18 | 0.50 | 0.38 | 0.26 |
| | Mayo-BAl-ASU | 27 | 52 | 0.691 | 0.624 | 5 | 4.98 | 4.98 | 0.32 | 19 | 0.52 | 0.52 | 0.40 |
| | BGU-RFFIX | 28 | 32 | 0.831 | 0.673 | 29-30 | 6.33 | 6.10 | 0.35 | 17-18 | 0.50 | 0.38 | 0.26 |
| | FortuneTellerFish-Control | 29 | 31 | 0.834 | 0.692 | 1 | 4.70 | 4.70 | 0.22 | 50 | 1.38 | 1.38 | 0.50 |
| | GlassFrog-LCMEM-HDR | 30 | 4-6 | 0.902 | 0.825 | 31 | 6.34 | 6.21 | 0.47 | 51 | 1.66 | 1.59 | 0.41 |
| | SBLA | 31 | 43 | 0.776 | 0.721 | 43 | 7.10 | 7.38 | 0.40 | 8 | 0.44 | 0.31 | 0.13 |
| | Chen-MCW-Stratify | 32 | 23 | 0.848 | 0.783 | 36-37 | 6.48 | 6.24 | 0.23 | 36-37 | 1.01 | 1.00 | 0.11 |
| | Rocket | 33 | 54 | 0.680 | 0.519 | 18 | 5.81 | 5.71 | 0.34 | 28 | 0.64 | 0.64 | 0.29 |
| | BenchmarkSVM | 34-35 | 30 | 0.836 | 0.764 | 40 | 6.82 | 6.82 | 0.42 | 32 | 0.86 | 0.84 | 0.50 |
| | Chen-MCW-Std | 34-35 | 29 | 0.836 | 0.778 | 36-37 | 6.48 | 6.24 | 0.23 | 36-37 | 1.01 | 1.00 | 0.11 |
| | DIKU-ModifiedMri-Custom | 36 | 36-37 | 0.807 | 0.670 | 32-35 | 6.44 | 6.44 | 0.27 | 34-35 | 0.92 | 0.92 | 0.01 |
| | DIKU-ModifiedMri-Std | 37 | 38-39 | 0.806 | 0.670 | 32-35 | 6.44 | 6.44 | 0.27 | 34-35 | 0.92 | 0.92 | 0.01 |
| | DIVE | 38 | 51 | 0.708 | 0.568 | 42 | 7.10 | 7.10 | 0.34 | 15 | 0.49 | 0.49 | 0.13 |
| | ITESMCEM | 39 | 53 | 0.680 | 0.657 | 26 | 6.26 | 6.26 | 0.35 | 33 | 0.92 | 0.92 | 0.43 |
| | BenchmarkLastVisit | 40 | 44-45 | 0.774 | 0.792 | 41 | 7.05 | 7.05 | 0.45 | 27 | 0.63 | 0.61 | 0.47 |
| | Sunshine-Conservative | 41 | 25 | 0.845 | 0.816 | 44-45 | 7.90 | 7.90 | 0.50 | 43-44 | 1.12 | 1.12 | 0.50 |
| | BravoLab | 42 | 46 | 0.771 | 0.682 | 47 | 8.22 | 8.22 | 0.49 | 24 | 0.58 | 0.58 | 0.41 |
| | DIKU-ModifiedLog-Custom | 43 | 36-37 | 0.807 | 0.670 | 32-35 | 6.44 | 6.44 | 0.27 | 47-48 | 1.17 | 1.17 | 0.06 |
| | DIKU-ModifiedLog-Std | 44 | 38-39 | 0.806 | 0.670 | 32-35 | 6.44 | 6.44 | 0.27 | 47-48 | 1.17 | 1.17 | 0.06 |
| | Sunshine-Std | 45 | 33 | 0.825 | 0.771 | 44-45 | 7.90 | 7.90 | 0.50 | 43-44 | 1.12 | 1.12 | 0.50 |
| | Billabong-UniAV45 | 46 | 49 | 0.720 | 0.616 | 48-49 | 9.22 | 8.82 | 0.29 | 41-42 | 1.09 | 0.99 | 0.45 |
| | Billabong-Uni | 47 | 50 | 0.718 | 0.622 | 48-49 | 9.22 | 8.82 | 0.29 | 41-42 | 1.09 | 0.99 | 0.45 |
| | ATRI-Biostat-JMM | 48 | 42 | 0.779 | 0.710 | 51 | 12.88 | 69.62 | 0.35 | 54 | 1.95 | 5.12 | 0.33 |
| | Billabong-Multi | 49 | 56 | 0.541 | 0.556 | 55 | 27.01 | 19.90 | 0.46 | 40 | 1.07 | 1.07 | 0.45 |
| | ATRI-Biostat-MA | 50 | 47 | 0.741 | 0.671 | 52 | 12.88 | 11.32 | 0.19 | 53 | 1.84 | 5.27 | 0.23 |
| | BIGS2 | 51 | 58 | 0.455 | 0.488 | 50 | 11.62 | 14.65 | 0.50 | 49 | 1.20 | 1.12 | 0.07 |
| | Billabong-MultiAV45 | 52 | 57 | 0.527 | 0.530 | 56 | 28.45 | 21.22 | 0.47 | 45 | 1.13 | 1.07 | 0.47 |
| | ATRI-Biostat-LTJMM | 53 | 55 | 0.636 | 0.563 | 54 | 16.07 | 74.65 | 0.33 | 52 | 1.80 | 5.01 | 0.26 |
| | Threedays | - | 2 | 0.921 | 0.823 | - | - | - | - | - | - | - | - |
| | ARAMIS-Pascal | - | 15 | 0.876 | 0.850 | - | - | - | - | - | - | - | - |
| | IBM-OZ-Res | - | 18 | 0.868 | 0.766 | - | - | - | - | 46 | 1.15 | 1.15 | 0.50 |
| | Orange | - | 44-45 | 0.774 | 0.792 | - | - | - | - | - | - | - | - |
| | SMALLHEADS-NeuralNet | - | 48 | 0.737 | 0.605 | 53 | 13.87 | 13.87 | 0.41 | - | - | - | - |
| | SMALLHEADS-LinMixedEffects | - | - | - | - | 46 | 8.09 | 7.94 | 0.04 | - | - | - | - |
| | Tohka-Ciszek-SMNSR | - | - | - | - | 19 | 5.87 | 5.87 | 0.14 | - | - | - | - |

Table 6.2: Ranked scores for all TADPOLE submissions and benchmarks using the longitudinal prediction data set (D2). Best scores in each category are bolded. Missing numerical entries indicate that submissions did not include forecasts for the corresponding target variable. The “Diagnosis” ranking uses multiclass area under the receiver operating characteristic curve (MAUC), those of ADAS-Cog13 and Ventricles use mean absolute error (MAE). The overall ranking on the left uses the sum of the ranks from the three target variables. The table also lists the secondary metrics: BCA – balanced classification accuracy, WES – weighted error score, CPA – coverage probability accuracy. See Online Methods section 6.5.2 for details on performance metrics.

D2 and D3, see Online Methods section 6.5.1, while for details on performance metrics see Online Methods section 6.5.2. Box-plots showing the distribution of scores, computed on 50 bootstraps of the test set, are shown in Supplementary Fig. 6.SF1, while the distribution of ranks is shown in Supplementary Figs. 6.SF8 – 6.SF10. Among the benchmark methods, *BenchmarkMixedEffectsAPOE* had the best overall rank of 18, obtaining rank 35 on clinical diagnosis prediction, rank 2 on ADAS-Cog13 and rank 23 on Ventricle volume prediction. Removing the APOE status as covariate proved to significantly increase the predictive performance (*BenchmarkMixedEffects*), although we do not show ranks for this entry as it was found during the evaluation phase. Among participant methods, the submission with the best overall rank was *Frog*, obtaining rank 1 for prediction of clinical diagnosis, rank 4 for ADAS-Cog13 and rank 10 for Ventricle volume prediction.

For clinical diagnosis, the best submitted forecasts (team *Frog*) scored better than all benchmark methods, reducing the error of the best benchmark methods by 58% for the multiclass area under the receiver operating characteristic curve (MAUC) and by 38% for balanced classification accuracy (BCA). Here, the best benchmarks obtained a MAUC of 0.846 (*BenchmarkMixedEffects*) and a BCA of 0.792 (*BenchmarkLastVisit*). Among participant methods, *Frog* had the best MAUC score of 0.931, significantly better than all entries other than *Threedays* according to the bootstrap test (p-value = 0.24, see Supplementary section 6.5.4 for details on significance testing). Supplementary Figure 6.SF8 further shows the variability in performance ranking over bootstrap samples and highlights that the top two entries consistently remain at the top of the ranking. In terms of BCA, *ARAMIS-Pascal* had the best score of 0.850. Moreover, ensemble methods (*ConsensusMedian*) achieved the second best MAUC score of 0.925 and the best BCA score of 0.857. In contrast, the best randomised prediction (*RandomisedBest*) achieved a much lower MAUC of 0.800 and a BCA of 0.803, suggesting entries below these scores did not perform significantly better than random guessing according to the bootstrap test (p-value = 0.01). MAUC and BCA performance metrics had a relatively high correlation across all submissions ($r = 0.88$, Supplementary Fig. 6.SF3).

For Ventricle volume, the best submitted forecasts among participants (team *EMC1*) also scored considerably better than all benchmark methods, reducing the error of the best benchmark methods by almost one third (29%) for mean absolute error (MAE) and around one half (51%) for weighted error score (WES). Here, the best benchmark method (*BenchmarkMixedEffects*) had an overall Ventricle MAE and WES of 0.56. Among participant submissions, *EMC1-Std/-Custom* had the best MAE of 0.41 (% ICV), significantly lower than all entries other than *ImaUCL-Covariates/-Std/-half-D1*, *BORREGOTECMTY* and *SBIA* according to the Wilcoxon signed-rank test (see Supplementary section 6.5.4) – this is also confirmed in Supplementary Fig. 6.SF10 by the variability in performance ranking over bootstrap samples. Team *EMC1* also had the best Ventricle WES of 0.29, while *DIKU-ModifiedMri-Custom/-Std*

had the best Ventricle coverage probability accuracy (CPA) of 0.01. Ensemble methods (*ConsensusMean*) achieved the best Ventricle MAE of 0.38. In contrast, the best randomised prediction (*RandomisedBest*) achieved a higher MAE of 0.46, WES of 0.45 and CPA of 0.33. MAE and WES scores showed high correlation ($r = 0.99$, Supplementary Fig. 6.SF3) and were often of equal value for many submissions ($n = 24$), as teams set equal weights for all subjects analysed. CPA did not correlate ($r \approx -0.01$, Supplementary Fig. 6.SF3) with either MAE or WES.

For ADAS-Cog13, the best submitted forecasts did not score significantly better than the simple benchmarks. Here, the simple *BenchmarkMixedEffects* model obtained the second-best MAE of 4.19, which was significantly lower than all other submitted entries according to the Wilcoxon signed-rank test. *BenchmarkMixedEffects* also had the best ADAS-Cog13 WES of 4.19, while *VikingAI-Sigmoid* had the best ADAS-Cog13 CPA of 0.02. Among participants' submissions, *FortuneTellerFish-Control* ranked first in ADAS-Cog13 prediction with a MAE of 4.70, which is 11% higher than the error of the best benchmark. Moreover, all participants' forecasts scored worse than the best randomised prediction (*RandomisedBest*), which here achieved a MAE of 4.52 and WES of 4.52. Nevertheless, the ensemble method *ConsensusMean* obtained the best ADAS scores for MAE (3.75), WES (3.54) and CPA (0.0), which along with *BenchmarkMixedEffects* were the only entries that performed significantly better than random guessing (p-value = 0.01). The MAE and WES scores for ADAS-Cog13 had relatively high correlation ($r = 0.97$, Supplementary Fig. 6.SF3) and were often of equal value for many submissions ($n = 25$). CPA had a weak but significant correlation with MAE ($r = 0.37$, p-value < 0.02) and WES ($r = 0.35$, p-value < 0.02).

6.2.4 Forecasts from the cross-sectional prediction set (D3) and custom prediction sets

Table 6.3 shows the ranking of the forecasts from the cross-sectional D3 prediction set. Box-plots showing the distribution of scores, computed on 50 bootstraps of the test set, are shown in Supplementary Fig 6.SF2, while the distribution of ranks is shown in Supplementary Figs. 6.SF11 – 6.SF13. Due to the lack of longitudinal data, most submissions had lower performance compared to their equivalents from the D2 longitudinal prediction set. Among submitted forecasts, *GlassFrog-Average* had the best overall rank, as well as rank 2-4 on diagnosis prediction, rank 5 on ADAS-Cog13 prediction and rank 3 on ventricle prediction.

For clinical diagnosis prediction on D3, the best prediction among TADPOLE participants (team *IBM-OZ-Res*) scored considerably better than all benchmark methods, reducing the error of the best benchmark method by 40% for MAUC and by 25% for BCA, and achieving error rates comparable to the best predictions from the longitudinal prediction set D2. The best benchmark methods obtained a MAUC of 0.839 (*BenchmarkMixedEffects*) and a

BCA of 0.771 (*BenchmarkLastVisit*). Among participant methods, *IBM-OZ-Res* had the best MAUC score of 0.905, significantly better than all entries other than *GlassFrog-SM/-Average/-LCMEM-HDR*, *BGU-RF/-RFFIX/-LSTM*, *VikingAI-Logistic*, *EMC-EB*, *Rocket* and *Tohka-Ciszek-RandomForestLin* according to the bootstrap hypothesis test (same methodology as in D2). This is further confirmed in Supplementary Fig. 6.SF11 by the variability of ranks under bootstrap samples of the dataset, as these teams often remain at the top of the ranking. *IBM-OZ-Res* also had the best BCA score of 0.830 among participants. Among ensemble methods, *ConsensusMean* obtained the best Diagnosis MAUC of 0.917. In contrast, the best randomised prediction (*RandomisedBest*) obtained an MAUC of 0.811 and a BCA of 0.783. MAUC and BCA performance metrics had a relatively high correlation across all submissions ($r = 0.9$, Supplementary Fig. 6.SF4).

For Ventricle volume prediction on D3, the best prediction (*GlassFrog-LCMEM-HDR*) scored considerably better than all benchmark methods, reducing the error of the best benchmark methods by 58% for MAE and 41% for WES, and achieving error rates comparable to the best predictions of D2. Here, the best benchmark methods had an overall Ventricle MAE of 1.13 (*BenchmarkMixedEffects*) and WES of 0.64 (*BenchmarkLastVisit*). Among participant submissions, *GlassFrog-LCMEM-HDR* had the best MAE of 0.48, significantly lower than all other submitted entries according to the Wilcoxon signed-rank test – this is also confirmed in Supplementary Fig. 6.SF13 by the rank distribution under dataset bootstraps. *GlassFrog-LCMEM-HDR* also had the best Ventricle WES of 0.38, while submissions by team *DIKU* had the best Ventricle CPA of 0.05. Among ensemble methods, *ConsensusMedian* obtained a Ventricle MAE of 0.71 (4th best) and WES of 0.65 (7th best). In contrast, the best randomised prediction (*RandomisedBest*) obtained a Ventricle MAE of 0.92, WES of 0.50 and CPA of 0. As in D2, MAE and WES scores in D3 for Ventricles had very high correlation ($r = 0.99$, Supplementary Fig. 6.SF4), while CPA showed weak correlation with MAE ($r = 0.24$, p-value = 0.17) and WES ($r = 0.37$, p-value < 0.032).

For ADAS-Cog13 on D3, the predictions submitted by participants again did not perform better than the best benchmark methods. *BenchmarkMixedEffects* had the best MAE of 4.23, which was significantly lower than all entries by other challenge participants. Moreover, the MAE of 4.23 was only marginally worse than the equivalent error (4.19) by the same model on D2. *BenchmarkMixedEffects* also had the best ADAS-Cog13 WES of 4.23, while *ATRI-Biostat-MA* had the best ADAS-Cog13 CPA of 0.04. Among participants' submissions, *CyberBrains* ranked first in ADAS-Cog13 prediction with a MAE of 4.72, an error 11% higher than the best benchmark. Among ensemble methods, *ConsensusMean* obtained an ADAS-Cog13 MAE of 4.58, WES of 4.34, better than all participants' entries. As in D2, the best randomised predictions (*RandomisedBest*) obtained an ADAS-Cog13 MAE of 4.54 (2nd best) and WES of 4.50 (3rd best). As in D2, MAE and WES scores for ADAS-Cog13 had high correlation ($r = 0.97$, Supplementary Fig. 6.SF4), while CPA showed weak, non-significant correlation with

MAE ($r = 0.34$, $p\text{-value} \approx 0.052$) or WES ($r = 0.33$, $p\text{-value} \approx 0.057$).

Results on the custom prediction sets are presented in Supplementary Table 6.ST1.

6.2.5 Algorithm characteristics associated with increased performance

To understand what characteristics of algorithms could have yielded higher performance, we show in Figure 6.1 associations from a general linear model between predictive performance and feature selection methods, different types of features, methods for data imputation, and methods for forecasting of target variables. For each type of feature/method and each target variable (clinical diagnosis, ADAS-Cog13 and Ventricles), we show the distribution of estimated coefficients from a general linear model, derived from the approximated inverse Hessian matrix at the maximum likelihood estimator (see Online Methods section 6.5.3). From this analysis we removed outliers, defined as submissions with ADAS MAE higher than 10 and Ventricle MAE higher than 1.15 (%ICV). For all plots, distributions to the right of the gray dashed vertical line denote increased performance compared to baseline (i.e. when those characteristics are not used).

For feature selection, Figure 6.1 shows that methods with manual selection of features tend to be associated with better predictive performance in ADAS-Cog13 and Ventricles. In terms of feature types, CSF and DTI features were generally associated with an increase in predictive performance for clinical diagnosis, while augmented features were associated with performance improvements for ventricle prediction. In terms of data imputation methods, while some differences can be observed, no clear conclusions can be drawn. In terms of prediction models, the only positive association that indicates increased performance is in the neural networks for ventricle prediction. However, given the small number of methods tested (<50) and the large number of degrees of freedom, these results should be interpreted with care.

6.3 Discussion

In this work, we presented the results of the TADPOLE Challenge. The results of the challenge provide important insights into the current state of the field, and how well current algorithms can predict progression of AD diagnoses and markers of disease progression both from rich longitudinal data sets and, comparatively, from sparser cross-sectional data sets typical of a clinical trial scenario. The challenge further highlights the algorithms, features and data-handling strategies that tend to lead to improved forecasts. In the following sections we discuss the key conclusions that we draw from our study and highlight important limitations.

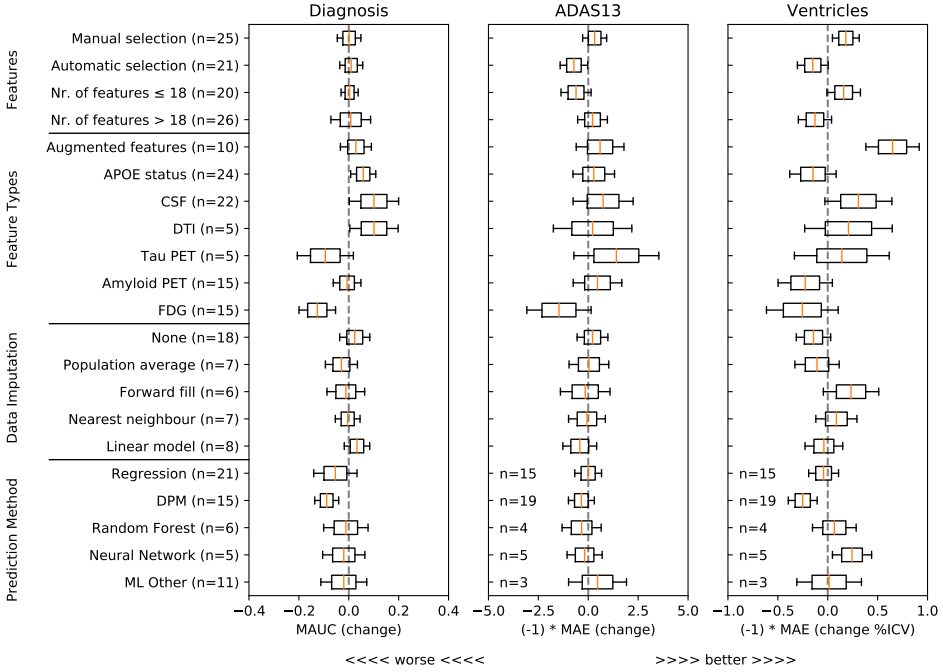


Figure 6.1: Associations between the prediction of clinical diagnosis, ADAS-Cog13 and Ventricle volume and different strategies of (top) feature selection, (upper-middle) types of features, (lower-middle) data imputation strategies and (bottom) prediction methods for the target variables. For each type of feature/method (rows) and each target variable (columns), we show the distribution of estimated coefficients from a general linear model. Positive coefficients, where distributions lie to the right of the dashed vertical line, indicate better performance than baseline (vertical dashed line). For ADAS-Cog13 and Ventricle prediction, we flipped the sign of the coefficients, to consistently show better performance to the right of the vertical line.

6.3.1 TADPOLE pushed forward performance on AD clinical diagnosis prediction

When comparing to previous state-of-the-art results in the literature, the best TADPOLE methods show similar or higher performance in AD diagnostic classification while also tackling a harder problem than most previous studies of predicting future, rather than estimating current, classification. A comparison of 15 studies presented by [262] reported lower performance (maximum AUC of 0.902 vs 0.931 obtained by the best TADPOLE method) for the simpler two-class classification problem of separating MCI-stable from MCI-converters in ADNI. A more recent method by [263] reported a maximum AUC of 0.932 and accuracy of 0.88 at the same MCI-stable vs -converter classification task. However, a) TADPOLE's discrimination of CN-converters from CN-stable subjects is harder as disease signal is weaker at such early stages, and b) the predictive performance drops in three-class problems like TADPOLE compared to two-class. Furthermore, the best out of 19 algorithms

in the CADDementia Challenge [258] obtained an MAUC of 0.78.

We are unaware of previous studies forecasting future ventricle volume or ADAS-Cog13, so TADPOLE sets a new benchmark state-of-the-art performance on these important prediction tasks.

6.3.2 No one-size-fits-all prediction algorithm

The results on the longitudinal D2 prediction set suggest no clear winner on predicting all target variables – no single method performed well on all tasks. While *Frog* had the best overall submission with the lowest sum of ranks, for each performance metric individually we had different winners: *Frog* (clinical diagnosis MAUC of 0.931), *ARAMIS-Pascal* (clinical diagnosis BCA of 0.850), *BenchmarkMixedEffects* (ADAS-Cog13 MAE and WES of 4.19), *VikingAI-Sigmoid* (ADAS-Cog13 CPA of 0.02), *EMC1-Std/EMC1-Custom* (ventricle MAE of 0.41 and WES of 0.29), and *DIKU-ModifiedMri-Std/-Custom* (ventricle CPA of 0.01). Moreover, on the cross-sectional D3 prediction set, the methods by *Glass-Frog* had the best performance. Associations of method-type with increased performance in Fig. 6.1 confirm no clear *increase* in performance for any types of prediction methods (with the exception of neural networks for ventricle volume prediction). This raises an important future challenge to algorithm designers to develop methods able to perform well on multiple forecasting tasks and also in situations with limited data, such as D3.

6.3.3 Ensemble methods perform strongly

Consistently strong results from ensemble methods (*ConsensusMean/ConsensusMedian* outperformed all others on most tasks) might suggest that different methods over-estimate future measurements for all subjects while others under-estimate them, likely due to the underlying assumptions they make. This is confirmed by plots of the difference between true and estimated measures (Supplementary Figures 6.SF5–6.SF7), where most methods systematically under- or over-estimate in *all subjects*. However, even if methods were completely unbiased, averaging over all methods could also help predictions by reducing the variance in the estimated target variables.

6.3.4 Predictability of ADAS-Cog13 scores

ADAS-Cog13 scores were more difficult to forecast than clinical diagnosis or ventricle volume. The only single method able to forecast ADAS-Cog13 better than informed random guessing (*RandomisedBest*) was the *BenchmarkMixedEffects*, a simple mixed effects model with no covariates and age as a regressor. The difficulty could be due to variability in administering the tests or practice effects. A useful target performance level comes from the 4 points change

generally used to identify responders to a drug treatment [264]. With the exception of the ensemble method, all submitted forecasts failed to produce mean error below 4, highlighting the substantial challenge of estimating change in ADAS-Cog13 over the 1.4 year interval – although over longer time periods, non-trivial forecasts are likely to improve in comparison to *RandomisedBest*, which is independent of time period. Nevertheless, for the longitudinal D2 prediction set, the MAE in ADAS-Cog13 from *ConsensusMean* was 3.75, which restores hope in forecasting cognitive score trajectories even over relatively short timescales.

6.3.5 Prediction errors from limited cross-sectional dataset mimicking clinical trials are similar to those from longitudinal dataset

For clinical diagnosis, the best performance on the limited, cross-sectional D3 prediction set was similar to the best performance on the D2 longitudinal prediction set: 0.917 vs 0.931 for MAUC (p-value = 0.14), representing a 3% error increase for D3 compared to D2. Slightly larger and significant differences were observed for ADAS MAE (3.75 vs 4.23, p-value < 0.01) and Ventricle MAE (0.38 vs 0.48, p-value < 0.01). It should be noted that Ventricle predictions for D3 were extremely difficult, given that only 25% of subjects to be forecasted had MRI data in D3. This suggests that, for clinical diagnosis, current forecast algorithms are reasonably robust to lack of longitudinal data and missing inputs, while for ADAS and Ventricle volume prediction, some degree of performance is lost. Future work is also required to determine the optimal balance of input data quality and quantity versus cost of acquisition.

6.3.6 DTI and CSF features appear informative for clinical diagnosis prediction, augmented features appear informative for ventricle prediction

DTI and CSF features are most associated with increases in clinical diagnosis forecast performance. CSF, in particular, is well established as an early marker of AD [5] and likely to help predictions for early-stage subjects, while DTI, measuring microstructure damage, may be informative for middle-stage subjects. On the other hand, for prediction of ventricle volume, augmented features had the highest association with increases in prediction performance. Future work is required to confirm the added value of these features and others in a more systematic way.

6.3.7 Challenge design and limitations

TADPOLE Challenge has several limitations that future editions of the challenge may consider addressing. One limitation is the reliability of the three target variables: clinical diagnosis, ADAS-Cog13 and Ventricle volume. First of all, clinical diagnosis has only moderate agreement with gold-standard neuropathological post-mortem diagnosis. In particular, one study [265] has shown that a clinical diagnosis of probable AD has sensitivity between

70.9% and 87.3% and specificity between 44.3% and 70.8%. With the advent of post-mortem confirmation in ADNI, future challenges might address this by evaluating the algorithms on subjects with pathological confirmation. Similarly, ADAS-Cog13 is known to suffer from low reliability across consecutive visits [264], and TADPOLE algorithms fail to forecast it reliably. However, this might be related to the short time-window (1.4 years), and more accurate predictions might be possible over longer time-windows, when there is more significant cognitive decline. Ventricle volume measurements depend on MRI scanner factors such as field strength, manufacturer and pulse sequences [266], although these effects have been removed to some extent by ADNI through data preprocessing and protocol harmonization. TADPOLE Challenge also assumes all subjects either remain stable or convert to Alzheimer’s disease, whereas in practice some of them might develop other types of neurodegenerative diseases.

For performance evaluation, we elected to use very simple yet reliable metrics as the primary performance scores: the multiclass area under the curve (mAUC) for the clinical categorical variable and the mean absolute error (MAE) for the two numerical variables. While the mAUC accounts for decision confidence, the MAE does not, which means that the confidence intervals submitted by participants do not contribute to the rankings computed in Tables 6.2 and 6.3. While the weighted error score (WES) takes confidence intervals into account, we consider it susceptible to “hacking”, e.g. participants might assign high confidence to only one or two data points and thereby skew the score to ignore most of the predictions – in practice, we did not observe this behaviour in any submission. For clinical relevance, we believe that confidence intervals are an extremely important part of such predictions and urge future studies to consider performance metrics that require and take account of participant-calculated confidence measures.

TADPOLE has other limitations related to the algorithms’ comparability and generalisability. First of all, the evaluation and training were both done on data collected by ADNI – in future work, we plan to assess how the models will generalise on different datasets. Another limitation is that we can only compare full methods submissions and not different types of features, and strategies for data imputation and prediction used within the full method. While we tried to evaluate the effect of these characteristics in Figure 6.1, in practice the numbers were small and hence most effects did not reach statistical significance. Moreover, the challenge format does not provide an exhaustive comparison of all combinations of data processing, predictive model, features, etc., so does not lead to firm conclusions on the best combinations but rather provides hypotheses for future testing. In future work, we plan to test inclusion of features and strategies for data imputation and prediction independently, by changing one such characteristic at a time.

Another limitation is that the number of controls and MCI converters in the D4 test set is

low (9 MCI converters and 9 control converters). However, these numbers will increase over time as ADNI acquires more data, and we plan to re-run the evaluation at a later stage with the additional data acquired after April 2019. A subsequent evaluation will also enable us to evaluate the TADPOLE methods on longer time-horizons, over which the effects of putative drugs would be higher.

6.4 Conclusion

In this work we presented the results of the TADPOLE Challenge. The results of the challenge provide important insights into the current state of the art in AD forecasting, such as performance levels achievable with current data and technology as well as specific algorithms, features and data-handling strategies that support the best forecasts. The developments and outcomes of TADPOLE Challenge can aid refinement of cohorts and endpoint assessment for clinical trials, and can support accurate prognostic information in clinical settings. The challenge website (<https://tadpole.grand-challenge.org>) will stay open for submissions, which can be added to our current ranking. The open test set remains available on the ADNI LONI website and also allows individual participants to evaluate future submissions. Through TADPOLE-SHARE <https://tadpole-share.github.io/>, we further plan to implement many TADPOLE methods in a common framework, to be made publicly available. TADPOLE provides a standard benchmark for evaluation of future AD prediction algorithms.

| | | Overall | Diagnosis | | ADAS-Cog13 | | | | Ventricles (% ICV) | | | | |
|------------------------------|----------------------------|---------|-----------|-------|------------|-------|-------|------|--------------------|-------|------|------|------|
| Submission | | Rank | Rank | MAUC | BCA | Rank | MAE | WES | CPA | Rank | MAE | WES | CPA |
| Tohka-Ciszek-RandomForestLin | ConsensusMean | - | - | 0.917 | 0.821 | - | 4.58 | 4.34 | 0.12 | - | 0.73 | 0.72 | 0.09 |
| | ConsensusMedian | - | - | 0.905 | 0.817 | - | 5.44 | 5.37 | 0.19 | - | 0.71 | 0.65 | 0.10 |
| | GlassFrog-Average | 1 | 2-4 | 0.897 | 0.826 | 5 | 5.86 | 5.57 | 0.25 | 3 | 0.68 | 0.55 | 0.24 |
| | GlassFrog-LCMEM-HDR | 2 | 2-4 | 0.897 | 0.826 | 9 | 6.57 | 6.56 | 0.34 | 1 | 0.48 | 0.38 | 0.24 |
| | GlassFrog-SM | 3 | 2-4 | 0.897 | 0.826 | 4 | 5.77 | 5.77 | 0.19 | 9 | 0.82 | 0.55 | 0.07 |
| | RandomisedBest | - | - | 0.811 | 0.783 | - | 4.54 | 4.50 | 0.26 | - | 0.92 | 0.50 | 0.00 |
| | ImaUCL-Std | 5-9 | 12-14 | 0.854 | 0.698 | 16-18 | 6.95 | 6.93 | 0.05 | 5-7 | 0.81 | 0.81 | 0.22 |
| | ImaUCL-Covariates | 5-9 | 12-14 | 0.854 | 0.698 | 16-18 | 6.95 | 6.93 | 0.05 | 5-7 | 0.81 | 0.81 | 0.22 |
| | ImaUCL-halfD1 | 5-9 | 12-14 | 0.854 | 0.698 | 16-18 | 6.95 | 6.93 | 0.05 | 5-7 | 0.81 | 0.81 | 0.22 |
| | Rocket | 5-9 | 10 | 0.865 | 0.771 | 3 | 5.27 | 5.14 | 0.39 | 23 | 1.06 | 1.06 | 0.27 |
| | VikingAI-Logistic | 5-9 | 8 | 0.876 | 0.768 | 6 | 5.94 | 5.91 | 0.22 | 22 | 1.04 | 1.01 | 0.18 |
| | EMC1-Std | 10 | 30 | 0.705 | 0.567 | 7 | 6.29 | 6.19 | 0.47 | 4 | 0.80 | 0.62 | 0.48 |
| | BenchmarkMixedEffects | - | - | 0.839 | 0.728 | - | 4.23 | 4.23 | 0.34 | - | 1.13 | 1.13 | 0.50 |
| | SBIA | 11 | 28 | 0.779 | 0.782 | 10 | 6.63 | 6.43 | 0.40 | 8 | 0.82 | 0.75 | 0.18 |
| | BGU-LSTM | 12-14 | 5-7 | 0.877 | 0.776 | 13-15 | 6.75 | 6.17 | 0.39 | 26-28 | 1.11 | 0.79 | 0.17 |
| | BGU-RFFIX | 12-14 | 5-7 | 0.877 | 0.776 | 13-15 | 6.75 | 6.17 | 0.39 | 26-28 | 1.11 | 0.79 | 0.17 |
| | BGU-RF | 12-14 | 5-7 | 0.877 | 0.776 | 13-15 | 6.75 | 6.17 | 0.39 | 26-28 | 1.11 | 0.79 | 0.17 |
| | BravoLab | 15 | 18 | 0.813 | 0.730 | 28 | 8.02 | 8.02 | 0.47 | 2 | 0.64 | 0.64 | 0.42 |
| | BORREGOTECMTY | 16-17 | 15 | 0.852 | 0.748 | 8 | 6.44 | 5.86 | 0.46 | 30 | 1.14 | 1.02 | 0.49 |
| | CyberBrains | 16-17 | 17 | 0.830 | 0.755 | 1 | 4.72 | 4.72 | 0.21 | 35 | 1.54 | 1.54 | 0.50 |
| | ATRI-Biostat-MA | 18 | 19 | 0.799 | 0.772 | 26 | 7.39 | 6.63 | 0.04 | 11 | 0.93 | 0.97 | 0.10 |
| | DIKU-GeneralisedLog-Std | 19-20 | 20 | 0.798 | 0.684 | 20-21 | 6.99 | 6.99 | 0.17 | 16-17 | 0.95 | 0.95 | 0.05 |
| | EMC-EB | 19-20 | 9 | 0.869 | 0.765 | 27 | 7.71 | 7.91 | 0.50 | 21 | 1.03 | 1.07 | 0.49 |
| | DIKU-GeneralisedLog-Custom | 21 | 21 | 0.798 | 0.681 | 20-21 | 6.99 | 6.99 | 0.17 | 16-17 | 0.95 | 0.95 | 0.05 |
| | DIKU-ModifiedLog-Std | 22-23 | 22-23 | 0.798 | 0.688 | 22-25 | 7.10 | 7.10 | 0.17 | 12-15 | 0.95 | 0.95 | 0.05 |
| | DIKU-ModifiedMri-Std | 22-23 | 22-23 | 0.798 | 0.688 | 22-25 | 7.10 | 7.10 | 0.17 | 12-15 | 0.95 | 0.95 | 0.05 |
| | DIKU-ModifiedMri-Custom | 24-25 | 24-25 | 0.798 | 0.691 | 22-25 | 7.10 | 7.10 | 0.17 | 12-15 | 0.95 | 0.95 | 0.05 |
| DIKU-ModifiedLog-Custom | 24-25 | 24-25 | 0.798 | 0.691 | 22-25 | 7.10 | 7.10 | 0.17 | 12-15 | 0.95 | 0.95 | 0.05 | |
| Billabong-Uni | 26 | 31 | 0.704 | 0.626 | 11-12 | 6.69 | 6.69 | 0.38 | 19-20 | 0.98 | 0.98 | 0.48 | |
| Billabong-UniAV45 | 27 | 32 | 0.703 | 0.620 | 11-12 | 6.69 | 6.69 | 0.38 | 19-20 | 0.98 | 0.98 | 0.48 | |
| ATRI-Biostat-JMM | 28 | 26 | 0.794 | 0.781 | 29 | 8.45 | 8.12 | 0.34 | 18 | 0.97 | 1.45 | 0.37 | |
| CBIL | 29 | 16 | 0.847 | 0.780 | 33 | 10.99 | 11.65 | 0.49 | 29 | 1.12 | 1.12 | 0.39 | |
| BenchmarkLastVisit | 30 | 27 | 0.785 | 0.771 | 19 | 6.97 | 7.07 | 0.42 | 33 | 1.17 | 0.64 | 0.11 | |
| Billabong-MultiAV45 | 31 | 33 | 0.682 | 0.603 | 30-31 | 9.30 | 9.30 | 0.43 | 24-25 | 1.09 | 1.09 | 0.49 | |
| Billabong-Multi | 32 | 34 | 0.681 | 0.605 | 30-31 | 9.30 | 9.30 | 0.43 | 24-25 | 1.09 | 1.09 | 0.49 | |
| ATRI-Biostat-LTJMM | 33 | 29 | 0.732 | 0.675 | 34 | 12.74 | 63.98 | 0.37 | 32 | 1.17 | 1.07 | 0.40 | |
| BenchmarkSVM | 34 | 36 | 0.494 | 0.490 | 32 | 10.01 | 10.01 | 0.42 | 31 | 1.15 | 1.18 | 0.50 | |
| DIVE | 35 | 35 | 0.512 | 0.498 | 35 | 16.66 | 16.74 | 0.41 | 34 | 1.42 | 1.42 | 0.34 | |
| IBM-OZ-Res | - | 1 | 0.905 | 0.830 | - | - | - | - | 36 | 1.77 | 1.77 | 0.50 | |

Table 6.3: Ranked prediction scores for all TADPOLE submissions that used the cross-sectional prediction data set (D3). Best scores in each category are bolded. Missing numerical entries indicate that submissions did not include predictions for the corresponding target variable. The “Diagnosis” ranking uses multiclass area under the receiver operating characteristic curve (MAUC), those of ADAS-Cog13 and Ventricles use mean absolute error (MAE). The overall ranking on the left uses the sum of the ranks from the three target variables. The table also lists the secondary metrics: BCA – balanced classification accuracy, WES – weighted error score, CPA – coverage probability accuracy.

6.5 Online Methods – Challenge design and prediction algorithms

6.5.1 Data

The challenge uses data from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) [260]. Specifically, the TADPOLE Challenge made four key data sets available to the challenge participants:

- **D1:** The TADPOLE standard training set draws on longitudinal data from the entire ADNI history. The data set contains measurements for every individual that has provided data to ADNI in at least two separate visits (different dates) across three phases of the study: ADNI1, ADNI GO, and ADNI2.
- **D2:** The TADPOLE longitudinal prediction set contains as much available data as we could gather from the ADNI rollover individuals for whom challenge participants are asked to provide predictions. D2 includes data from all available time-points for these individuals. It defines the set of individuals for which participants are required to provide forecasts.
- **D3:** The TADPOLE cross-sectional prediction set contains a single (most recent) time point and a limited set of variables from each rollover individual in D2. Although we expect worse predictions from this data set than D2, D3 represents the information typically available when selecting a cohort for a clinical trial.
- **D4:** The TADPOLE test set contains visits from ADNI rollover subjects that occurred after 1 Jan 2018 and contain at least one of the three outcome measures: diagnostic status, ADAS-Cog13 score, or ventricle volume.

While participants were free to use any training datasets they wished, we provided the D1-D3 datasets in order to remove the need for participants to pre-process the data themselves, and also to be able to evaluate the performance of different algorithms on the same standardised datasets. Participants that used custom training data sets were asked also to submit results using the standard training data sets to enable direct performance comparison. We also included the D3 cross-sectional prediction set in order to simulate a clinical trial scenario. For information on how we created the D1-D4 datasets, see Supplementary section 6.5.4. The software code used to generate the standard datasets is openly available on Github: <https://github.com/noxtoby/TADPOLE>.

Table 6.4 shows the demographic breakdown of each TADPOLE data set as well as the proportion of biomarker data available in each dataset. Many entries are missing data, especially for certain biomarkers derived from exams performed on only subsets of subjects, such as tau imaging (AV1451). D1 and D2 also included demographic data typically available in ADNI (e.g. education, marital status) as well as standard genetic markers (e.g. Alipoprotein E – APOE epsilon 4 status).

| Demographics | | | | | |
|--|-------------------------|--------------|--------------|-------------|-------------|
| | | D1 | D2 | D3 | D4 |
| Overall number of subjects | | 1667 | 896 | 896 | 219 |
| Controls [†] | Number (% all subjects) | 508 (30.5%) | 369 (41.2%) | 299 (33.4%) | 94 (42.9%) |
| | Visits per subject | 8.3 ± 4.5 | 8.5 ± 4.9 | 1.0 ± 0.0 | 1.0 ± 0.2 |
| | Age | 74.3 ± 5.8 | 73.6 ± 5.7 | 72.3 ± 6.2 | 78.4 ± 7.0 |
| | Gender (% male) | 48.6% | 47.2% | 43.5% | 47.9% |
| | MMSE | 29.1 ± 1.1 | 29.0 ± 1.2 | 28.9 ± 1.4 | 29.1 ± 1.1 |
| | Converters* | 18 | 9 | - | - |
| MCI [†] | Number (% all subjects) | 841 (50.4%) | 458 (51.1%) | 269 (30.0%) | 90 (41.1%) |
| | Visits per subject | 8.2 ± 3.7 | 9.1 ± 3.6 | 1.0 ± 0.0 | 1.1 ± 0.3 |
| | Age | 73.0 ± 7.5 | 71.6 ± 7.2 | 71.9 ± 7.1 | 79.4 ± 7.0 |
| | Gender (% male) | 59.3% | 56.3% | 58.0% | 64.4% |
| | MMSE | 27.6 ± 1.8 | 28.0 ± 1.7 | 27.6 ± 2.2 | 28.1 ± 2.1 |
| | Converters* | 117 | 37 | - | 9 |
| AD [†] | Number (% all subjects) | 318 (19.1%) | 69 (7.7%) | 136 (15.2%) | 29 (13.2%) |
| | Visits per subject | 4.9 ± 1.6 | 5.2 ± 2.6 | 1.0 ± 0.0 | 1.1 ± 0.3 |
| | Age | 74.8 ± 7.7 | 75.1 ± 8.4 | 72.8 ± 7.1 | 82.2 ± 7.6 |
| | Gender (% male) | 55.3% | 68.1% | 55.9% | 51.7% |
| | MMSE | 23.3 ± 2.0 | 23.1 ± 2.0 | 20.5 ± 5.9 | 19.4 ± 7.2 |
| | Converters* | - | - | - | 9 |
| Number of clinical visits for all subjects with data available (% of total visits) | | | | | |
| | | D1 | D2 | D3 | D4 |
| Cognitive | | 8862 (69.9%) | 5218 (68.1%) | 753 (84.0%) | 223 (95.3%) |
| MRI | | 7884 (62.2%) | 4497 (58.7%) | 224 (25.0%) | 150 (64.1%) |
| FDG-PET | | 2119 (16.7%) | 1544 (20.2%) | - | - |
| AV45 | | 2098 (16.6%) | 1758 (23.0%) | - | - |
| AV1451 | | 89 (0.7%) | 89 (1.2%) | - | - |
| DTI | | 779 (6.1%) | 636 (8.3%) | - | - |
| CSF | | 2347 (18.5%) | 1458 (19.0%) | - | - |

Table 6.4: Summary of TADPOLE datasets D1-D4. ([†]) Diagnosis at first visit with available data. For D3 and D4, 192 and 6 subjects respectively did not have a diagnosis at any clinical visit, so numbers don't add up to 100%. (*) For D4, converters are ADNI3 subjects who are MCI, but were previously CN, or who are AD, but were previously CN or MCI in their last visit in ADNI2. For D1, D2 and D3, converters are CN or MCI at their earliest available visit, who progress to a later classification of MCI/AD within 1.4 years (same duration as D4)

6.5.2 Forecast Evaluation

For evaluation of clinical status predictions, we used similar metrics to those that proved effective in the CADDementia challenge [258]: (i) the multiclass area under the receiver operating characteristic curve (MAUC) and (ii) the overall balanced classification accuracy (BCA). For ADAS-Cog13 and ventricle volume, we used three metrics: (i) mean absolute error (MAE), weighted error score (WES) and coverage probability accuracy (CPA). BCA and MAE focus purely on prediction accuracy ignoring confidence, MAUC and WES account for accuracy and confidence, while CPA assesses the confidence interval only. The formulas for each performance metric are summarised in Table 6.5. See the TADPOLE white paper [69] for further rationale for choosing these performance metrics. In order to characterise the distribution of these metric scores, we compute scores based on 50 bootstraps with replacement on the test dataset.


| Formula | Definitions |
|--|--|
| $\frac{1}{L(L-1)} \sum_{i=2}^L \sum_{j=1}^i \hat{A}(c_i c_j) + \hat{A}(c_j c_i)$ <p>where $\hat{A}(c_i c_j) = \frac{S_i - n_i(n_i+1)/2}{n_i n_j}$</p> | n_i, n_j – number of points from class i and j . S_{ij} – 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 . L – number of classes. c_i – class i . |
| $BCA = \frac{1}{2L} \sum_{i=1}^L \left[\frac{TP_i}{TP_i + FN_i} + \frac{TN_i}{TN_i + FP_i} \right]$ | TP_i, FP_i, TN_i, FN_i – the number of true positives, false positives, true negatives and false negatives for class i . L – number of classes |
| $MAE = \frac{1}{N} \sum_{i=1}^N \tilde{M}_i - M_i $ | M_i is the actual value in individual i in future data. \tilde{M}_i is the participant’s best guess at M_i and N is the number of data points |
| $WES = \frac{\sum_{i=1}^N \tilde{C}_i \tilde{M}_i - M_i }{\sum_{i=1}^N \tilde{C}_i}$ | M_i, \tilde{M}_i and N defined as above. $\tilde{C}_i = (C_+ - C_-)^{-1}$, where $[C_-, C_+]$ is the 50% confidence interval |
| $CPA = ACP - 0.5 $ | actual coverage probability (ACP) - the proportion of measurements that fall within the 50% confidence interval. |

Table 6.5: TADPOLE performance metric formulas and definitions for the terms.

6.5.3 Statistical Analysis of Method Attributes with Performance

To identify which features and types of algorithms enable good predictions, we annotated each TADPOLE submission with a set of 21 attributes related to (i) feature selection (manual/automatic and large vs. small number of features), (ii) feature types (e.g. “uses Amyloid PET”), (iii) strategy for data imputation (e.g. “patient-wise forward-fill”) and (iv) prediction method (e.g. “neural network”) for clinical diagnosis and ADAS/Ventricles separately. To understand which of these annotations were associated with increased performance, we applied a general linear model [267], $Y = X\beta + \epsilon$, where Y is the performance metric (e.g. diagnosis MAUC), X is the nr_submissions \times 21 design matrix of binary annotations, and β show the contributions of each of the 21 attributes towards achieving the performance measure Y .

6.5.4 Prediction Algorithms*

Team: EMC1  (Members: Vikram Venkatraghavan, Esther Bron, Stefan Klein, Institution: Erasmus MC, The Netherlands)

Feature selection: Automatic – Only the subjects who had converted to AD were used for feature selection. Features with the largest changes over time after correcting for age, gender, education and ICV were selected

Selected features: 250 features from the set of FDG, AV45, DTI, MRI (cross-sectional Freesurfer volumes), Arterial Spin Labelling (ASL) MRI, CSF and cognitive tests.

*This part of the manuscript has been edited to provide only the details of Vikram Venkatraghavan’s algorithm. For details of the other submitted methods, please read the pre-print of the manuscript published on arxiv: <https://arxiv.org/abs/2002.03419>

Missing data: Imputed using nearest-neighbour interpolation. For D2, visits with missing diagnosis were excluded. For the D3 subjects with no known diagnosis, this was estimated using a nearest-neighbour search based on disease severity

Confounder correction: Corrected for age, gender, education and ICV using linear regression based on data from controls.

Method category: Data-driven disease progression model and machine learning

Prediction method: Authors hypothesize that aging and progression of AD are the primary causes for the change in biomarker values with time and that these changes eventually lead to a change in clinical status. To predict biomarker values at future timepoints, the rate of AD progression is estimated in each subject. This is followed by estimating the interactions of aging and AD progression in the progression of different biomarkers. Lastly, authors use the biomarkers estimated at the future timepoint to predict the change in clinical status. These steps are elaborated below:

Rate of Progression of AD: To assess the severity of AD, we estimated the sequence in which the selected features became abnormal in AD using a Discriminative Event-Based Model [93] and used it to estimate the disease severity at all the timepoints for each subject. A linear mixed effect model was fit to estimate the rate of change of disease severity for different subjects. This model was used for predicting the disease severity at all the future timepoints.

Interactions of aging and AD progression: For predicting the biomarker values at the future timepoint, we fit linear mixed effect models for each biomarker considering interactions between the estimated disease severity and age, with gender and ICV as additional covariates. This model was used to forecast the future values of all 250 selected features, including ADAS-Cog13 scores and Ventricle volumes.

Predicting the change in clinical status: For the diagnosis prediction, the forecasted values of the biomarkers and the last known clinical diagnosis of the subject were used as inputs for a soft-margin SVM classifier with a radial basis function kernel. Two separate submissions were made:

- EMC1-Std (ID 1): ASL based features were excluded in this model
- EMC1-Custom (ID 2): ASL based features were included in this model

Confidence Intervals: Standard errors of the predicted values of Ventricles and ADAS-Cog-13 were estimated by repeating the prediction procedure, including the estimation of disease severity, for 10 repetitions of bootstrap sampling. These standard errors were used to define the confidence intervals.

Publication link: <https://doi.org/10.1016/j.neuroimage.2018.11.024>

Repository link:

https://github.com/88vikram/TADPOLE_submission_with_debm

Supplementary material

Creating the D1-D4 datasets

The data used from ADNI consists of: (1) CSF markers of amyloid-beta and tau deposition; (2) various imaging modalities such as magnetic resonance imaging (MRI), positron emission tomography (PET) using several tracers: Fluorodeoxyglucose (FDG, hypometabolism), AV45 (amyloid), AV1451 (tau) as well as diffusion tensor imaging (DTI); (3) cognitive assessments acquired in the presence of a clinical expert; (4) genetic information such as apolipoprotein E4 (APOE4) status extracted from DNA samples; and (5) general demographic information. Extracted features from this data were merged together into a final spreadsheet and made available on the LONI ADNI website.

The imaging data has been pre-processed with standard ADNI pipelines. For MRI scans, this included correction for gradient non-linearity, B1 non-uniformity correction and peak sharpening. [ADNI MRI pre-processing]. Meaningful regional features such as volume and cortical thickness were extracted using the Freesurfer cross-sectional and longitudinal pipelines [268]. Each PET image (FDG, AV45, AV1451) had their frames co-registered, averaged across the six five-minute frames, standardised with respect to the orientation and voxel size, and smoothed to produce a uniform resolution of 8mm full-width/half-max (FWHM) (see <http://adni.loni.usc.edu/methods/pet-analysis/pre-processing/>). Standardised uptake value ratio (SUVR) measures for relevant regions-of-interest were extracted after registering the PET images to corresponding MR images using the SPM5 software [269]. Further details have been provided in the ADNI procedures manual. DTI scans were corrected for head motion and eddy-current distortion, skull-stripped, EPI-corrected, and finally aligned to the T1 scans using the pipeline from [270]. Diffusion tensor summary measures were estimated based on the Eve white-matter atlas [271].

In addition to the standard datasets, we also created three leaderboard datasets LB1, LB2 and LB3 which mimic the D1, D2 and D4 datasets. These datasets were used by participants to preliminarily evaluate their algorithms before the competition deadline, and to compare their results on the leaderboard system (<https://tadpole.grand-challenge.org/Leaderboard/>).

Statistical testing

Differences in MAUC scores

For analysing whether the MAUC scores obtained by top algorithms are significantly different, we performed a bootstrapped hypothesis test [272], since the significance test for

comparing two AUC scores [112] does not extend to multiple classes. For two TADPOLE entries A and B , where A scored better than B on the full D4 test set, we want to confirm either the null hypothesis $H_0 : \text{mauc}(A) = \text{mauc}(B)$ or the alternative hypothesis $H_1 : \text{mauc}(A) > \text{mauc}(B)$. We then proceed as follows:

- Sample $N = 50$ random bootstraps D_i of the D4 test set with replacement.
- Compute the $\text{mauc}(A, D_i)$ and $\text{mauc}(B, D_i)$ based on the bootstrapped dataset. Repeat for all N bootstraps.
- Compute the p-value as $\sum_i I[\text{mauc}(A, D_i) < \text{mauc}(B, D_i)]/N$, which is the proportion of bootstrapped datasets where A performed worse than B .
- Accept/reject H_0 based on a 5% significance level.

Differences in MAE scores

For comparing differences in MAE scores, we applied the non-parametric Wilcoxon signed-rank test on paired samples of absolute errors across all visits of the D4 subjects. We chose the non-parametric Wilcoxon test because the input samples are not normally distributed, as they represent absolute errors and are always positive. We report results based on a 5% significance level.

Differences between D2 and D3 forecasts

For comparing differences between the scores obtained by two algorithms on D2 vs D3 forecasts, we use an approach similar to comparing MAUC scores (section 8.4.2).

Supplementary Results







| Submission | Overall Rank | Diagnosis | | | ADAS-Cog13 | | | | Ventricles (% ICV) | | | |
|---|--------------|-----------|--------------|--------------|------------|-------------|-------------|-------------|--------------------|-------------|-------------|-------------|
| | | Rank | MAUC | BCA | Rank | MAE | WES | CPA | Rank | MAE | WES | CPA |
| Billabong-UniAV45  | 1 | 1 | 0.719 | 0.624 | 1-2 | 8.71 | 8.55 | 0.33 | 3-4 | 3.49 | 3.40 | 0.50 |
| Billabong-Uni  | 2 | 2 | 0.717 | 0.621 | 1-2 | 8.71 | 8.55 | 0.33 | 3-4 | 3.49 | 3.40 | 0.50 |
| Billabong-MultiAV45  | 3 | 3 | 0.661 | 0.562 | 3-4 | 12.95 | 12.71 | 0.42 | 1-2 | 3.16 | 3.08 | 0.47 |
| Billabong-Multi  | 4 | 4 | 0.658 | 0.552 | 3-4 | 12.95 | 12.71 | 0.42 | 1-2 | 3.16 | 3.08 | 0.47 |
| Simple-SPMC-Plymouth2  | - | 5 | 0.500 | 0.504 | - | - | - | - | - | - | - | - |
| Simple-SPMC-Plymouth1  | - | 6 | 0.500 | 0.499 | - | - | - | - | - | - | - | - |

Table 6.ST1: Results on custom prediction sets from two teams: Billabong and SPMC-Plymouth. SPMC-Plymouth predicted fewer subjects due to an incomplete submission, while Billabong used a prediction set similar to D3, but filled in missing data for cognitive tests and MRI with the last available measurement. SPMC-Plymouth only submitted predictions for clinical diagnosis, and obtained an MAUC score of 0.5. Results from Billabong show higher MAUC and BCA in diagnosis prediction compared to D3, but lower performance for ADAS-Cog13 and Ventricle volume prediction. Bold entries show best scores in this category.

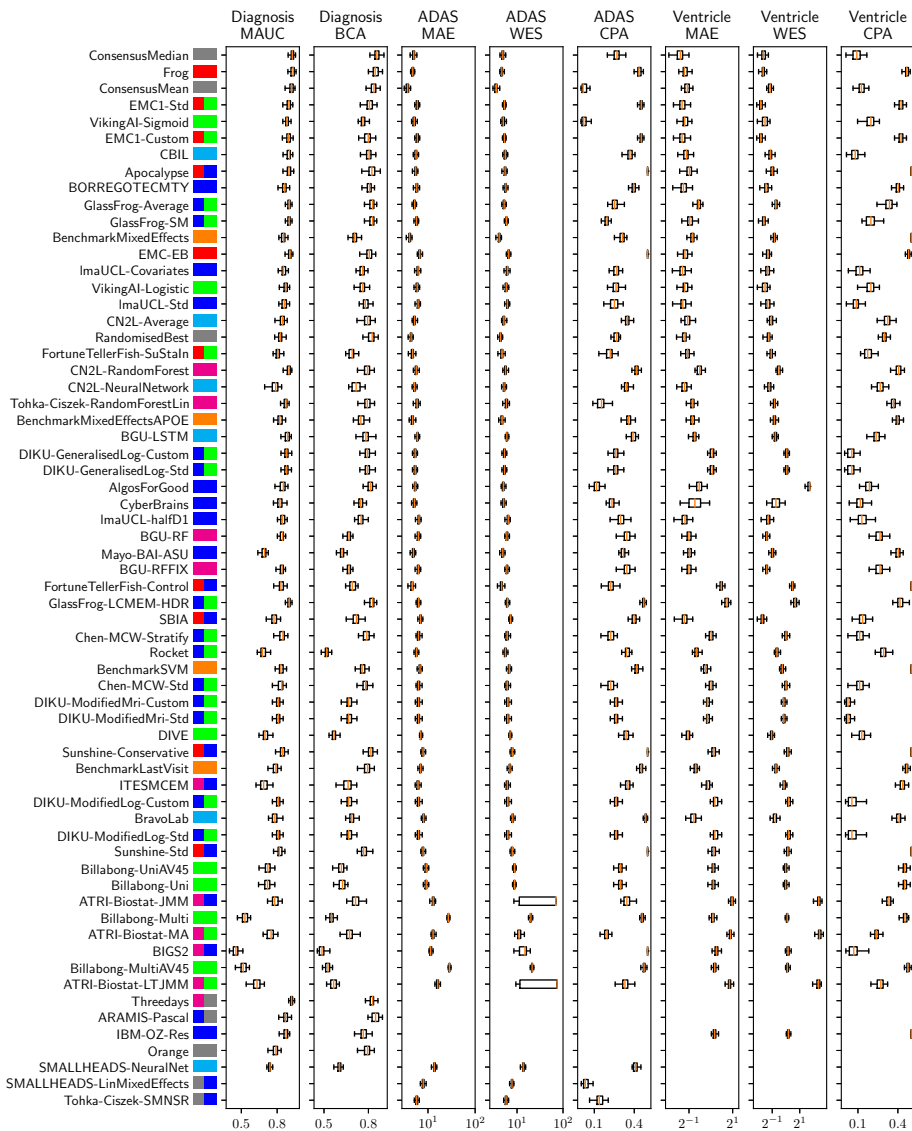


Figure 6.SF1: Distribution of performance metrics for clinical diagnosis (MAUC and BCA), ADAS-Cog13 (MAE, WES and CPA) and ventricle volume (MAE, WES and CPA) on the longitudinal D2 prediction set. For each entry, we plot the distribution of performance metrics derived using 50 bootstrap data sets drawn from the D4 test set. The submissions (rows) are in the same order as in Table 6.2. Entries are missing where teams did not make predictions for a particular target variable.

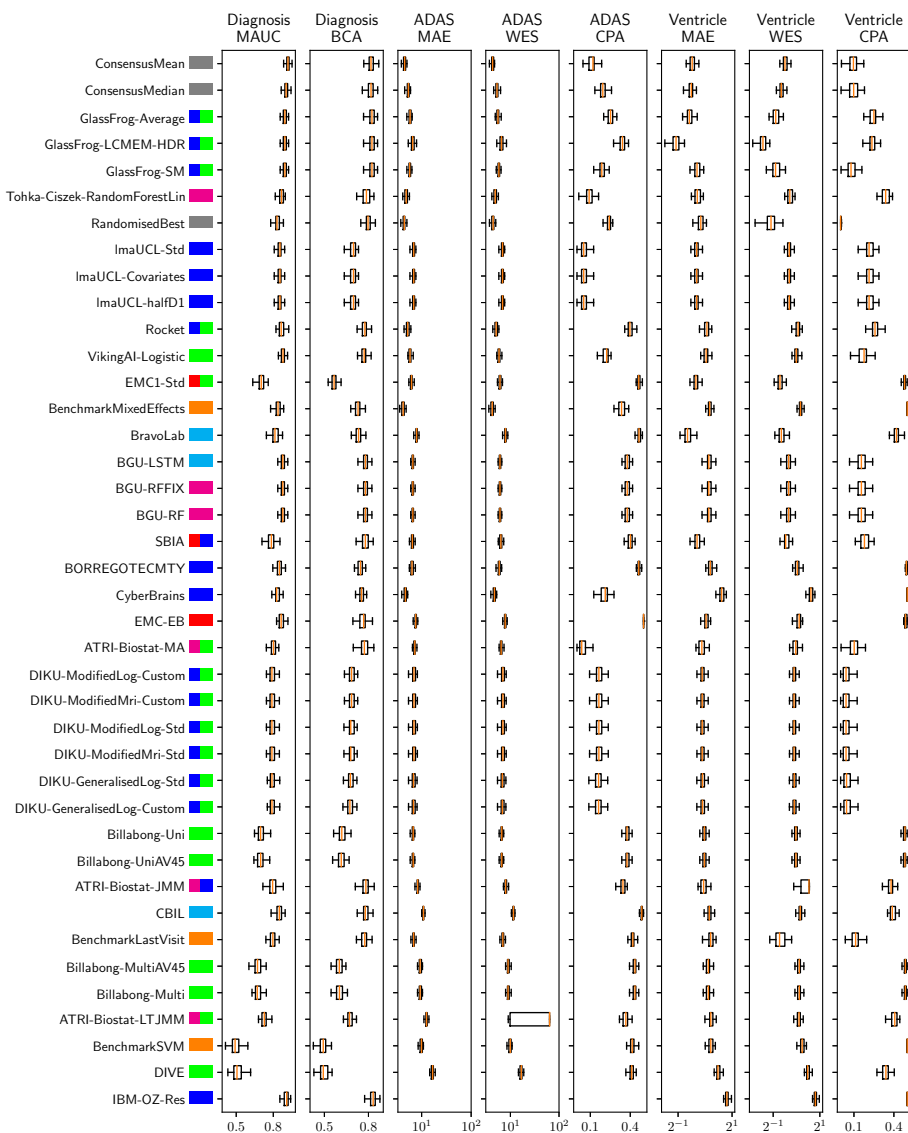


Figure 6.SF2: Box plots of performance metrics for clinical diagnosis (MAUC and BCA), ADAS-Cog13 (MAE, WES and CPA) and ventricle volume (MAE, WES and CPA) on the cross-sectional D3 prediction set. The submissions (rows) are in the same order as in Table 6.3. Some entries are missing because teams did not make predictions for those target variables.

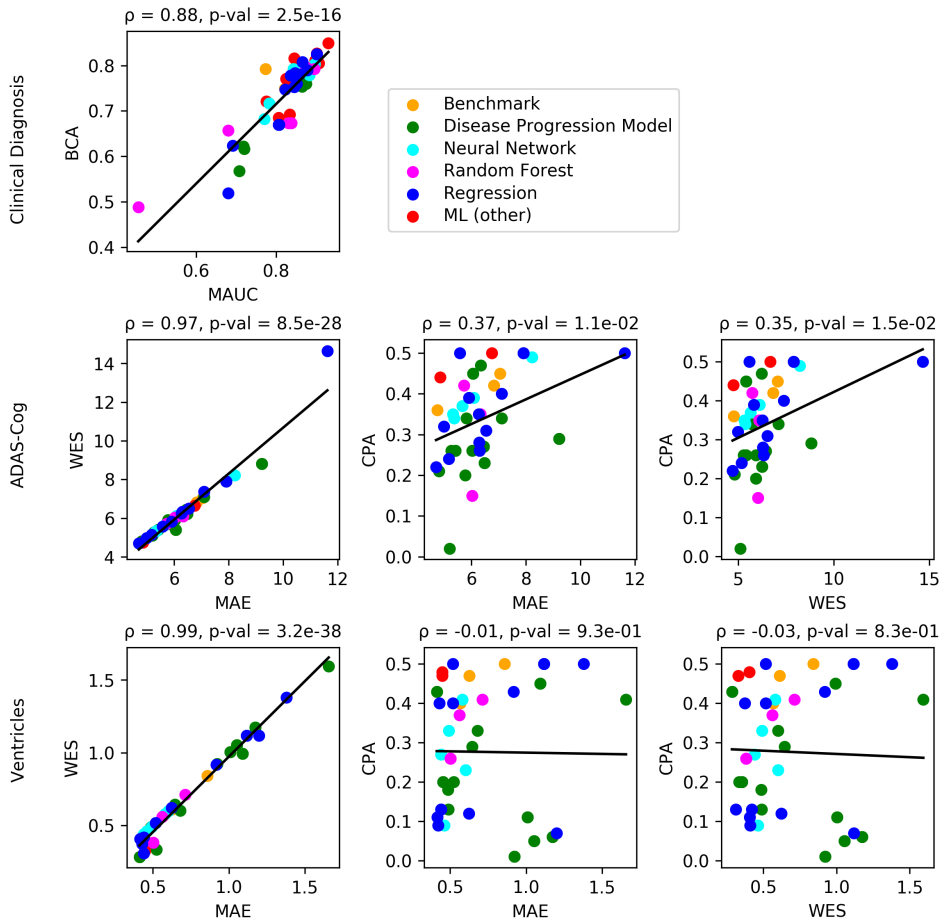


Figure 6.SF3: For D2 submissions, we show scatter plots of pairs of performance metrics for (top row) clinical diagnosis, (middle row) ADAS-Cog13 and (bottom row) Ventricles. Each dot is a participant submission, coloured according to the type of prediction algorithm used. Correlation coefficients and p-values are given above each subplot. A few outlier submissions with ADAS MAE > 20, ADAS WES > 40 or Ventricle WES > 3 were excluded from the analysis.

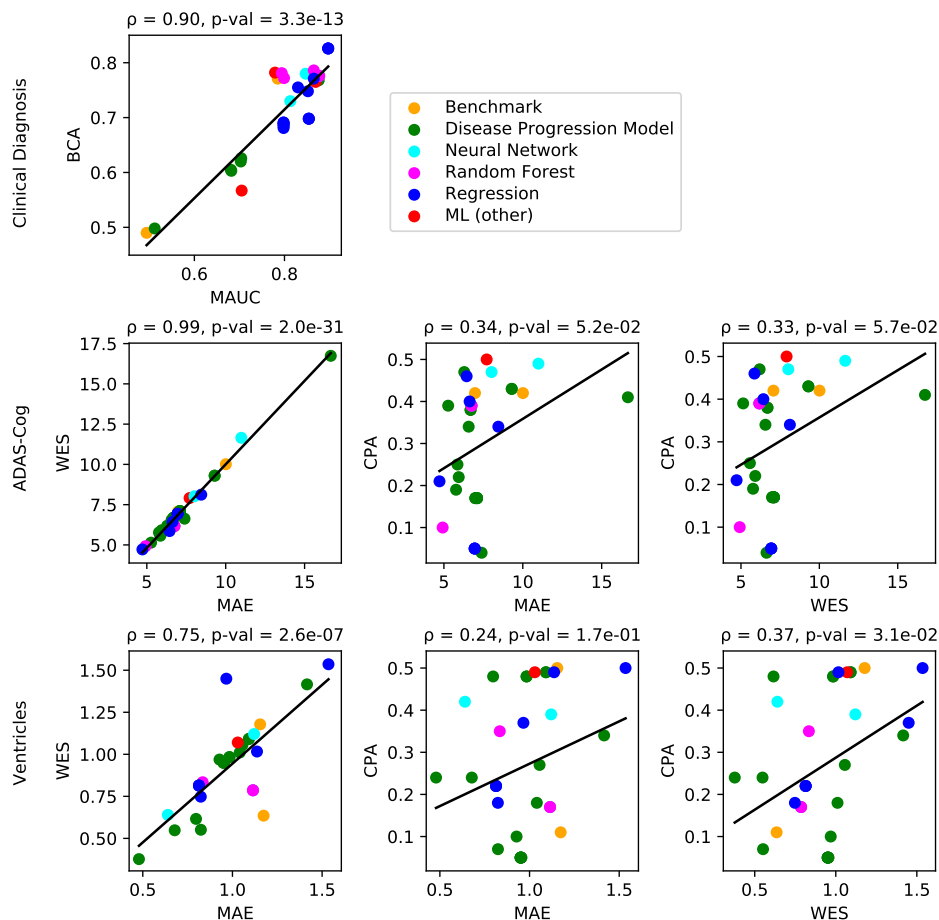


Figure 6.SF4: For D3 submissions, we show scatter plots of pairs of performance metrics for (top row) clinical diagnosis, (middle row) ADAS-Cog13 and (bottom row) Ventricles. Each dot is a participant submission, coloured according to the type of prediction algorithm used. Correlation coefficients and p-values are given above each subplot. A few outlier submissions with ADAS MAE > 20, ADAS WES > 40 or Ventricle WES > 3 were excluded from the analysis.

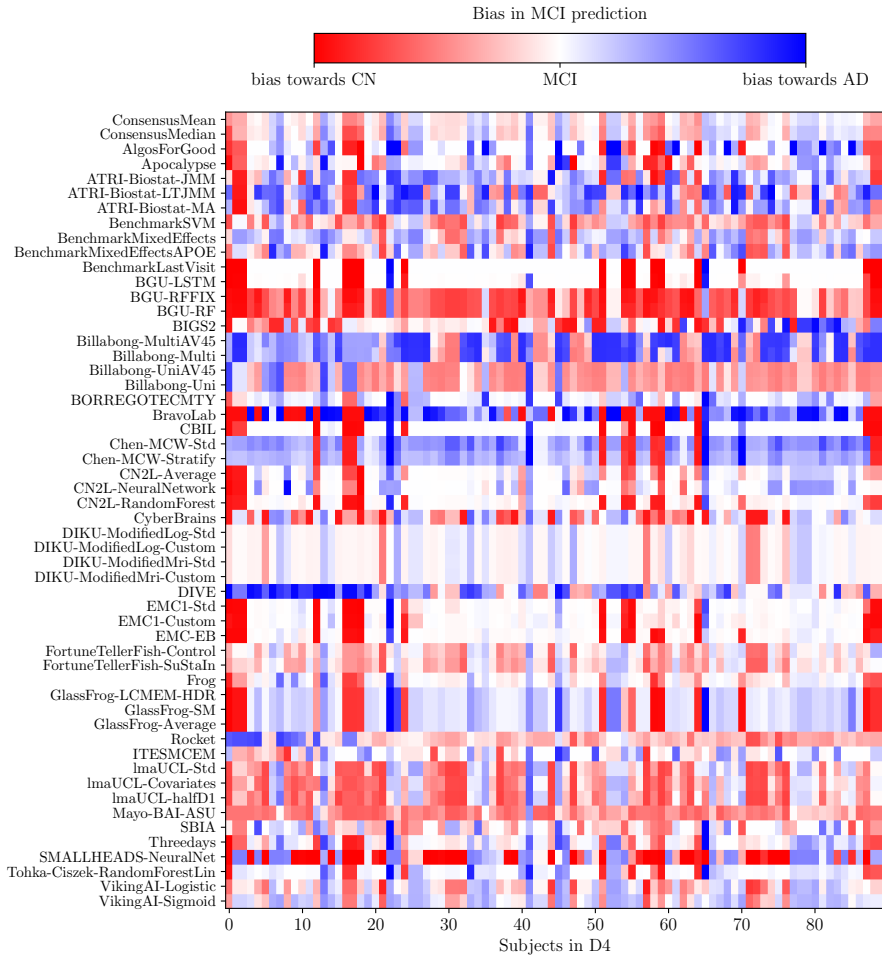


Figure 6.SF5: Bias in prediction of clinical diagnosis for MCI subjects only. X-axis shows individual subjects with designated MCI status at the clinical visit in D4, while the Y-axis shows TADPOLE algorithms. Red represents subjects which were predicted as CN with true diagnosis of MCI, while blue represents subjects predicted as AD with true diagnosis of MCI. Some algorithms show systematic biases either towards CN or AD.

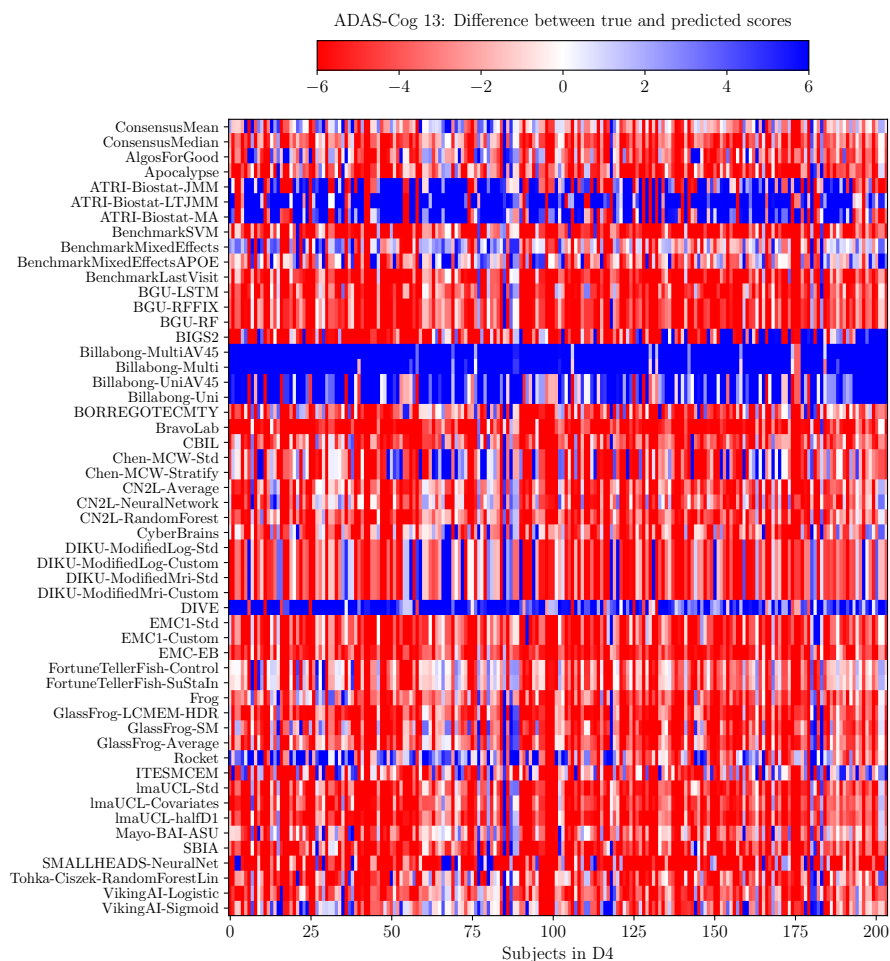


Figure 6.SF6: Bias in prediction of ADAS-Cog13. X-axis shows individual subjects with ADAS-Cog measurements in D4, while Y-axis shows TADPOLE algorithms. Red represents under-estimates while blue represents over-estimates. Most algorithms under-estimate ADAS-Cog measurements.

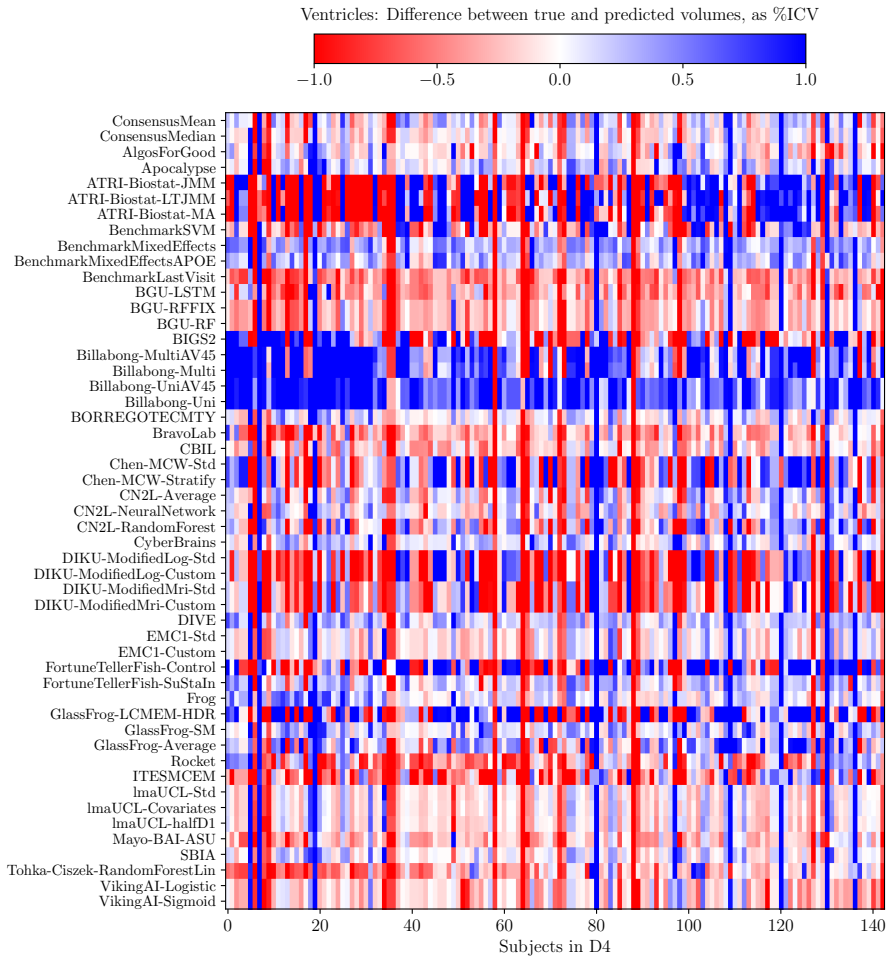


Figure 6.SF7: Bias in prediction of ventricle volume. X-axis shows individual subjects with Ventricle volume measurements in D4, while Y-axis shows TADPOLE algorithms. Red represents under-estimates while blue represents over-estimates. Some algorithms systematically under-estimate or over-estimate ventricle volume.

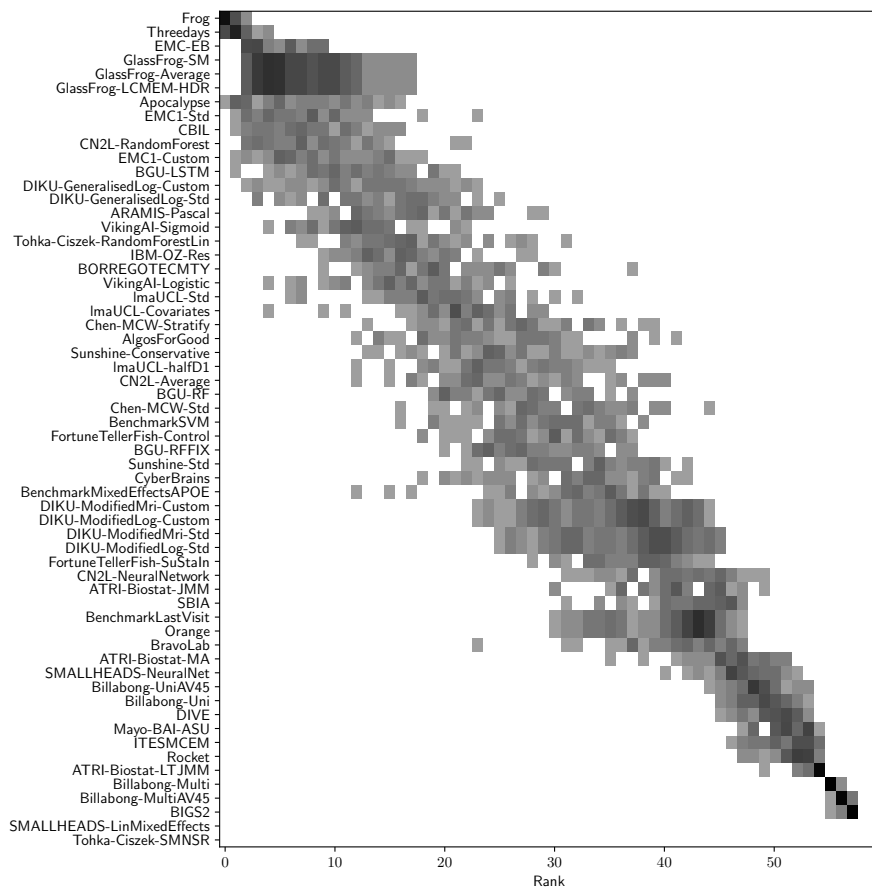


Figure 6.SF8: Distribution of ranks in clinical diagnosis MAUC for TADPOLE submissions using the longitudinal prediction set (D2), obtained from $N = 50$ bootstraps of the test set (D4). More precisely, we computed the MAUC ranks given a specific bootstrap of the test set, and then for each TADPOLE submission (Y-axis) we plotted the number of times it achieved a specific rank. Figures 6.SF9 – 6.SF13 use the same methodology.

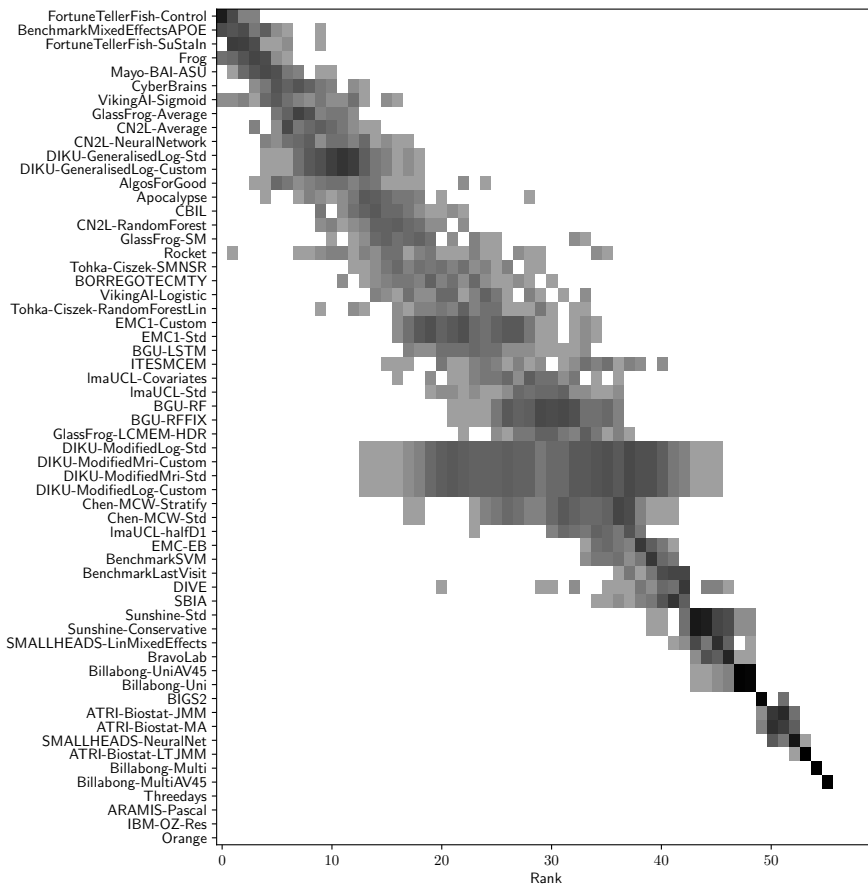


Figure 6.SF9: Distribution of ranks in ADAS-Cog13 MAE for TADPOLE submissions using the longitudinal prediction set (D2)

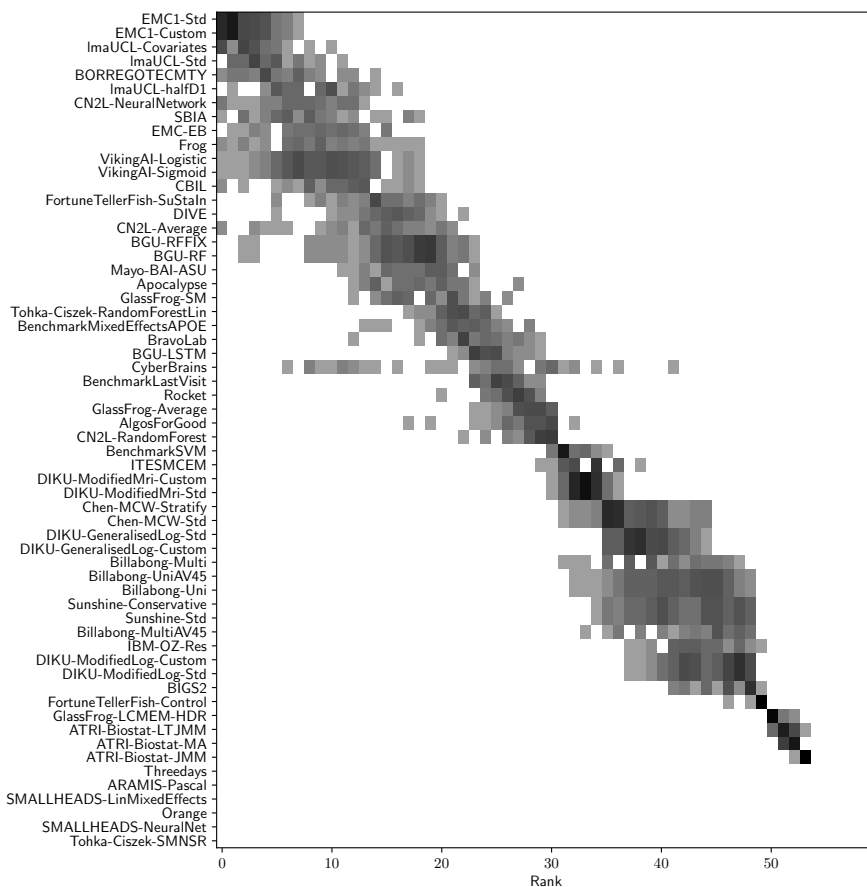


Figure 6.SF10: Distribution of ranks in Ventricle Volume MAE for TADPOLE submissions using the longitudinal prediction set (D2).

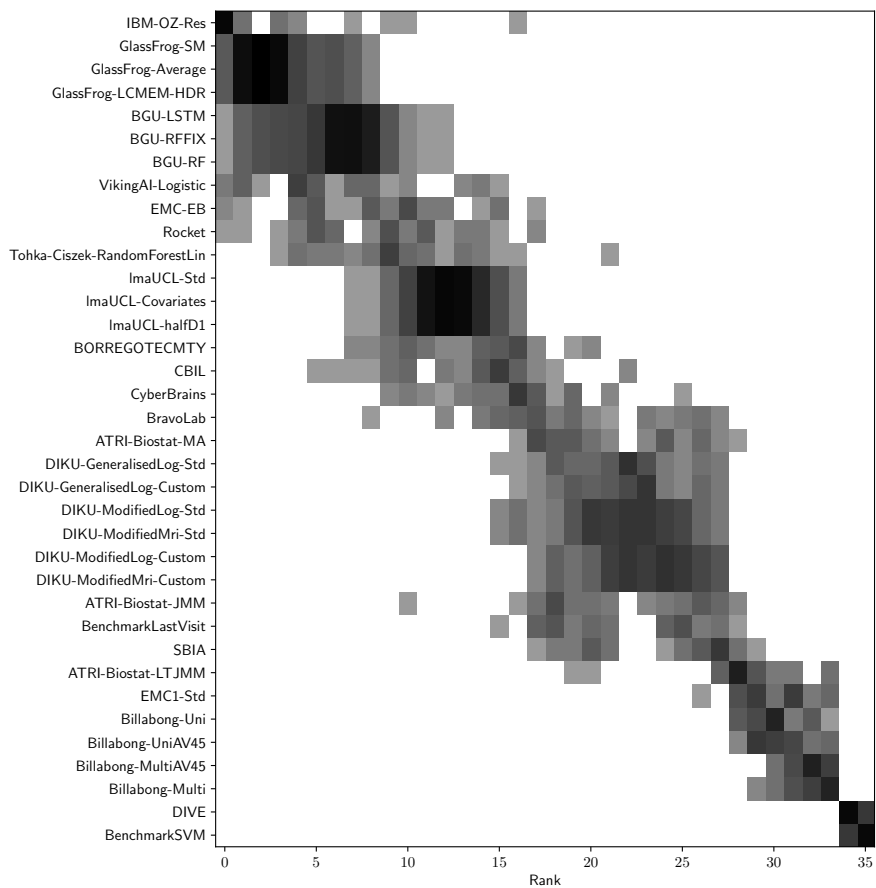


Figure 6.SF11: Distribution of ranks in clinical diagnosis MAUC for TADPOLE submissions using the cross-sectional prediction set (D3).

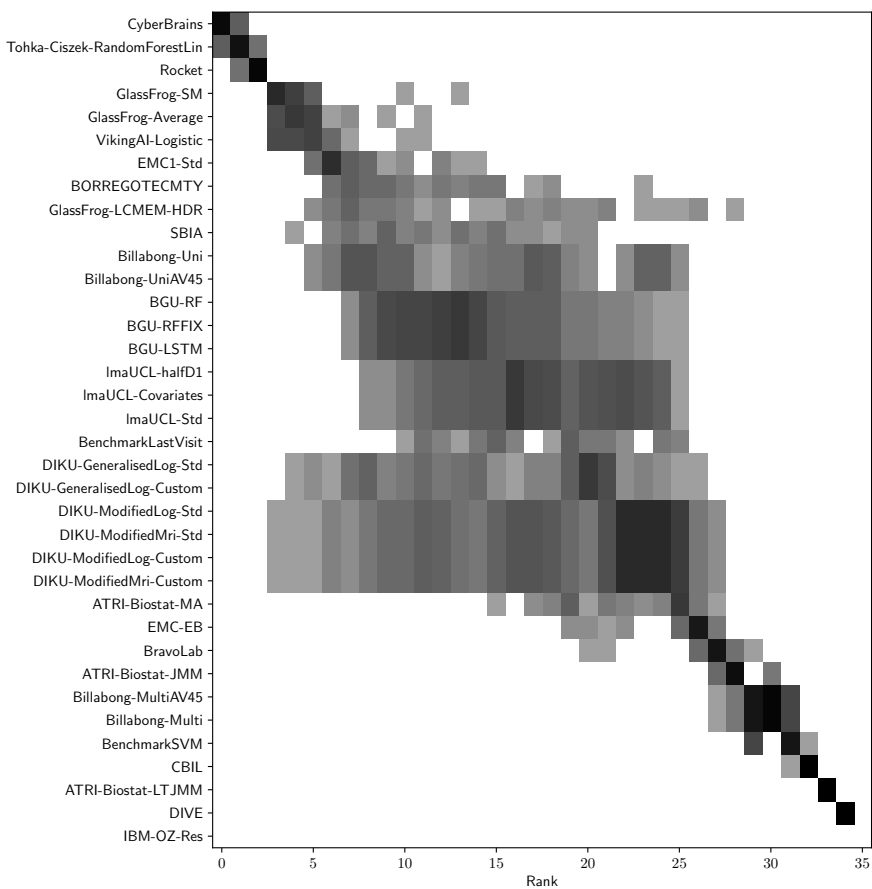


Figure 6.SF12: Distribution of ranks in ADAS-Cog13 MAE for TADPOLE submissions using the cross-sectional prediction set (D3).

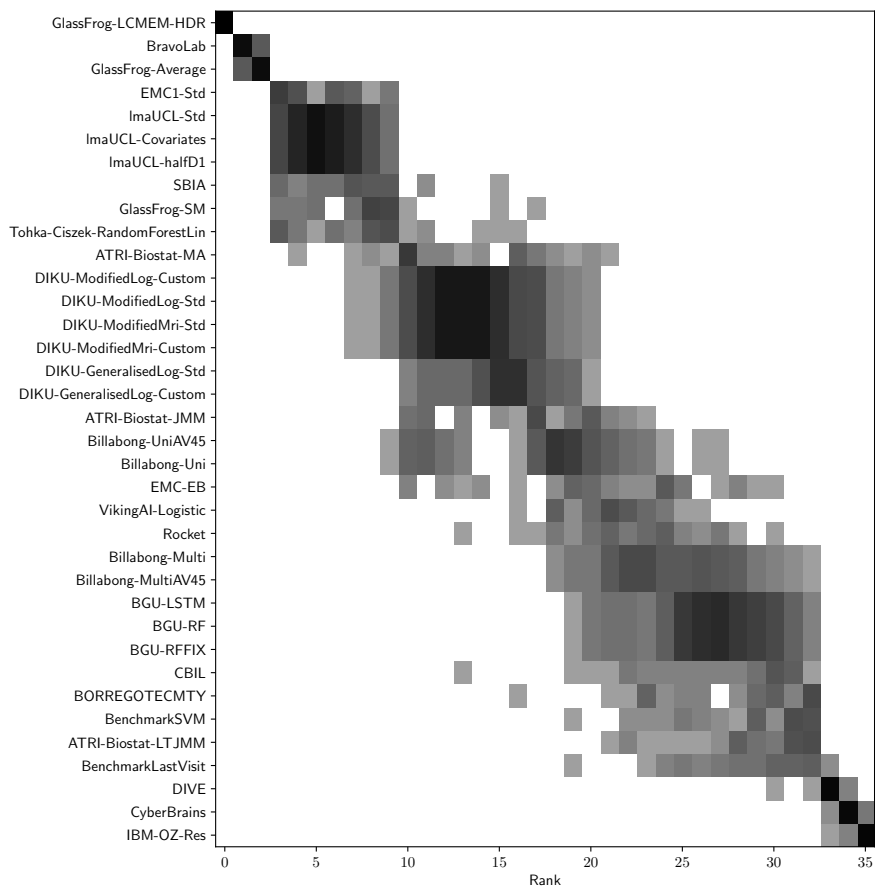


Figure 6.SF13: Distribution of ranks in Ventricle Volume MAE for TADPOLE submissions using the cross-sectional prediction set (D3).

PART II

**Extensions of DEBM: high-dimensional biomarkers,
stratified populations, and subtypes**

Chapter 7

Event-based modeling with high-dimensional imaging biomarkers for estimating spatial progression of dementia

This chapter contains the content of the manuscript 'Event-based modeling with high-dimensional imaging biomarkers for wstimating spatial progression of dementia. **Vikram Venkatraghavan***, Florian Dubost*, Esther E. Bron, Wiro J. Niessen, Marleen de Bruijne, Stefan Klein, In: Lecture Notes in Computer Science, vol 11492: pp. 169-180, Springer, Cham.' It was presented (oral) at the International Conference on Information Processing in Medical Imaging (IPMI) 2019.

* Contributed equally in this work

Vikram Venkatraghavan's contributions in this chapter are: Conceptualization, methodology, software, validation, and writing.

Abstract

Event-based models (EBM) are a class of disease progression models that can be used to estimate temporal ordering of neuropathological changes from cross-sectional data. Current EBMs only handle scalar biomarkers, such as regional volumes, as inputs. However, regional aggregates are a crude summary of the underlying high-resolution images, potentially limiting the accuracy of EBM. Therefore, we propose a novel method that exploits high-dimensional voxel-wise imaging biomarkers: n-dimensional discriminative EBM (nDEBM). nDEBM is based on an insight that mixture modeling, which is a key element of conventional EBMs, can be replaced by a more scalable semi-supervised support vector machine (SVM) approach. This SVM is used to estimate the degree of abnormality of each region which is then used to obtain subject-specific disease progression patterns. These patterns are in turn used for estimating the mean ordering by fitting a generalized Mallows model. In order to validate the biomarker ordering obtained using nDEBM, we also present a framework for Simulation of Imaging Biomarkers' Temporal Evolution (SlmBioTE) that mimics neurodegeneration in brain regions. SlmBioTE trains variational auto-encoders (VAE) in different brain regions independently to simulate images at varying stages of disease progression. We also validate nDEBM clinically using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). In both experiments, nDEBM using high-dimensional features gave better performance than state-of-the-art EBM methods using regional volume biomarkers. This suggests that nDEBM is a promising approach for disease progression modeling.

7.1 Introduction

In 2015, approximately 46.8 million people were estimated to be living with dementia, and by 2050 this number is expected to have increased to 131.5 million [62]. Dementia is characterized by a cascade of neuropathological changes which are quantified using several imaging and non-imaging biomarkers. Understanding how the different biomarkers progress from normal to abnormal state after disease onset enables precise estimation of disease severity in an objective and quantitative way. This can help in identifying individuals at risk of developing dementia as well as monitor the effectiveness of preventive and supportive therapies.

Event-based models (EBM) are a class of disease progression models that estimate the order in which biomarkers become abnormal during disease progression using cross-sectional data [53, 59, 61, 93]. It was reported in a recent paper on discriminative EBM (DEBM) [93] that the EBMs are very sensitive to the quality of biomarkers used for building the model. Hence, to infer the neuropathological changes that occur during dementia accurately, good quality biomarkers are important.

An essential step in an EBM involves mixture modeling to obtain biomarker distributions in normal and abnormal classes [53, 93]. This restricts the current EBMs to only handle scalar biomarkers. In case of imaging biomarkers, regional volumes from structural MRIs are often used [53, 61, 80, 86, 93]. However, regional volumes are a crude summary of the high-dimensional information available from structural MRI, resulting in suboptimal EBM performance, as shall be demonstrated later in this paper. Therefore, we propose a novel

method that exploits voxel-wise imaging biomarkers: n-dimensional discriminative EBM (nDEBM).

Estimating the accuracy of ordering obtained by EBMs is not feasible as ground-truth ordering is not known for a disease. In order to validate the proposed method and compare its accuracy with that of existing state-of-the-art EBM methods, we also present a framework for Simulation of Imaging Biomarkers' Temporal Evolution (SimBioTE). SimBioTE uses variational auto-encoders (VAE) to simulate neurodegeneration in brain regions. These regions are represented by a vector in the latent space of the VAE. Synthetic brain regions were created by sampling latent representations corresponding to target degrees of abnormality which were determined by a ground-truth ordering of disease progression. The generated synthetic brain regions were used as inputs for nDEBM, and the regional aggregates were used as inputs for state-of-the-art EBMs to evaluate the accuracies.

7.2 nDEBM

In Section 8.2.1, a brief introduction to the current DEBM [93] model is given. Section 7.2.2, presents a novel framework to use semi-supervised SVMs in DEBM for estimating posterior probabilities of abnormality for high-dimensional biomarkers. In Section 7.2.3, we use these posterior probabilities to estimate severity of disease progression in an individual.

7.2.1 DEBM

In a cross-sectional dementia dataset (X) of M subjects (consisting of cognitively normal (CN) and patients with dementia (DE)), let X_j denote a measurement of biomarkers for subject $j \in [1, M]$, consisting of N scalar biomarker values $x_{j,i}$. As dementia is characterized by a cascade of neuropathological changes that occurs over several years, even CN subjects can show some abnormal biomarker values. On the other hand, in DE subjects, a proportion of biomarkers may still have normal values, especially in patients at an early disease stage. This leads to label noise in the data and hence clinical labels cannot directly be propagated to individual biomarkers. The DEBM model introduced in [93], similar to previously proposed EBMs [53, 59, 61], fits a Gaussian mixture model (GMM) to construct the normal and abnormal distributions. These are used to compute pre-event and post-event likelihoods $p(x_{j,i} | \neg E_i)$ and $p(x_{j,i} | E_i)$ respectively, where an event E_i is defined as the corresponding biomarker becoming abnormal. The mixing parameters are used as prior probabilities to convert these likelihoods to posterior probabilities $p(\neg E_i | x_{j,i})$ and $p(E_i | x_{j,i})$.

$p(E_i|x_{j,i})\forall i$ are used to estimate the subject-specific orderings s_j . s_j is established such that:

$$s_j \ni p(E_{s_j(1)}|x_{j,s_j(1)}) > p(E_{s_j(2)}|x_{j,s_j(2)}) > \dots > p(E_{s_j(N)}|x_{j,s_j(N)}) \quad (7.1)$$

Finally, DEBM computes the central event ordering S from the subject-specific estimates s_j . To describe the distribution of s_j , a generalized Mallows model is used. The central ordering is defined as the ordering that minimizes the sum of distances to all subject-specific orderings s_j , with probabilistic Kendall's Tau being the distance measure.

7.2.2 n-Dimensional Biomarker Progression

It was reported in [93] that the accuracy of EBMs depends on the quality of biomarkers used to build the model. Greater separability of individual biomarkers results in estimation of more accurate event ordering. We hypothesize that high-dimensional imaging biomarkers can increase the separability between the normal and abnormal groups, thus improving the accuracy when used as inputs to EBMs. The use of GMM in EBMs however restricts it to using only scalar or low-dimensional biomarkers as GMMs do not scale well to high-dimensional features. SVMs do scale well to high-dimensional features, but a supervised soft-margin SVM cannot be used because of the large amounts of label noise (upto one third of the elderly CN population could be in pre-symptomatic stages of DE [71]). In this section, we present a way in which scalable semi-supervised SVM classifiers can be used within the DEBM framework with high-dimensional inputs.

Let $X_{j,i}$ denote the high-dimensional imaging biomarker for brain region i . Since the clinical diagnosis of the subject cannot be propagated to each region, the labels cannot be trusted while training a classifier. If we were to train a classifier trusting these labels, independently on each biomarker ($X_{\forall j,i}$), we hypothesize that labels of the data close to the decision boundary or on either side of it cannot be completely trusted for that biomarker. For identifying the labels that cannot be trusted for a biomarker, we propose to train a linear classifier assuming equal class-priors. Fitting a non-linear classifier risks over-fitting to the wrongly-labeled data whereas class-priors derived from labeled data could be misleading as some of the labels might be wrong, for that biomarker.

For biomarker $X_{\forall j,i}$, subjects whose labels are preserved are considered as labeled data ($X_{\mathbb{L},i}$). Subjects whose labels have been rejected, along with any prodromal subjects in the dataset are considered as unlabeled data ($X_{\mathbb{U},i}$). Semi-supervised classifiers can be used in this context for obtaining the decision boundary for each biomarker.

To identify the subjects for whom labels can be trusted when considering $X_{\forall j,i}$, we first train

a linear SVM ($f_{0;i}$) based on CN and DE subjects. After rejecting labels that cannot be trusted (with distance $d_{0;i} < |d_t|$ from the decision boundary), we use semi-supervised learning with EM [273] using linear SVM with subject-specific costs [274] ($f_{1;i}, \dots, f_{k+1;i}$) to iteratively refine the decision boundary. The algorithm for this semi-supervised classification is given below:

Algorithm 2 Semi-Supervised SVM Learning with Subject-specific weights

```

1: for  $i \in \{1 \dots N\}$  do
2:   Train  $f_{0;i}$  with  $X_{\forall j \in \{CN, DE\}, i}$  as inputs
3:    $d_{0;\forall j, i} \leftarrow$  prediction of  $X_{\forall j, i}$  using  $f_{0;i}$ 
4:   for  $j \in \{1 \dots M\}$  do
5:     if  $d_{0;j, i} > |d_t|$  then:  $X_{L, i} \leftarrow X_{j, i}$ 
6:     else:  $X_{U, i} \leftarrow X_{j, i}$ 
7:   Estimate  $\hat{p}_0(E_i|X_{U, i})$  from  $d_{0;U, i}$  (using Platt scaling [275]).
8:   Train  $f_{1;i}$  using  $X_{\forall j, i}$  using  $|\hat{p}_0(E_i|X_{U, i}) - \hat{p}_0(\neg E_i|X_{U, i})|$  as weights of  $X_{U, i}$ .
9:   Estimate  $\hat{p}_1(E_i|X_{U, i})$  from  $d_{1;U, i}$ 
10:   $k \leftarrow 1$ 
11:  while  $\|\hat{p}_k(E_i|X_{U, i}) - \hat{p}_{k-1}(E_i|X_{U, i})\|^2 < \epsilon$  do
12:    Train  $f_{k+1;i}$  using  $X_{\forall j, i} \ni |\hat{p}_k(E_i|X_{U, i}) - \hat{p}_k(\neg E_i|X_{U, i})|$  are weights of  $X_{U, i}$ .
13:    Estimate  $\hat{p}_{k+1}(E_i|X_{U, i})$  from  $d_{k+1;U, i}$ .
14:     $k \leftarrow k + 1$ 
15:  Estimate  $\hat{p}_{k+1}(E_i|X_{\forall j, i})$  from  $d_{k+1;\forall j, i}$ 
16:   $p(E_i|X_{j, i}) \leftarrow \hat{p}_{k+1}(E_i|X_{j, i})$ 

```

d_t was chosen such that such that 5% of correctly classified data closest to decision boundary are treated as unlabeled. Weighing $X_{U, i}$ based on $|\hat{p}_k(E_i|X_{U, i}) - \hat{p}_k(\neg E_i|X_{U, i})|$ is motivated based on [276]. It is done because unlabeled data close to the decision boundary are not the ideal support vectors. The samples which are farther away from the decision boundary of the previous iteration can be trusted more as support vectors for the next iteration of training.

7.2.3 Patient Staging

Patient staging refers to the process of positioning individuals on a disease progression timeline characterized by the obtained event ordering. Patient stage (Υ_j) is computed as an expectation of event-centers (λ_n) with respect to $p(n, S, X_j)$, where n denotes the possible discrete stages in the timeline characterized by N biomarker events. Event-centers are the positions of the biomarker events on a normalized disease progression timeline $[0, 1]$, that

capture relative distances between events.

$$\Upsilon_j = \frac{\sum_{n=1}^N \lambda_n p(n, S, X_j)}{\sum_{n=1}^N p(n, S, X_j)} \quad (7.2)$$

$p(k, S, X_j)$ can be expressed in-terms of posterior probabilities of events obtained from semi-supervised SVM as:

$$p(n, S, X_j) \propto \prod_{i=1}^n p(E_{S(i)} | X_{j,S(i)}) \times \prod_{i=n+1}^N p(\neg E_{S(i)} | X_{j,S(i)}) \quad (7.3)$$

7.3 SImBioTE: A Validation Framework

For validating classical EBM and nDEBM in a unified framework, we extend the framework developed in [79] for simulating datasets consisting of scalar biomarkers, to be capable of generating datasets with realistic voxel-wise imaging biomarkers. It was built on the assumption that the trajectory of biomarker progression follows a sigmoid. Using a similar assumption, we consider the degree of abnormality in different regions $(a_{j,i})$ follows a sigmoidal trajectory.

$$a_{j,i}(\Psi) = \frac{1}{1 + \exp(-\rho_i(\Psi - \xi_{j,i}))} + \epsilon \quad (7.4)$$

Ψ denotes disease stage of a subject which we take to be a random variable distributed uniformly throughout the disease timeline. ϵ is the equivalent of measurement noise, which represents randomness in the measurement of abnormality. ρ_i signifies the rate of progression of a biomarker, which we take to be equal for all subjects for all biomarkers. It was shown in [93] that the performance of EBMs is similar for equal $\rho_i \forall i$ and unequal ρ_i . $\xi_{j,i}$ denotes the disease stage at which the biomarker becomes abnormal.

After randomly choosing degrees of abnormalities for different regions, we use a variational autoencoder (VAE) [277] for each region i , to generate 3D images of these brain regions at a target degree of abnormality $a_{j,i}(\Psi)$. VAEs are neural networks consisting of two main components: an encoder E which projects input images into a lower dimensional space \mathbb{R}^K called the latent space, and a decoder D which generates images from their hidden representation in the latent space $Z \in \mathbb{R}^K$. Once the VAE has been trained using a large dementia dataset, a latent representation $Z_{j,i;t}$ corresponding to the target degree of abnormality $a_{j,i}(\Psi)$ can be sampled in the latent space. The decoder D then generates a 3D

image $D(Z_{j,i;t})$ corresponding to $a_{j,i}(\Psi)$. Below we describe the VAE used in this work, and the sampling strategy in the latent space.

7.3.1 Implementation of the Convolutional Variational Autoencoder

Figure 7.1 summarizes the architecture of our VAE. We use a ReLU activation after each convolutional layer, except after the last $1*1*1$ convolutional layer. We implemented the loss function as proposed by Kingma and Welling [277], with mean-square-error (MSE) and Kullback-Leibler divergence. We optimized the network with Adadelata [278].

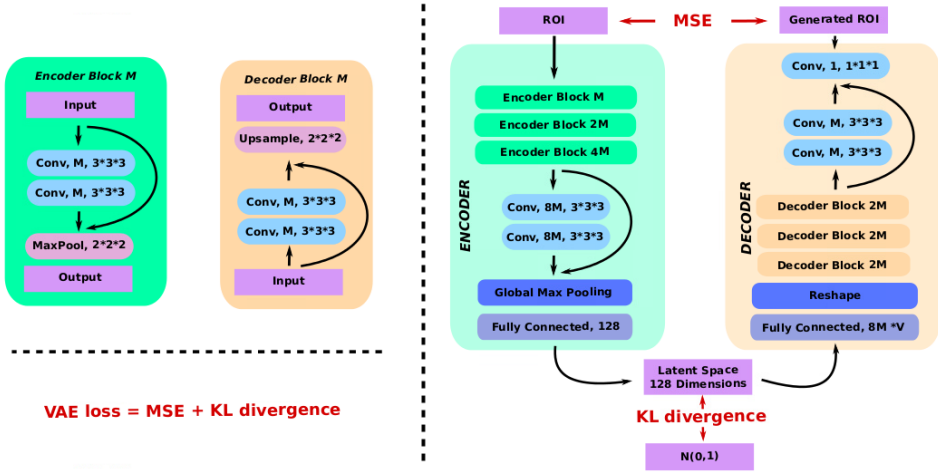


Figure 7.1: Architecture of the Variational Autoencoder.

7.3.2 Sampling Strategy in the Latent Space

To navigate in the latent space \mathbb{R}_i^K of region i , we use Euclidean geometry. We first build a scale vector U_i in the latent space to describe the range of the disease from CN to DE. In order to generate a point $Z_{j,i;t} \in \mathbb{R}_i^K$ at the target degree of abnormality $a_{j,i}(\Psi)$, we first randomly sample a point $Z_{j,i;s} \in \mathbb{R}_i^K$, and translate it along the direction of the scale vector U_i until we reach the target abnormality $a_{j,i}(\Psi)$.

Scale Vector from Cognitively Normal to Dementia.

To build the scale vector U_i , we first compute the latent representations of all the images of region i in the training dataset by projecting these images in the latent space \mathbb{R}_i^K using the encoder E . Then we use the binary labels – CN and DE – of each subject j to compute the means $\mu_{i;CN} \in \mathbb{R}_i^K$ and $\mu_{i;DE} \in \mathbb{R}_i^K$, and standard deviations $\sigma_{i;CN} \in \mathbb{R}_i^K$ and $\sigma_{i;DE} \in \mathbb{R}_i^K$ for each of the two categories respectively.

This is followed by computing the vector joining the two mean points as $u_i = \mu_{i;DE} - \mu_{i;CN}$. The idea is to create a vector U_i spanning the range of the disease progression, from CN to DE. However, u_i joins only the means, if we want to capture the whole distribution, we need to lengthen this vector by a multiple of the standard deviations, on both sides: for instance by $3\sigma_{i;CN}$ in the CN side, and $3\sigma_{i;DE}$ on the DE side. To do so, we compute the scalar projections of the standard deviations as $\sigma_{i;CNp} = |\sigma_{i;CN} \cdot \hat{u}_i|$ and $\sigma_{i;DEp} = |\sigma_{i;DE} \cdot \hat{u}_i|$, where $\hat{u}_i = u_i / \|u_i\|_2$. Now we can compute the new origin point (CN) as $O = \mu_{i;CN} - 3\sigma_{i;CNp} \hat{u}_i$, and the new end point (DE) as $M = \mu_{i;DE} + 3\sigma_{i;DEp} \hat{u}_i$. Finally, we can compute $U_i = M - O$. Note that $\hat{U}_i = U_i / \|U_i\|_2 = \hat{u}_i$.

Navigation for generation

We first randomly sample a point $Z_{j,i;s}$ using the mean and standard deviation of the latent representations of all subjects j for region i . The degree of abnormality $a_{j,i;s}$ of this randomly sampled point $Z_{j,i;s}$ can be computed as $a_{j,i;s} = OZ_{j,i;s} \cdot \hat{U}_i / \|U_i\|_2$. To reach the target point $Z_{j,i;t}$, we need to translate the randomly sampled point $Z_{j,i;s}$. This now can be done by computing $Z_{j,i;t} = Z_{j,i;s} + (a_{j,i;t} - a_{j,i;s})U_i$. To generate the corresponding brain region we can now use the decoder and compute $D(Z_{j,i;t})$.

7.4 Experiments and Results

This section describes the experiments performed to validate the proposed nDEBM algorithm and also compare it with classical EBM [53] and DEBM [93] algorithms.

7.4.1 ADNI Data

We considered 1737 ADNI subjects (417 CN, 106 with significant memory concern (SMC), 872 with mild cognitive impairment (MCI) and 342 AD subjects) who had a 1.5T structural MRI (T1w) scan at baseline. This was followed by multi-atlas brain extraction using the method described in [9]. Gray matter (GM) volumes of segmented regions were regressed on age, sex and intra-cranial volume (ICV) and the effects of these factors were subsequently corrected for. Student's t-test between CN and AD was performed on these confounding factor corrected GM volumes and 15 regions with smallest p-values were retained. They were subsequently used as inputs for DEBM and EBM [53] models. The optimization routine proposed in [93] was used to train the GMM in these two models.

The T1w images were registered to a common template space based on the method used in [9]. Probabilistic tissue segmentations were obtained for white matter (WM), GM, and cerebrospinal fluid on the T1w image using the unified tissue segmentation method [77]. The voxel-wise GM density maps were computed based on the Jacobian of the local deformation

map and the probabilistic GM volume. The GM density maps from the corresponding 15 regions were used as inputs for nDEBM.

Model Validation

Since the groundtruth ordering is not known in a clinical setting, validation of these models was done based on the resulting patient stages for classifying AD subjects from CN as well as for classifying MCI non-converters (MCI-nc) from converters (MCI-c)*. We performed 10-fold cross-validation with 10 repetitions. The training set was used to train the three models. The disease timeline created during training was used to stage the patients in the test-set.

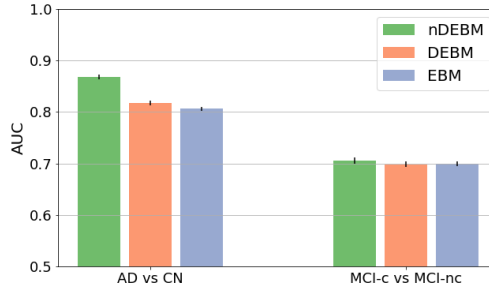


Figure 7.2: AUC measures when patient stages of nDEBM, DEBM and EBM were used for classifying AD vs CN (left) and MCI-c vs MCI-nc (right). The error bar represents the standard deviation in 10 random repetitions.

Figure 7.2 shows the results of 10 random repetitions of 10-fold cross-validation on ADNI dataset. The error-bar shows the standard deviation of the AUCs when the patient stages obtained from nDEBM, DEBM and EBM were used to classify AD vs CN and MCI-c vs MCI-nc.

Uncertainty in Estimation

Variation of the positions of the biomarker events on a normalized disease progression timeline (event-centers) estimated by nDEBM and DEBM was studied by creating 50 bootstrapped samples of the data and applying nDEBM on those samples[†].

Figure 7.3 shows event-centers estimated by nDEBM and DEBM along with the uncertainty in their estimations. The biomarkers are ordered along the y-axis based on the event-ordering obtained by nDEBM. 3D visualization of the estimated central ordering is shown in Figure 7.4.

*MCI converters are subjects who convert to AD within 3 years of baseline measurement

[†]EBM was left out of this experiment as the concept of event-centers was not introduced for EBM.

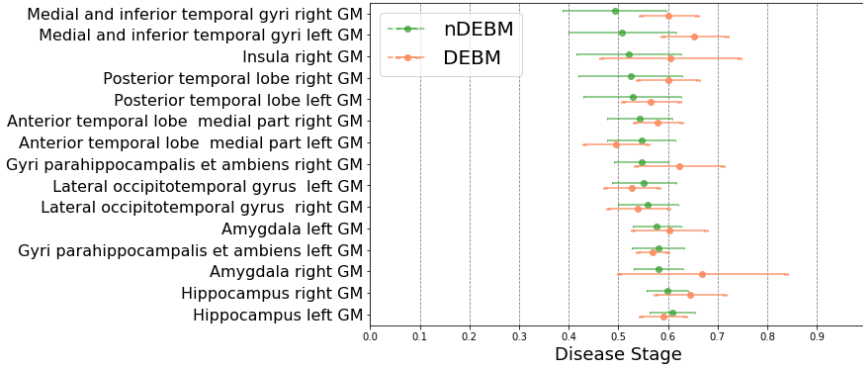


Figure 7.3: Variation of event-centers estimated by nDEBM and DEBM in 50 bootstrapped samples of the ADNI data. The error bar represents the standard deviation of the respective event-centers.

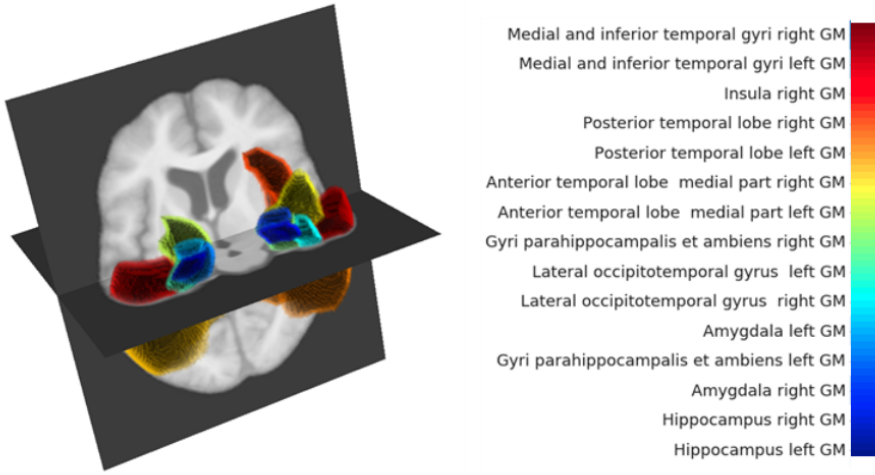


Figure 7.4: 3D visualization of the central ordering estimated by nDEBM. The position of a region in the estimated ordering is encoded using a distinct color. The color-map used for this purpose is on the right-side of the figure, where the early biomarkers are at the top of the color-map and late biomarkers are at the bottom.

7.4.2 Simulation Data

In our experiments, $\xi_{j,i} \forall j$ are random variables with Normal distribution $\mathcal{N}(\mu_{\xi_i}, \Sigma_{\xi_i})$. μ_{ξ_i} were equally spaced for different i . The value of Σ_{ξ_i} was set to be $\Delta\xi$ where $\Delta\xi$ is the difference in μ_{ξ_i} of adjacent events. ρ_i was considered to be equal for all biomarkers. Ψ of the simulated subjects were distributed uniformly throughout the disease timeline.

We first trained 15 VAEs (one per selected region) on the GM density maps of the ADNI dataset. Then we generated - as detailed in Section 7.3 - images for these 15 regions and for 1737 artificial subjects according to pre-computed degrees of abnormality as defined in Equation 7.4. These degrees of abnormality are different for each region and each subject. We repeated this process 10 times, with different random simulations. The voxel-wise GM density maps of regions were used for obtaining the ordering using nDEBM. The GM volume of the simulated regions (computed by integrating the GM density map over the region of interest) were used as biomarkers for DEBM and EBM.

SimBioTE results depicting Lateral occipitotemporal gyrus atrophy in simulated images is shown in Figure 7.5. The images thus generated were used for validating different EBM methods.

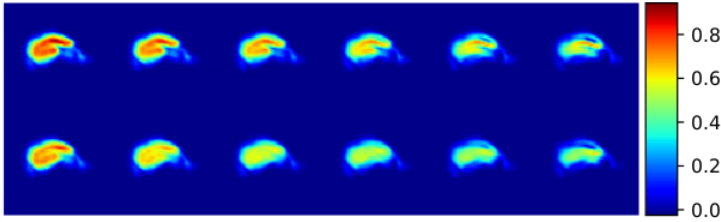


Figure 7.5: An example of Lateral occipitotemporal gyrus (right) atrophy as simulated by SimBioTE. The interpolation spans the full range U_i , as described in section 7.3. Left is normal (CN) and right is abnormal (DE). The two rows shows disease progression in two different simulated subjects.

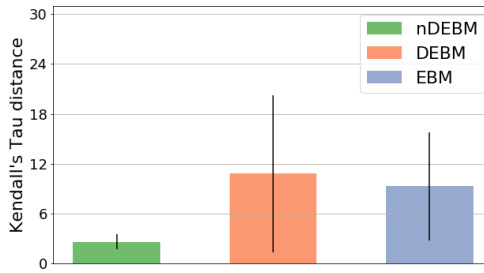


Figure 7.6: Inaccuracies, as measured by Kendall's Tau distance from groundtruth, of nDEBM, DEBM and EBM. The error bar represents the standard deviation of the errors made in 10 repetitions of simulations.

The errors made by different EBM methods on SimBioTE data are shown in Figure 7.6. The estimated ordering and the ground-truth orderings were compared using Kendall's Tau distance.

7.5 Discussions

We proposed a novel method (nDEBM) that exploits high-dimensional voxel-wise imaging biomarkers for event-based modeling using semi-supervised SVM. This was validated based on ADNI dataset, where the spatial spread of structural abnormality was estimated based on a cross-sectional dataset. However this is an indirect validation of the orderings based on accuracy of the estimated patient stages, since the ground-truth ordering for clinical data is unknown.

To unambiguously validate the orderings obtained, we also proposed a new simulation framework (SlmBioTE) to simulate voxel-wise imaging biomarkers based on training VAEs on different regions. It is known that GM tissue is lost in AD progression. Therefore the voxel-wise GM density maps will become darker as the disease progresses, as can be observed in Figure 7.5. It was also observed in Figure 7.5 that simulated regions for different subjects shows considerable variations. This shows that the simulation framework is capable of generating datasets with realistic atrophy and with good inter-subject variability. This, in combination with the scalar biomarkers' simulation framework, results in images where the disease progression in different regions can be controlled. However, a more thorough validation of the simulation framework by comparing the atrophy patterns of the simulated data with that of real-life longitudinal data is needed to understand the effect of different model parameters. Possible extensions of SlmBioTE includes simulating whole brain images from these independent regions, which can be used to validate wider range of disease progression models.

The datasets simulated by SlmBioTE were used for inputs for different EBMs. It was observed in Figure 7.6 that the orderings obtained by nDEBM are much closer to the ground-truth as compared to DEBM and EBM. It was also observed in Figure 7.2 that the patient stages obtained by nDEBM delineates AD and CN subjects much better than the ones obtained by DEBM and EBM. The AUCs of classifying MCI-c vs MCI-nc are also marginally better for nDEBM as compared to the other two methods. These experiments serve as a validation for our initial hypothesis that increasing the dimensionality of the inputs helps in better delineation of normal and abnormal regions, which increases the accuracy of the resulting ordering. It can hence be concluded that the voxel-wise data helps nDEBM in estimating the disease progression more accurately than regional volumes. However, the choice of hyper-parameters in nDEBM (for e.g. d_t , SVM slack parameters) was done ad-hoc. The effect they have on the accuracy of the resulting ordering needs to be studied through more rigorous validation experiments.

The difference in event orderings obtained by nDEBM and DEBM as observed in Figure 7.3 suggests that the two types of inputs can lead to very different results. Hence, computing regional aggregates, such as volumes, and using that as inputs for EBMs as done in [53, 61, 80,

86, 93] is not an optimal choice for estimating the spatial progression of disease.

7.6 Conclusion

We hypothesized that high-dimensional imaging biomarkers would result in better delineation of normal and abnormal regions thus leading to more accurate event-based models. We hence proposed a novel method (nDEBM) that exploits high-dimensional voxel-wise imaging biomarkers based on semi-supervised SVM to estimate temporal ordering of neuropathological changes in the brain structure using cross-sectional data. We also proposed a simulation framework (SImBioTE) using variational auto-encoders that mimics neurodegeneration in brain regions to validate nDEBM. Furthermore, we applied nDEBM framework to a set of 1737 subjects from ADNI dataset for clinically validating the method. In both experiments, nDEBM using high-dimensional features gave better performance than state-of-the-art EBM methods using regional volume biomarkers. This served as a validation for our initial hypothesis. nDEBM thus presents a new paradigm for estimating spatial progression of dementia.

Chapter 8

Analyzing the effect of *APOE* on Alzheimer's disease progression using an event-based model for stratified populations

This chapter contains the content of the manuscript 'Analyzing the effect of *APOE* on Alzheimer's disease progression using an event-based model for stratified populations. **Vikram Venkatraghavan**, Stefan Klein, Lana Fani, Leontine S. Ham, Henri Vrooman, M. Kamran Ikram, Wiro J. Niessen, Esther E. Bron, for the Alzheimer's Disease Neuroimaging Initiative. *NeuroImage*, published online ahead of print, 2020'. An earlier version of this chapter was also presented (poster) at the Computational approaches for ageing and age-related diseases (CompAge) workshop 2020.

Vikram Venkatraghavan's contributions in this chapter are: Conceptualization, methodology, software, validation, and writing.

Abstract

Alzheimer's disease (AD) is the most common form of dementia and is phenotypically heterogeneous. *APOE* is a triallelic gene which correlates with phenotypic heterogeneity in AD. In this work, we determined the effect of *APOE* alleles on the disease progression timeline of AD using a discriminative event-based model (DEBM). Since DEBM is a data-driven model, stratification into smaller disease subgroups would lead to more inaccurate models as compared to fitting the model on the entire dataset. Hence our secondary aim is to propose and evaluate novel approaches in which we split the different steps of DEBM into group-aspecific and group-specific parts, where the entire dataset is used to train the group-aspecific parts and only the data from a specific group is used to train the group-specific parts of the DEBM. We performed simulation experiments to benchmark the accuracy of the proposed approaches and to select the optimal approach. Subsequently, the chosen approach was applied to the baseline data of 417 cognitively normal, 235 mild cognitively impaired who convert to AD within 3 years, and 342 AD patients from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset to gain new insights into the effect of *APOE* carriership on the disease progression timeline of AD. In the $\epsilon 4$ carrier group, the model predicted with high confidence that CSF Amyloid β_{42} and the cognitive score of Alzheimer's Disease Assessment Scale (ADAS) are early biomarkers. Hippocampus was the earliest volumetric biomarker to become abnormal, closely followed by the CSF Phosphorylated Tau $_{181}$ (PTAU) biomarker. In the homozygous $\epsilon 3$ carrier group, the model predicted a similar ordering among CSF biomarkers. However, the volume of the fusiform gyrus was identified as one of the earliest volumetric biomarker. While the findings in the $\epsilon 4$ carrier and the homozygous $\epsilon 3$ carrier groups fit the current understanding of progression of AD, the finding in the $\epsilon 2$ carrier group did not. The model predicted, with relatively low confidence, CSF Neurogranin as one of the earliest biomarkers along with cognitive score of Mini-Mental State Examination (MMSE). Amyloid β_{42} was found to become abnormal after PTAU. The presented models could aid understanding of the disease, and in selecting homogeneous group of presymptomatic subjects at-risk of developing symptoms for clinical trials.

8.1 Introduction

Dementia affects roughly 5% of the world's elderly population of whom 60 – 70% are affected by Alzheimer's Disease (AD), which is the most common form of dementia [279]. There are several neurobiological subtypes of AD [38] and each subtype potentially needs a different strategy to prevent or slow the progression of AD. Understanding the pathophysiological processes in AD is thus crucial for selecting novel preventive or therapeutic targets for clinical trials of disease modifying treatments, identifying target groups for such trials and tracking the disease progression in patients.

While several studies have looked into the pathophysiology of AD [82, 117, 280], it is still not completely understood. Although it has been observed that AD is phenotypically heterogeneous [281, 282, 283] with potentially different pathways for disease progression, these pathways remain unclear. There is hence a need to understand the phenotypic heterogeneity in AD while leveraging neuroimaging, fluid and cognitive biomarkers.

APOE is a triallelic gene in which the $\epsilon 2$ allele reduces the risk of AD [284], the $\epsilon 3$ allele acts as a reference allele and the $\epsilon 4$ allele is a major genetic risk factor of AD [285, 286, 287]. *APOE* has been shown to correlate with phenotypic heterogeneity in AD [288]. Hence we hypothesize that the pathophysiology of AD can be better understood when considering the effect of *APOE* carriership on biomarker changes.

In the context of data-driven methods for understanding AD pathophysiology, disease progression models have been used to study the trajectories of individual biomarkers [51, 67, 68] as well as their progression with respect to each other [53, 59, 61, 70]. Unlike typical machine learning approaches, these models are interpretable by design and provide insight for understanding the mechanisms of disease progression. Event-based models (EBMs) are a class of such interpretable disease progression models that estimate the timeline of neuropathologic change during AD progression using cross-sectional data [53, 93].

Our primary aim is to use the discriminative event-based model (DEBM), which was shown to be more accurate than previously proposed EBMs [93], to understand the effect of different *APOE* alleles on the disease timeline of AD. To shed light on different aspects of neurodegeneration and identify the earliest brain regions affected, we included commonly studied cerebrospinal fluid (CSF) biomarkers, cognitive scores, and volumetric biomarkers from neuroimaging.

The default approach for estimating the disease progression timeline would be to stratify the population based on their *APOE* $\epsilon 2 - 4$ carrier status and independently train the DEBM model on the stratified populations [61]. However, since DEBM is a data-driven model, stratification into smaller groups would lead to less accurate models than those obtained by the original method on the entire dataset. Hence our secondary aim is to propose and evaluate a novel approach in which we split the different steps of DEBM into group-aspecific and group-specific parts, where the entire dataset is used to train the group-aspecific parts and only the data from a specific group is used to train the group-specific parts of the DEBM. We present two different variations of this approach and we hypothesize that the optimal split of the DEBM steps into the group-aspecific and group-specific parts would result in better accuracy of the estimated disease progression timeline. Since the ground-truth timelines are unknown in a clinical setting, we evaluate the accuracy of the proposed variations using simulation experiments and we select the optimal method for the analysis on the effect of *APOE* on the AD progression timeline on patient data.

To summarize, our contributions in this paper include proposing and evaluating a novel approach for using DEBM in stratified populations and estimating a comprehensive timeline of AD progression, in terms of biomarker changes, in the presence of different *APOE* alleles.

8.2 Methods

An introduction to the DEBM model [93] is provided in Section 8.2.1. In Section 8.2.2 we propose our novel approach for using DEBM in stratified populations with its two variations.

8.2.1 Discriminative event-based modeling

In a cross-sectional dataset (X) of M subjects, including cognitively normal individuals (CN), subjects with mild cognitive impairment (MCI) and patients with AD, let X_j denote a measurement of biomarkers for subject $j \in [1, M]$, consisting of scalar biomarker values $x_{j,i}$ for $i \in [1, N]$. $x_{\cdot,i}$ denotes the i -th biomarker for any unspecified j . DEBM estimates the posterior probabilities of individual biomarkers being abnormal. These posterior probabilities are used to estimate the ordering of biomarker changes for each subject independently. The central ordering and disease progression timeline for the entire dataset are estimated based on these subject-specific orderings. The resulting disease progression timeline is used

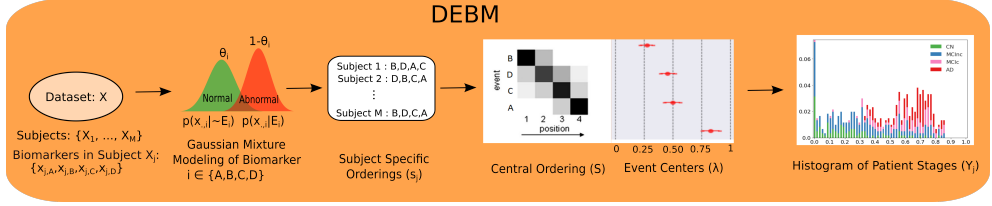


Figure 8.1: Overview of the steps involved in DEBM. Input for the DEBM model is a cross-sectional dataset X with M subjects and various biomarkers (A, B, C and D) representing different aspects of neurodegeneration. Using Gaussian mixture modeling (GMM), mixing parameters (θ_i) and probability density functions of normal ($p(x_{\cdot,i}|\neg E_i)$) and abnormal ($p(x_{\cdot,i}|E_i)$) levels are estimated for each biomarker. This is followed by the estimation of subject-specific orderings (s_j), for each subject in the dataset. Disease progression timeline consisting of central ordering (S) and event-centers (λ) are estimated based on these subject-specific orderings. Based on the constructed disease progression timeline, patient stages (Y_j) of subjects in an independent test-set can be estimated.

for assessing the severity of disease in an individual based on his/her biomarker values. Figure 8.1 shows the different steps involved in DEBM.

Step 1 - Mixture Modeling: As AD is characterized by a cascade of neuropathological changes that occurs over several years, presymptomatic CN subjects can have some abnormal biomarker values. On the other hand, in some clinically diagnosed AD subjects, a proportion of biomarkers may still have normal values, as they might not have an underlying AD pathology or could have atypical AD. Hence clinical labels cannot directly be propagated to individual biomarkers to label normal and abnormal biomarker values. We shall refer to this as biomarker label noise in the rest of the paper. In order to estimate the posterior probabilities of individual biomarkers being abnormal, DEBM, similar to previously proposed EBM [53, 59, 61], fits a Gaussian mixture model (GMM) to construct the normal / pre-event probability density function (PDF), $p(x_{\cdot,i}|\neg E_i)$, and abnormal / post-event PDF, $p(x_{\cdot,i}|E_i)$. Event E_i in this notation is used to denote the corresponding biomarker becoming abnormal and $\neg E_i$ denotes the corresponding biomarker being normal. The aforementioned PDFs can be expressed as:

$$p(x_{\cdot,i}|\neg E_i) = \mathcal{N}(\mu_{i,\neg E}; \sigma_{i,\neg E}) \quad (8.1)$$

$$p(x_{\cdot,i}|E_i) = \mathcal{N}(\mu_{i,E}; \sigma_{i,E}) \quad (8.2)$$

Where, $\mathcal{N}(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ .

For estimating these parameters robustly in the presence of biomarker label noise, the normal and abnormal PDF estimates are first initialized using the mean and standard deviations after truncating the overlapping tails of the observed distributions in CN and AD subjects. This can be observed in Figure 8.2, where the initialization is performed only based on the non-overlapping parts of green and red curves, while the overlapping part is left out to account for biomarker label noise. At this stage of GMM initialization, MCI subjects are left out as well, because it is unsure a priori whether their biomarkers are normal or abnormal. The resulting initialized PDFs are denoted as $\hat{p}(x_{\cdot,i}|\neg E_i)$ and $\hat{p}(x_{\cdot,i}|E_i)$.

This is followed by an alternating GMM maximum likelihood optimization scheme until both the Gaussian parameters as well as the mixing parameters converge. All the subjects, including MCI, are used for GMM optimization. After convergence, these Gaussians are used to represent the PDFs $p(x_{\cdot,i}|\neg E_i)$ and $p(x_{\cdot,i}|E_i)$. The mixing parameters (θ_i) are used as prior probabilities to convert these PDFs to posterior probabilities $p(\neg E_i|x_{\cdot,i})$ and $p(E_i|x_{\cdot,i})$. Figure 8.2 shows an overview of this optimization scheme.

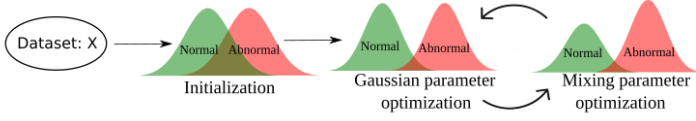


Figure 8.2: Overview of GMM optimization in DEBM.

Step 2 - Subject-specific Orderings: $p(E_i|x_{j,i})\forall i$ are used to estimate the subject-specific orderings s_j . s_j is established such that:

$$s_j \ni p(E_{s_j(1)}|x_{j,s_j(1)}) > \dots > p(E_{s_j(N)}|x_{j,s_j(N)}) \quad (8.3)$$

Step 3 - Central Ordering: DEBM computes the central event ordering S from the subject-specific estimates s_j . To describe the distribution of s_j , a generalized Mallows model is used [72]. The central ordering is defined as the ordering that minimizes the sum of distances to all subject-specific orderings s_j , with probabilistic Kendall's Tau being the distance measure [93]. While S denotes the sequence of biomarker events, the relative position of these events (event-centers) in a normalized scale of $[0, 1]$ is denoted by the vector λ . The pair $\{S, \lambda\}$ together forms a disease progression timeline.

Step 4 - Patient Staging: Once the disease progression timeline is created, subjects in an independent test set (T) can be placed on this timeline to estimate disease severity. This is achieved by converting the biomarker values of the test subjects to posterior probabilities $p(E_i|x_{j,i})$, $\forall j \in T$. These can be used to estimate disease severities in test subjects by first estimating the conditional distribution $p(i|S, X_j)$, which estimates the probability that the first i events of S have occurred for a test-subject and the rest are yet to occur.

$$p(i|S, X_j) \propto \prod_{l=1}^i p(E_{S(l)}|x_{j,S(l)}) \times \prod_{l=i+1}^N p(\neg E_{S(l)}|x_{j,S(l)}) \quad (8.4)$$

The patient stage of a test subject (Υ_j) is defined as the expectation of $\lambda(i)$ with respect to the conditional distribution $p(i|S, X_j)$.

$$\Upsilon_j = \frac{\sum_{i=1}^N \lambda(i)p(i|S, X_j)}{\sum_{i=1}^N p(i|S, X_j)} \quad (8.5)$$

8.2.2 Group-specific and group-aspecific parts of DEBM

We propose extensions of DEBM for stratified populations, i.e., when the dataset X can be subdivided in groups $g \in [1, G]$, based on, e.g., genotype or phenotype of the subjects. Since DEBM is a data-driven model, data stratification into smaller groups would lead to more inaccurate models [93]. To obtain better DEBM accuracies in such scenario, we propose to co-train DEBM for estimating disease timelines $\forall g$ by splitting DEBM into group-aspecific and group-specific parts. The group-aspecific parts of DEBM are estimated using the entire dataset and group-specific parts are estimated for each group independently.

We first discuss the default way of independently training DEBM in the different groups and then propose two different approaches for splitting DEBM into group-aspecific and group-specific parts.

Approach 1: Independent DEBM

In this default approach, each group is considered as an independent dataset and the disease progression timeline in each group is estimated independently. GMM in such a scenario is illustrated in Figure 8.3a.

Approach 2: Coupled DEBM

$$\text{DEBM} \rightarrow \begin{cases} p(x_{\cdot,i} | \neg E_i), p(x_{\cdot,i} | E_i) & \text{group-aspecific} \\ \theta_{i,g}, \{S_g, \lambda_g\}, & \text{group-specific} \end{cases} \quad (8.6)$$

In this approach, we assume that the different groups share the normal and abnormal PDFs, but the ordering in which these biomarkers become abnormal are different. The mixing parameters $(\theta_{i,g})$ are considered as group-specific part of the DEBM algorithm because the proportion of subjects with normal and abnormal biomarker values in each group g is correlated with the position of the biomarker along the ordering S_g , which we expect to be different in each group.

Hence, in our approach, we modify the alternating GMM optimization scheme to jointly optimize the GMM parameters of multiple groups. First, the GMM algorithm is initialized without considering the groups, as explained in Section 8.2.1. Secondly, as with the default DEBM, Gaussian parameters and mixing parameters are alternately optimized. In contrast in coupled DEBM, the Gaussian parameters are estimated jointly for all groups, while mixing parameters are estimated separately for each group. This has been illustrated in Figure 8.3b.

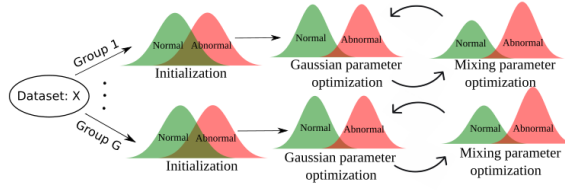
Once the GMM optimization has been performed, S_g and λ_g are estimated in each group. Patient staging (Υ_j) of the test-subjects in group g are computed based on the disease progression timeline $\{S_g, \lambda_g\}$.

Approach 3: Co-init DEBM

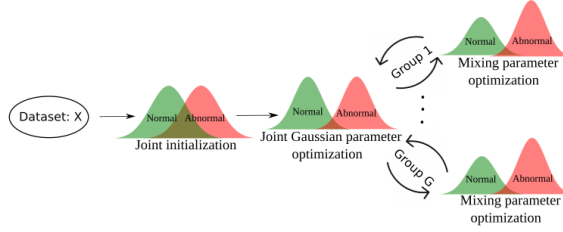
$$\text{DEBM} \rightarrow \begin{cases} \hat{p}(x_{\cdot,i} | \neg E_i), \hat{p}(x_{\cdot,i} | E_i) & \text{group-aspecific} \\ p_g(x_{\cdot,i} | \neg E_i), p_g(x_{\cdot,i} | E_i) & \text{group-specific} \\ \theta_{i,g}, \{S_g, \lambda_g\} & \text{group-specific} \end{cases} \quad (8.7)$$

In this approach, we assume that the different groups do not share the normal and abnormal PDFs, but that they are close to each other. Hence, in co-init DEBM, we relax the constraint on $p(x_{\cdot,i}|\neg E_i)$ and $p(x_{\cdot,i}|E_i)$ and instead consider the initialized values of normal and abnormal PDFs ($\hat{p}(x_{\cdot,i}|\neg E_i)$ and $\hat{p}(x_{\cdot,i}|E_i)$) to be group-specific part of DEBM. We estimate $p_g(x_{\cdot,i}|\neg E_i)$ and $p_g(x_{\cdot,i}|E_i)$ independently for each group. This is illustrated in Figure 8.3c.

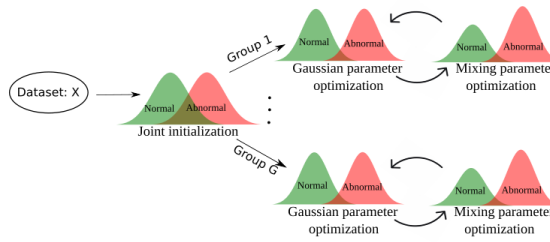
As with the previous approach, S_g , λ_g and the patient staging of the test-subjects in group g are computed independently for each group.



(a) GMM in independent DEBM



(b) GMM in Coupled DEBM



(c) GMM in Co-init DEBM

Figure 8.3: Overview of GMM optimization strategies in the different approaches for DEBM analysis in stratified populations. (a) The default approach in which GMM in each group is trained independently. (b) GMM in coupled DEBM, where the different groups share the Gaussian parameters, but the mixing parameters are estimated independently. (c) GMM in co-init DEBM in which the different groups are jointly initialized before the GMM optimization, but the optimization is done independently for each group.

8.3 Experiments

Section 8.3.1 describes the experiments to evaluate the proposed DEBM approaches on a stratified population. Since ground-truth orderings are unknown in real clinical data, we use simulated datasets for evaluating the methods. After evaluating the proposed approaches, we select the best approach for analyzing the effect of *APOE* on AD progression using subjects from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database. Section 8.3.2 describes the details of these experiments.

8.3.1 Simulation Experiments

We used the framework developed by [79] for simulating cross-sectional data consisting of scalar biomarker values for CN, MCI and AD subjects in two groups. In this framework, disease progression in a subject is modeled by a series of biomarker changes representing the temporal cascade of biomarker abnormality as estimated by an EBM. Individual biomarker trajectories are represented by sigmoids varying from the biomarker’s normal value to its abnormal value. To account for inter-subject variability, the normal and abnormal values for different subjects are drawn randomly from Gaussian distributions.

The simulation dataset used in our experiments are based on a set of seven biomarkers as described in the simulation experiments of [93]. The simulated datasets were stratified into two groups, with each group having its own distinct disease progression patterns. There are two ways in which the progression of disease in the groups can differ: 1. difference in ground-truth orderings S_1 and S_2 ; 2. difference in the abnormal biomarker PDFs in the two groups *i.e.* $p_1(x_{\cdot,i}|E_i)$ and $p_2(x_{\cdot,i}|E_i)$. Each of these differences could affect the accuracy of the proposed approaches. Hence, we evaluated the proposed approaches in the presence of each of these differences. Normalized Kendall’s Tau distance between the estimated ordering (S) and the ground-truth ordering (S_{gt}) was used as an evaluation measure in these experiments:

$$\varepsilon_S = K(S, S_{gt}) / \binom{N}{2} \quad (8.8)$$

where $K(A, B)$ is the number of swaps required to obtain ordering B from ordering A.

The normalization ensures that ε_S falls in the range $[0, 1]$, with 0 as the distance when the two orderings are the same, and 1 as the distance when the two orderings are the reverse of each other.

Experiment 1: The first simulation experiment studied the effect of the difference in ordering between the two groups. The ordering in the first group (Group 1) was fixed and the ordering in the second group (Group 2) was selected randomly such that the normalized Kendall’s Tau distance between the two groups was a fixed number, say ε_O . ε_O was varied from 0 to 1 in steps of 0.2. The number of subjects in Group 2 was kept constant at 900. The number of subjects in Group 1 was varied from 100 to 900 in steps of 200, to study how the different approaches perform in small as well as large groups. The normal and abnormal biomarkers levels in the two groups were sampled from the same Gaussian distribution for this experiment. We generated 50 random repetitions of the simulated datasets, and reported mean and standard deviation of ε_S for independent DEBM, coupled DEBM, and co-init DEBM in groups 1 and 2.

Experiment 2: This experiment studied the performance of the proposed approaches with the $\mu_{g,i,E}$ parameter of the $p_g(x_{\cdot,i}|E_i)$ distribution being different in the two groups. $\mu_{1,i,E}$ was fixed, and

$\mu_{2,i,E}$ was varied such that the difference $\mu_{2,i,E} - \mu_{1,i,E} (\varepsilon_G)$ was one of $\{-0.2d, 0, +0.2d\}$ where $d = \mu_{1,i,E} - \mu_{1,i,\neg E}$. 0 is considered the reference level, where the abnormal Gaussians are the same in the two groups. $\mu_{g,i,\neg E}$ were kept the same in the two groups. Hence, when $\varepsilon_G = -0.2d$, the abnormal biomarker levels are closer to the normal biomarker levels in Group 2 than in Group 1. This results in Group 2 biomarkers being weaker than their Group 1 counterparts when $\varepsilon_G = -0.2d$ and stronger when $\varepsilon_G = +0.2d$. The number of subjects in Group 2 was kept a constant at 900, while the subjects in Group 1 increased from 100 to 900. ε_O between the two groups was fixed at 0.4. We again generated 50 random repetitions of the simulated datasets, and reported mean and standard deviation of ε_S for coupled DEBM, co-init DEBM and DEBM.

These experiments were used to evaluate the different approaches mentioned in Section 9.2 and select the best method for analyzing the effect of *APOE* alleles in AD progression.

8.3.2 Studying the effect of *APOE*

We considered the baseline measurements from 417 CN, 235 MCI converters and 342 AD subjects in ADNI1, ADNIGO and ADNI2 studies*. The MCI converters are subjects who had MCI at baseline but converted to AD within 3 years of baseline measurement. We excluded subjects with significant memory concerns (without a diagnosis of AD or MCI) and MCI non-converters in our experiments to select a more phenotypically homogeneous group of subjects with prevalent or incident AD. In each of the experiments, the dataset was divided into three groups ($\varepsilon 2$ carriers, homozygous $\varepsilon 3$ carriers, and $\varepsilon 4$ carriers) based on the subject's *APOE* carriership [284]. Subjects with *APOE* $\varepsilon 2, 4$ ($n=34$) were not included in either group because of the presence of both $\varepsilon 2$ and $\varepsilon 4$ alleles.

Subject demographics and their *APOE* carrierships are summarized in Table 8.1. The modalities considered were structural imaging biomarkers, biomarkers extracted from cerebrospinal fluid (CSF), and cognitive biomarkers. Structural imaging biomarkers were obtained from T1-weighted MRI acquired at 1.5T or 3T. Details of the MRI acquisition protocols of ADNI can be found in [289, 290].

Imaging biomarkers were estimated from T1-weighted MRI scans analysed with FreeSurfer software v6.0 cross-sectional stream and outputs were visually checked. We assumed a symmetric pattern of atrophy in AD and averaged imaging biomarkers between the left and right hemisphere.

Experiment 3: For this experiment, the selected imaging biomarkers were: hippocampal volume, volume of the entorhinal cortex, fusiform gyrus volume, middle-temporal gyrus volume, precuneus volume, together with whole brain volume and volume of the ventricles [105, 106, 291]. The selected CSF based biomarkers were: CSF concentrations of Amyloid- β_{42} (ABETA), total Tau (TAU) and phosphorylated Tau₁₈₁ (PTAU) proteins [23, 103], Neurogranin [292] and Neurofilament light chain [293, 294]. Mini mental state examination (MMSE) and Alzheimer's Disease Assessment Scale - Cognitive (13 items) (ADAS13) were used as cognitive biomarkers. The availability of these multimodal biomarkers in the

*The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). For up-to-date information, see www.adni-info.org.

| Demographics | | | |
|------------------------------------|----------------|----------------|----------------|
| Diagnosis | CN | MCIc | AD |
| n | 417 | 235 | 342 |
| $APOE$ 2*/33/*4 | 57/244/110 | 6/66/156 | 12/101/219 |
| Sex M/F | 209/208 | 145/90 | 189/153 |
| Age [yrs.] ($\mu \pm \sigma$) | 74.8 \pm 5.7 | 73.7 \pm 7.0 | 75.0 \pm 7.8 |
| Edu [yrs.] ($\mu \pm \sigma$) | 16.3 \pm 2.7 | 15.9 \pm 2.7 | 15.2 \pm 3.0 |

Table 8.1: Demographics for the used population. 2* represents the subjects with $APOE$ alleles $\varepsilon 2$, 2 and $\varepsilon 2$, 3. 33 represents the subjects with reference $APOE$ allele $\varepsilon 3$, 3. *4 represents the subjects with $APOE$ alleles $\varepsilon 3$, 4 and $\varepsilon 4$, 4. Subjects with both $\varepsilon 2$ and $\varepsilon 4$ alleles were excluded from this study ($n=34$). Edu. is an abbreviation used for Education.

ADNI database is summarized in Table 2.

We downloaded the CSF measurements from the ADNI database. The measurements of ABETA, TAU and PTAU had been made using the microbead-based multiplex immunoassay, the INNO-BIA AlzBio3 RUO [295]. The measurement of NFL had been made with enzyme-linked immunosorbent assay NF-light ELISA kit [296]. NG had been measured by electrochemiluminescence technology (Meso Scale Discovery) using a monoclonal antibody specific for NG (Ng7) for coating together with a detector antibody polyclonal neurogranin anti-rabbit (ab 23570, Upstate) [297]. As described previously in [93], the TAU and PTAU measurements were transformed to logarithmic scales to make the distributions less skewed and more suitable for DEBM analysis.

The volumes of the selected regions were regressed with age, sex and intra-cranial volume (ICV) and the effects of these factors were subsequently corrected for, before being used as biomarkers. The effects of age and sex were regressed out of CSF features, whereas effects of age, sex and education were regressed out of cognitive scores.

For the 12 selected biomarkers, we estimated the disease timelines in the three aforementioned groups using the method selected after simulation experiments. We studied the positional variance of the estimated orderings by creating 100 bootstrapped samples of the data. In order to evaluate if the estimated orderings in the three groups were significantly different from one another, we used permutation testing and estimated the distribution of the Kendall's Tau distance under the null hypothesis. To compute this distribution, we generated 10,000 random permutations of the three groups. We then computed the one-sided p -values for the actual Kendall's Tau distances between the orderings of the three groups, calculated as the proportion of sampled permutations where the distance was greater than or equal to the actual distance, and using Bonferroni correction to account for multiple testing.

Experiment 4: In this experiment, we validated the disease stage (Υ_j) by computing its correlation with the subjects' MMSE and ADAS13 values. We used a 10-fold cross validation, where the training set was used to estimate the disease timeline in the aforementioned groups and the test subjects' disease stage was evaluated by placing them on this disease timeline. We used the volume-based and CSF-based biomarkers

| Biomarker | Biomarker Availability | | |
|-----------|--|---|---|
| | $\varepsilon 2$ carriers ($N = 75$) | Homozygous $\varepsilon 3$ carriers ($N = 411$) | $\varepsilon 4$ carriers ($N = 485$) |
| Imaging | 74 | 408 | 481 |
| ABETA | 57 | 301 | 357 |
| PTAU | 57 | 301 | 357 |
| TAU | 57 | 299 | 348 |
| NG | 21 | 113 | 131 |
| NFL | 23 | 118 | 137 |
| MMSE | 75 | 411 | 485 |
| ADAS | 74 | 410 | 477 |

Table 8.2: Biomarker availability in number of subjects in the *APOE* based groups of $\varepsilon 2$ carriers, homozygous $\varepsilon 3$ carriers, and $\varepsilon 4$ carriers

from Experiment 3, but excluded MMSE and ADAS13 scores from the model.

8.4 Results

8.4.1 Simulations

Experiment 1: Figures 8.4 (a) and (b) show the ordering errors (ε_S) in Group 1 of the simulation datasets for DEBM, coupled DEBM and co-init DEBM as a function of number of subjects in Group 1, when ε_O between the two groups changes from 0 to 1. Figures 8.4 (c), (d) and (e) show ε_S in Group 2 of the simulation datasets for the aforementioned methods, as a function of number of subjects in Group 1. In our experiments, Group 1 dataset remains the same while Group 2 dataset changes as ε_O increase. Hence DEBM results do not change with change in ε_O in Figure 8.4 (a) and (b), whereas in Figure 8.4 (c), DEBM results do not change with increase in number of subjects in Group 1.

It can be seen that both coupled-training methods (i.e., co-init DEBM and coupled DEBM) outperform the default method of independently training DEBM models. It can also be observed that in both co-init DEBM and coupled DEBM the ordering errors decrease as ε_O increases and that co-init DEBM outperforms coupled DEBM for lower values of ε_O , whereas the performance is on par with coupled DEBM for higher values of ε_O .

Experiment 2: Figures 8.5 (a) and (b) show ε_S in Group 1 and Figures 8.5 (c), (d) and (e) show the same in Group 2, when varying ε_G . Even with $\varepsilon_G \neq 0$, coupled training (i.e., co-init DEBM and coupled DEBM) outperformed the default method of independently training DEBM models. Co-init DEBM showed negligible change in the errors when $\varepsilon_G \neq 0$. The performance of coupled DEBM in Group 1 worsened for $\varepsilon_G = +0.2d$ (Figure 8.5 (a)) and in Group 2 for $\varepsilon_G = -0.2d$ (Figure 8.5 (d)).

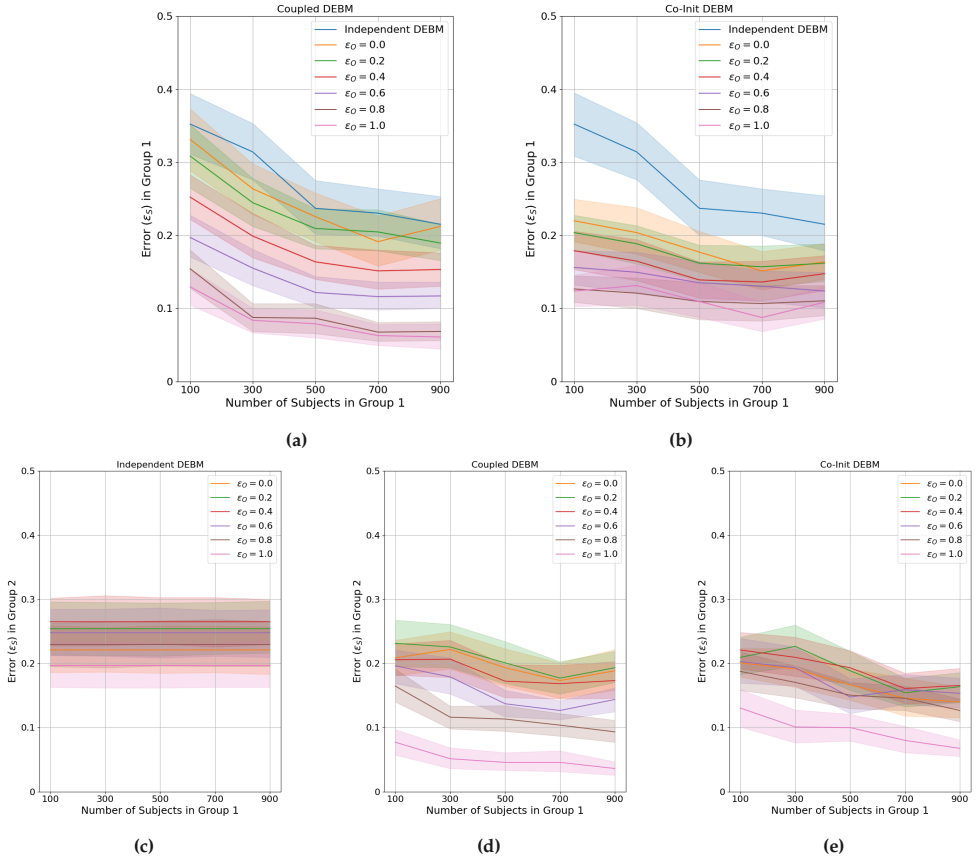


Figure 8.4: Experiment 1: The effect of ε_O (the difference in groundtruth event orderings in the two groups) on the performance of the proposed methods. The shaded region in these plots represents standard deviation of the error in estimation of the proposed methods in 50 random iterations of simulations. The plots in (a) and (b) show the ordering errors in Group 1 using Coupled DEBM and Co-init DEBM with independent DEBM shown in both (a) and (b), as a function of number of subjects in Group 1. The plots in (c), (d) and (e) show the ordering errors in Group 2 using independent DEBM, Coupled DEBM and Co-init DEBM respectively as a function of number of subjects in Group 1.

8.4.2 Studying the effect of APOE

The results in Experiments 1 and 2 show that the performance of co-init DEBM is more accurate and robust than coupled DEBM in most scenarios. We hence analyzed Experiments 3 and 4 using co-init DEBM.

Experiment 3: Figure 8.6 shows orderings of CSF, global cognition and volumetric biomarkers in the APOE based groups of $\varepsilon 2$ carriers, homozygous $\varepsilon 3$ carriers, and $\varepsilon 4$ carriers along with their uncertainty estimates. It can be seen that the uncertainty of the ordering in the $\varepsilon 2$ carriers group was high. Despite this uncertainty, some biomarkers (i.e. MMSE, NG and PTAU) seem to occur earlier than the other biomarkers in this group.

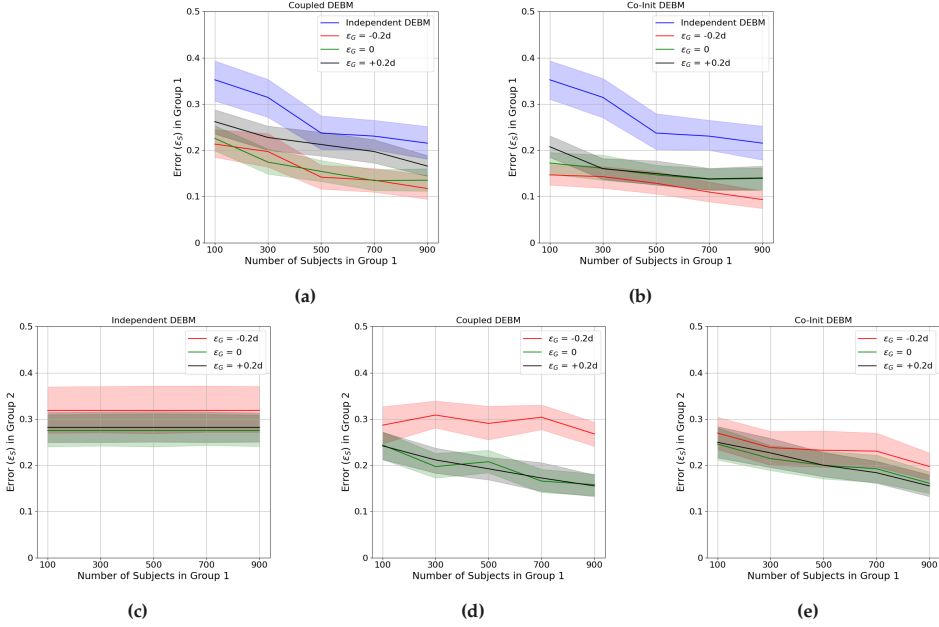


Figure 8.5: Experiment 2: The effect of ϵ_G (difference in abnormal biomarker levels in the two groups), on the performance of the proposed methods. The shaded region represents standard deviation of the error in 50 random iterations. The plots in (a) and (b) show the ordering errors in Group 1 using Coupled DEBM and Co-init DEBM with independent DEBM shown in both (a) and (b), as a function of number of subjects in Group 1. The plots in (c), (d) and (e) show the ordering errors in Group 2 using independent DEBM, Coupled DEBM and Co-init DEBM respectively as a function of number of subjects in Group 1.

In the homozygous ϵ_3 carrier group, ABETA was very prominently the earliest biomarker, followed by cognitive scores of MMSE and ADAS13. Among the CSF biomarkers, PTAU followed immediately after ABETA, which was in turn followed by TAU. NFL and NG were late biomarkers. Among the structural biomarkers, volumes of fusiform and middle-temporal gyri were the first to become abnormal, followed by ventricular volume and wholebrain volume. Hippocampus, precuneus and entorhinal volumes were late biomarkers in this group.

In the ϵ_4 carrier group, the CSF biomarkers followed a pattern that was similar to that of the homozygous ϵ_3 carrier group. The cognitive biomarkers were early biomarkers in this group as well. However the ordering in structural biomarkers was very different from that in the homozygous ϵ_3 carrier group. Hippocampus and entorhinal volumes were early biomarkers in this group, followed by middle-temporal and fusiform gyri volumes. Wholebrain, ventricular and precuneus volumes were late biomarkers.

The ordering of the ϵ_2 carrier group was significantly different from that of the homozygous ϵ_3 carrier group ($p = 0.0156$, after Bonferroni correction for multiple testing). Similarly, the orderings for the other two groups were significant as well: $p = 0.0147$ for the difference between ϵ_2 carrier group and ϵ_4 carrier group and $p = 0.0003$ for the difference between the homozygous ϵ_3 carrier group and ϵ_4 carrier group.

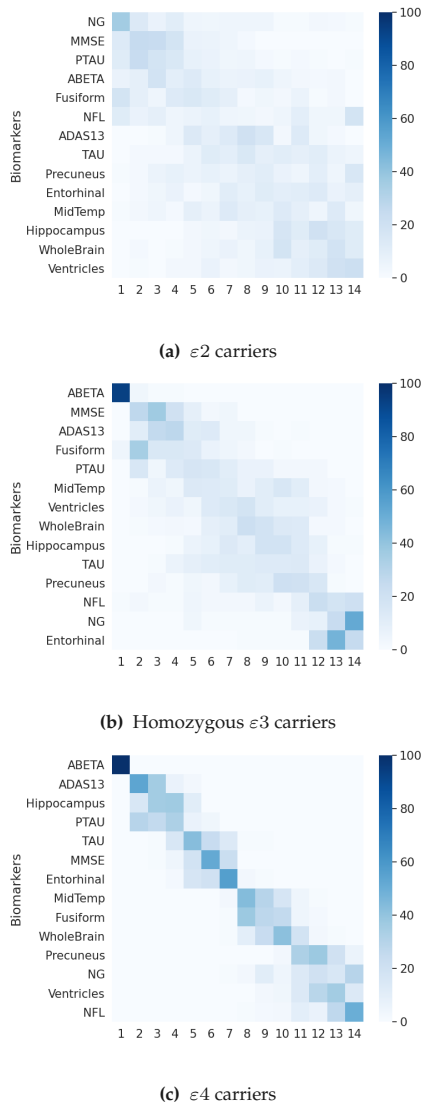


Figure 8.6: Experiment 3: Orderings of CSF, global cognition and volumetric biomarkers in the $APOE$ based groups of $\epsilon 2$ carriers, homozygous $\epsilon 3$ carriers, and $\epsilon 4$ carriers along with their uncertainty estimates. Uncertainty in the estimation of the ordering was measured by 100 repetitions of bootstrapping, in the three $APOE$ based groups. The color-map is based on the number of times a biomarker is at a position in 100 repetitions of bootstrapping. The number of subjects in the three groups were 75, 411 and 485 respectively. The orderings were obtained using Co-init DEBM.

Experiment 4: The variation of MMSE and ADAS13 scores with respect to the estimated disease stages has been plotted in Figure 8.7, for all three groups. The patient stages showed a significant correlation with both MMSE and ADAS13 scores. The correlation coefficients were also comparable in the three

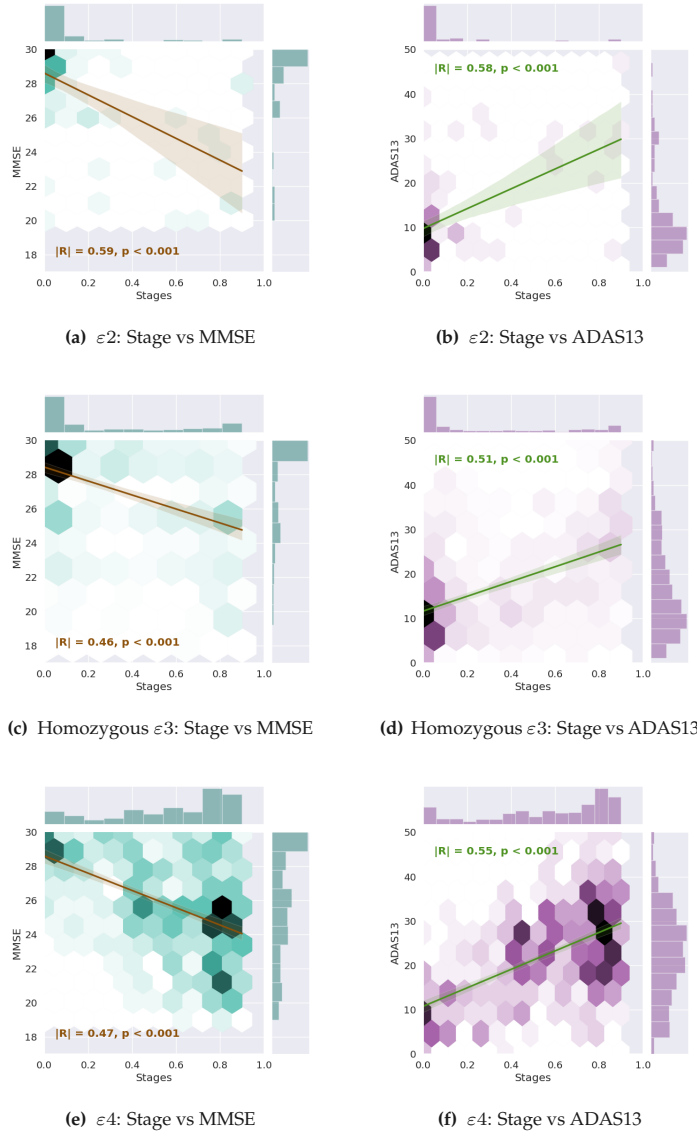


Figure 8.7: Experiment 4: Correlation of estimated disease stages with MMSE and ADAS scores in the APOE based groups of $\epsilon 2$ carriers, homozygous $\epsilon 3$ carriers, and $\epsilon 4$ carriers. The plot on top of each subfigure shows the probability density function of the disease stages, and the plot on the right of each subfigure shows the probability density function of the cognitive score in the subfigure. The 2D plot in each subfigure shows the joint density function of the two axes. The line in each subfigure shows the linear regression of MMSE / ADAS scores with the estimated disease stage and the shaded area around the line shows its 95% confidence interval. Figures (a),(c) and (e) depict correlation between MMSE score and obtained disease stages in the three APOE based groups. Figures (b), (d) and (f) depict correlation between ADAS13 score and the obtained disease stages in the three APOE based groups.

groups.

8.5 Discussion

DEBM models have been shown to be effective in determining the temporal cascade of biomarker abnormality as AD progresses, from cross-sectional data. In this work, we introduced a novel concept of splitting the different steps of DEBM into group-specific and group-aspecific parts for coupled training in stratified population. We considered two novel variations to split the steps of DEBM in this manner and through thorough experimentation in simulation datasets we observed that co-init DEBM helps in obtaining more accurate orderings in a stratified population. Using this method, we estimated the biomarker cascades in AD progression with $\epsilon 2$ alleles, homozygous $\epsilon 3$ alleles, and $\epsilon 4$ alleles of *APOE*, based on cross-sectional ADNI data. While the findings in the homozygous $\epsilon 3$ carrier and $\epsilon 4$ carrier groups fit the current understanding of progression of AD with high-confidence, the finding in the $\epsilon 2$ carrier group shows evidence for an alternative pathway (with relatively low confidence). In this section, we discuss the insights provided by the simulation experiments (Section 8.5.1) used for method selection as well as the insights into the AD progression pathways provided by our experiments on the ADNI dataset (Section 8.5.2).

8.5.1 Choice of the method

Coupled DEBM and co-init DEBM both split DEBM into group-specific and group-aspecific steps for coupled training of an EBM in stratified populations. Experiment 1 and 2 showed that coupled training of the group-aspecific parts of DEBM and independently training the group-specific parts of DEBM results in more accurate orderings in the groups better than the default approach of independently training a DEBM model in each group.

While splitting DEBM into group-specific and group-aspecific parts, we started with the assumption that the latent true normal and abnormal biomarker distributions in the groups are either same or similar. The difference between co-init DEBM and coupled DEBM is that, co-init DEBM accounts for slight differences in the underlying biomarker distributions between the groups whereas coupled DEBM does not.

The simulation dataset generated in Experiment 1 had the same true normal and abnormal biomarker distributions in the different groups, from which the simulated subjects were randomly sampled, aligning well with the assumption of coupled DEBM. However, this did not result in overall better accuracies for coupled DEBM than that of co-init DEBM. Co-init DEBM was also more robust than coupled DEBM as its accuracy was less dependent on ϵ_O , the distance between the ground-truth orderings in the two groups.

Another observation in Experiment 1, which was rather counter-intuitive, was that the errors made by the co-init and coupled DEBM models decreased as the distance between the ground-truth orderings in the two groups increased. When the orderings are further apart, the combined biomarker distributions in CN and AD groups have a larger overlap. The non-overlapping initialization (before the GMM optimization) thus results in the normal and abnormal distributions to be further apart. We hypothesize that this results in a better estimation of the mixing parameters during GMM optimization and in-turn resulted in more accurate orderings, as mixing-parameters are dependent on the biomarker's position in the ordering.

In Experiment 2, we checked the performance of our approaches when the assumption (true normal and abnormal biomarker distributions being same across groups) is violated in the dataset. This experiment showed that the orderings obtained using co-init DEBM are more robust to differences between the abnormal Gaussians across groups than those obtained with coupled DEBM. With coupled DEBM, the error increased in the group with weaker biomarkers i.e., Group 1 in the case of $\varepsilon_G = +0.2d$ and Group 2 in the case of $\varepsilon_G = -0.2d$. This shows that coupled DEBM introduces a systematic bias in the estimation of ordering that is detrimental to the group with weaker biomarkers. Co-init DEBM also showed a similar bias, but to a much lesser extent.

We hence selected co-init DEBM as the preferred approach for splitting and performed our analysis on ADNI dataset using this approach. We expect that this idea of splitting DEBM into group-specific and group-aspecific parts can be easily extended to the EBM introduced by [53].

8.5.2 Cascade of biomarker changes in the APOE based groups

Dividing the total population into groups based on APOE carriership enabled us to create more phenotypically homogeneous groups [288], each with potentially specific disease progression timeline. In this section, we discuss our results in these APOE carriership based groups.

Our findings show that the three APOE-carriership based groups have significantly different temporal cascades of disease progression. This suggests that the underlying pathways of progression are different for the three genotypes. Among the CSF biomarkers in the homozygous $\varepsilon 3$ carrier and the $\varepsilon 4$ carrier groups, ABETA abnormality is the earliest biomarker event followed by PTAU. This fits current understanding of AD progression [117]. It also confirms the need for preventing the accumulation of ABETA in high-risk patients. NFL and NG are late biomarkers in the homozygous $\varepsilon 3$ carrier and $\varepsilon 4$ carrier groups, which suggests that axonal [298] and synaptic [292] degeneration do not occur until very late in the disease process in these groups. NG being abnormal after PTAU and TAU in the homozygous $\varepsilon 3$ carrier and $\varepsilon 4$ carrier groups is also consistent with the previous findings that Tau mediates synaptic damage in AD [299].

In the $\varepsilon 2$ carrier group, we found that the abnormal NG and PTAU are the earliest CSF events, even before ABETA becomes abnormal. This could hint at the existence of an alternative pathway for the formation of tau tangles in the brain before ABETA accumulation, as suggested in [280], but needs more extensive validation.

Among the volumetric biomarkers, Entorhinal cortex is one of the early biomarkers in the $\varepsilon 4$ carrier group which is supported by the findings in [300], but is one of the last biomarkers to become abnormal in the homozygous $\varepsilon 3$ carrier group. Ventricular volume is a late biomarker in the $\varepsilon 4$ carrier group but it becomes abnormal quite early in the homozygous $\varepsilon 3$ carrier group as also observed by [301]. Hippocampus volume is the earliest biomarker in the $\varepsilon 4$ carrier group, but is a relatively late biomarker in the homozygous $\varepsilon 3$ carrier and $\varepsilon 2$ carrier groups. This suggests that incidence of hippocampal sparing AD [302] could correlate with APOE carriership.

The findings related to these orderings of biomarker events were validated by correlating the patient stages derived from these orderings with MMSE and ADAS13 scores. Patient stages of subjects in all

three groups, when used as test-subjects in a cross-validated manner, showed a significant correlation ($p < 0.001$) with these scores. These correlations validate our findings and suggest that these genotype-specific disease progression timelines could be used for patient monitoring.

8.6 Conclusion and Future work

We conclude that co-init DEBM provides the best accuracy and robustness when estimating orderings in stratified populations. Future work on co-init DEBM can focus on extending the approach for high-dimensional imaging biomarkers [303]. This work also provides groundwork for extending the method towards hypothesis-free, data-driven stratification of phenotypes.

We gained new insights into the disease progression timeline of AD in the *APOE* based groups of $\epsilon 2$ carriers, homozygous $\epsilon 3$ carriers, and $\epsilon 4$ carriers. While we observed that the estimated disease progression timelines in the $\epsilon 4$ carrier and the homozygous $\epsilon 3$ carrier groups fit the current understanding of AD progression with high confidence, the estimated timelines in the $\epsilon 2$ carrier group may suggest an alternative pathway for the formation of tau tangles in the brain before amyloid β accumulation, albeit with relatively low confidence. We expect that these genotype-specific disease progression timelines will benefit patient monitoring in the future, and may help optimize selection of eligible subjects for clinical trials.

Chapter 9

Progression along *APOE*-specific data-driven temporal cascades is predictive of Alzheimer's disease in a population-based cohort

This chapter contains the content of the manuscript 'Progression along *APOE*-specific data-driven temporal cascades is predictive of Alzheimer's disease in a population-based cohort. **Vikram Venkatraghavan**^{*}, Elisabeth J. Vinke^{*}, Esther E. Bron, Wiro J. Niessen, M. Arfan Ikram, Stefan Klein[†], Meike W. Vernooij[†], for the Alzheimer's Disease Neuroimaging Initiative. *Under review.*' Earlier version of this chapter was also presented (poster) at the Computational approaches for ageing and age-related diseases (CompAge) workshop 2020.

^{*} Contributed equally in this work

[†] Contributed equally in this work

Vikram Venkatraghavan's contributions in this chapter are: Methodology, software, validation, and writing.

Abstract

Data-driven disease progression models have provided important insight into the temporal cascade of brain changes in AD phenotypes. However, their utility in predicting the progression of pre-symptomatic AD in a population-based setting has not yet been investigated. In this study, we investigated if the temporal cascades constructed in a case-controlled setting, with subjects stratified according to *APOE* status, are generalizable to a population-based cohort, and if progression along these disease cascades is predictive of AD. Seven volumetric biomarkers derived from structural MRI were considered. We estimated *APOE*-specific temporal cascades of changes in these biomarkers using a recently proposed method called co-initialized discriminative event-based modeling (co-init DEBM). This method can also estimate a disease stage for new subjects by calculating their position along the temporal cascade. The model was trained and cross-validated on the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset, and tested on the population-based Rotterdam Study (RS) cohort. We compared the diagnostic and prognostic value of the disease stage in the two cohorts. Furthermore, we investigated if the rate of change of disease stage in RS participants with longitudinal MRI data was predictive of AD. In ADNI, the estimated disease stage distinguished AD subjects from controls with an AUC of 0.83 in both *APOE* ϵ 4 non-carriers and carriers. In the RS cohort, we obtained an AUC of 0.83 and 0.85 in ϵ 4 non-carriers and carriers, respectively. Progression along the temporal cascades as estimated by the rate of change of disease stage showed a significant difference ($p < 0.005$) for subjects with pre-symptomatic AD as compared to the general aging population in RS. It distinguished pre-symptomatic AD subjects with an AUC of 0.81 in *APOE* ϵ 4 non-carriers and 0.88 in carriers, which was better than any individual volumetric biomarker, or its rate of change, could achieve. Our results suggest that co-init DEBM trained on case-controlled data is generalizable to a population-based cohort setting and that progression along the temporal cascades is predictive of the development of AD in the general population. We expect that this approach can help to identify at-risk individuals from the general population for targeted clinical trials as well as to provide biomarker based objective assessment in such trials.

9.1 Introduction

Alzheimer’s disease (AD) is a chronic neurodegenerative disease that affects roughly 3% of the world’s elderly population (above 60 years old) [279]. Genetic variations and in particular *APOE* genotypes are not only the risk factors for the disease [27] but also affect its clinical [288, 304] and biological phenotypes [38].

Neuroimaging biomarkers play an important role in disentangling these phenotypes [86, 305]. They could also play an important role in finding disease modifying treatments [306]. There has been evidence that selection of the study population at its pre-symptomatic stage is also crucial for the success of potential modifying treatments for AD [44, 307]. Hence there is a crucial need for a way to objectively assess the progression of pre-symptomatic AD (or lack thereof).

Biomarkers extracted from neuroimaging data in combination with machine learning approaches have been shown to objectively assess the progression of AD in research cohorts [308] as well as in clinical cohorts [309]. However, machine learning approaches are not explainable by default and the lack of transparency in such approaches could hinder clinical decision making [310].

Disease progression models are data-driven approaches that are interpretable by design and can thus

aid not only in predicting AD but also in explaining the decision and facilitating transparency and trust [311]. In recent years, many disease progression models have emerged to provide insight into neurodegenerative diseases such as AD [52, 53]. Such insights have also been shown to aid in objective assessment of AD progression [58]. An example of such a model is the discriminative event-based model (DEBM) [93], which estimates a cascade of AD related biomarker abnormality events in a data-driven way. Such data-driven temporal cascades were also shown to be generalizable to multiple clinical cohorts [291]. However, the generalizability of such models to population-based cohorts and their utility in predicting the progression of pre-symptomatic AD in a population-based setting have not yet been investigated.

In this work, we investigate if i) *APOE*-specific temporal cascades constructed in a case-controlled setting are generalizable to a population-based cohort, and ii) if progression along these disease cascades is predictive of AD. For constructing the *APOE*-specific temporal cascades, we use a recently developed approach called co-initialized (co-init) DEBM [312] meant for obtaining temporal cascades in stratified cross-sectional datasets. We demonstrate the potential of the method's fine-grained disease stage estimation in predicting the subjects with pre-symptomatic AD in the general population.

9.2 Methods

We first describe the inclusion criteria for participants and the method for obtaining the volumetric biomarkers in the case-controlled Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and the population-based Rotterdam study (RS) dataset. This is followed by the description of co-init DEBM used to construct *APOE*-specific temporal cascades of volumetric biomarkers from baseline scans of the participants in the ADNI. We validated the temporal cascades constructed on ADNI by assessing their generalizability to the population-based RS cohort, and by predicting the participants at-risk of becoming symptomatic in the RS cohort.

9.2.1 Participants

ADNI

We considered the baseline measurements of 335 cognitively normal (CN), 565 non-AD, 167 incident-AD and 223 AD participants (prevalent-AD) who had imaging data available in ADNI1, ADNIGO and ADNI2 studies*. The non-AD cases were defined as ADNI participants who were either mild cognitively impaired (MCI) or had subjective memory complaints at the time of the baseline MRI scan, and did not develop AD within 3 years of follow-up. The incident-AD cases presented with MCI at baseline but developed AD within 3 years. Characteristics of the subjects and their volumetric measures in the ADNI dataset included in our study are shown in Table 9.1(a).

*ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). For up-to-date information, see www.adni-info.org.

Rotterdam Study

We considered participants from the population-based RS cohort, a prospective longitudinal study among community-dwelling subjects aged 45 years and over [313]. Participants were screened for dementia at baseline and at follow-up examinations with the Mini-Mental State Examination and the Geriatric Mental Schedule organic level. Those with a Mini-Mental State Examination score < 26 or Geriatric Mental Schedule score > 0 underwent further investigation and informant interview, including the Cambridge Examination for Mental Disorders of the Elderly. In addition, the entire cohort was continuously under surveillance for dementia through electronic linkage of the study database with medical records from general practitioners and the regional institute for outpatient mental health care. Available information on cognitive testing and clinical neuroimaging was used when required for diagnosis of dementia subtype. A consensus panel led by a consultant neurologist established the final diagnosis of AD.

In this work, we included participants from the RS who had at least one MRI scan, who completed cognitive testing, and were interviewed for the presence of subjective cognitive complaints at the time of the MRI. The included participants were categorized into 4 groups: participants that were cognitively normal at the time of the scan (CN), participants that had subjective memory complaints and/or objective cognitive impairment [314], but who did not develop AD at follow-up (non-AD), participants with AD at the time of the scan (prevalent-AD) and participants who developed AD after the MRI scan (incident-AD). Unlike in ADNI, we did not set a threshold of conversion within 3 years to be included as an incident-AD participant, since we wanted to assess the utility of our method in monitoring the progression of both pre-clinical and prodromal AD subjects. Participants with clinical stroke were excluded.

In our experiments, we used two subsets of the RS cohort: the *generalizability set* and the *prediction set*. The generalizability set consisted of the last MRI scan available for each participant in the RS cohort. This subset consisted of 998 CN, 2710 non-AD, 97 incident-AD, and 25 prevalent-AD cases and were used for experiments validating the generalizability of the *APOE*-specific temporal cascades constructed using co-init DEBM. The characteristics of the subjects in this subset are shown in Table 9.1(b). The prediction set consisted of the last two MRI scans available for each participant, which were used to assess the progression (or lack thereof) of pre-symptomatic AD in the participants. This subset consisted of 183 CN, 852 non-AD and 31 incident-AD cases. For the incident-AD cases, both the included scans were performed before the AD diagnosis. Participants with prevalent-AD were excluded in this subset. The characteristics of the subjects in this subset are shown in Table 9.1(c).

9.2.2 MRI acquisition and imaging biomarker extraction

The imaging biomarkers used in this study were estimated from T1-weighted (T1w) MRI scans. Details of the MRI acquisition protocol can be found in [289, 290] (ADNI) and [315] (RS). The MRI scans were analyzed with FreeSurfer software v6.0 cross-sectional stream (<http://surfer.nmr.mgh.harvard.edu>). Outputs were visually checked for the ADNI dataset. In the RS dataset, an automated quality metric was used to exclude scans with insufficient quality, which was visually verified in a randomly selected subset of both selected and rejected scans [316].

The selected imaging markers were the same markers as that of [291], namely volumetric measures of: total brain, ventricles, hippocampus, precuneus, middle temporal gyrus, fusiform gyrus and entorhinal cortex. The volumes were defined as the summed volumes of the structure in the left and right

hemisphere. The volumetric measures of CN subjects in ADNI were used to regress against age, sex and intracranial volume to estimate their confounding effects. These estimates were used for confounding factor correction in the remaining subjects in ADNI as well as in the RS cohort. The resultant volumetric measures will be referred to as biomarkers in the remainder of the manuscript.

9.2.3 Construction of *APOE*-specific temporal cascades using co-init DEBM

The co-init DEBM model introduced in [312] constructs genotype-specific AD related temporal cascades of biomarker changes, based on *cross-sectional* datasets. Such an estimation from cross-sectional data is feasible because, in a cohort consisting of subjects encompassing a wide spectrum of severity, early biomarkers have a higher prevalence of abnormal biomarker values as compared to biomarkers that become abnormal later in the disease timeline. The co-init DEBM model estimates this timeline without strictly considering the diagnostic labels of the subjects. The model uses a coupled mixture model to jointly fit normal and abnormal distributions in the dataset stratified by (*APOE*) genotypes. The model assumes that the normal and abnormal biomarker distributions in the different genotypes can be approximately represented by Gaussians. It also assumes that the different genotypes' abnormal (and normal) biomarker distributions are close to each other.

After the estimation of the normal and abnormal biomarker distributions, the model computes the probability of abnormality of each biomarker for each subject in the training dataset. Based on the assumption that a biomarker that becomes abnormal earlier in the temporal cascade of events would be more abnormal than the biomarker that becomes abnormal later, it estimates a subject-specific cascade of biomarker changes in each subject of the dataset. A generalized Mallows model is used to average the subject-specific biomarker cascades over the subjects within each genotypic group of the training set, to construct average temporal cascades for *APOE* $\epsilon 4$ non-carriers and carriers. Along with the sequence of the biomarker abnormality events, the model also estimates the relative positioning of such events with respect to each other (event-centers). Absolute magnitudes for these event-centers are irrelevant as they only convey relative (temporal) distances and in this study, they were normalized such that the first event and the last event coincided at a value of 0.1 and 0.9 respectively.

To construct the temporal cascades, the co-init DEBM was trained on CN, incident-AD, and prevalent-AD subjects from ADNI. The non-AD subjects in ADNI were excluded for training the model, to reduce the chances of disorders unrelated to AD affecting the estimated cascades. The variance in the estimated temporal cascades was computed using 100 independent bootstrap samples.

9.2.4 Estimating *APOE*-specific disease stages

After training the co-init DEBM model, the constructed *APOE*-specific temporal cascades were used to estimate the disease stage at multiple timepoints for subjects of the RS cohort. For estimating the disease stages of ADNI subjects, we used a 10-fold cross validation. The training set was used for constructing the temporal cascades and the disease stages were estimated in the test set, including the non-AD subjects excluded in the training phase. Disease stage quantifies the severity of the disease in a subject by positioning them along the pre-constructed temporal cascade and is normalized between 0 and 1. The estimated disease stages were used in two sets of experiments.

| ADNI dataset | CN | non-AD | incident-AD | prevalent-AD |
|---|--------------------|--------------------|--------------------|--------------------|
| Number of subjects | 335 | 565 | 167 | 223 |
| Number of women, % | 174, 51.9 | 268, 47.4 | 68, 40.7 | 104, 46.6 |
| Age (years) | 74.3 \pm 5.6 | 71.82 \pm 7.2 | 73.1 \pm 7.1 | 74.0 \pm 7.9 |
| Number of APOE ϵ 4 carriers, % | 92, 27.5 | 238, 42.1 | 121, 72.5 | 151, 67.7 |
| Intracranial volume (ml) | 1504.0 \pm 155.8 | 1520.9 \pm 152.8 | 1546.2 \pm 180.2 | 1524.2 \pm 183.9 |
| Total brain volume (ml) | 1030.7 \pm 98.7 | 1043.3 \pm 100.0 | 1017.7 \pm 111.7 | 991.8 \pm 114.1 |
| Ventricle volume (ml) | 38.4 \pm 18.1 | 41.0 \pm 21.3 | 49.1 \pm 23.9 | 51.4 \pm 21.9 |
| Hippocampus volume (ml) | 7.3 \pm 0.9 | 7.1 \pm 1.0 | 6.3 \pm 1.0 | 6.0 \pm 1.0 |
| Precuneus volume (ml) | 16.7 \pm 2.2 | 17.4 \pm 2.4 | 16.2 \pm 2.6 | 15.4 \pm 2.5 |
| Middle temporal gyrus volume (ml) | 20.4 \pm 2.7 | 20.4 \pm 2.7 | 18.5 \pm 2.9 | 17.6 \pm 3.0 |
| Fusiform gyrus volume (ml) | 17.5 \pm 2.1 | 17.6 \pm 2.2 | 16.3 \pm 2.4 | 15.5 \pm 2.4 |
| Entorhinal cortex volume (ml) | 4.0 \pm 0.7 | 3.9 \pm 0.8 | 3.4 \pm 0.8 | 3.2 \pm 0.8 |
| Time before AD diagnosis (years)* | | | 1.4 \pm 0.7 | |

(a)

| RS dataset - generalizability set | CN | non-AD | incident-AD | prevalent-AD |
|---|--------------------|--------------------|--------------------|--------------------|
| Number of subjects | 998 | 2710 | 97 | 25 |
| Number of women, % | 500, 50.1 | 1200, 44.3 | 39, 40.2 | 10, 40.0 |
| Age (years) | 67.4 \pm 8.3 | 70.9 \pm 9.3 | 79.6 \pm 5.7 | 80.2 \pm 6.3 |
| Number of APOE ϵ 4 carriers, % | 255, 25.6 | 745, 27.5 | 45, 46.4 | 11, 44.0 |
| Intracranial volume (ml) | 1512.3 \pm 157.6 | 1475.8 \pm 155.3 | 1437.5 \pm 156.6 | 1403.0 \pm 163.9 |
| Total brain volume (ml) | 1050.3 \pm 107.5 | 1012.6 \pm 105.6 | 936.6 \pm 94.9 | 884.5 \pm 105.0 |
| Ventricle volume (ml) | 33.7 \pm 17.3 | 36.5 \pm 19.3 | 49.1 \pm 21.1 | 59.9 \pm 28.3 |
| Hippocampus volume (ml) | 7.9 \pm 0.8 | 7.6 \pm 0.8 | 6.7 \pm 0.9 | 6.0 \pm 1.0 |
| Precuneus volume (ml) | 18.2 \pm 2.1 | 17.6 \pm 2.0 | 16.8 \pm 1.9 | 15.4 \pm 2.2 |
| Middle temporal gyrus volume (ml) | 20.6 \pm 2.7 | 19.9 \pm 2.7 | 17.6 \pm 2.5 | 16.2 \pm 2.7 |
| Fusiform gyrus volume (ml) | 17.7 \pm 2.2 | 17.2 \pm 2.1 | 15.8 \pm 2.0 | 14.5 \pm 2.7 |
| Entorhinal cortex volume (ml) | 3.7 \pm 0.6 | 3.6 \pm 0.7 | 3.1 \pm 0.8 | 2.6 \pm 0.7 |
| Time before AD diagnosis (years)* | | | 2.8 \pm 2.3 | |

(b)

| RS dataset - prediction set | CN | non-AD | incident-AD |
|---|--------------------|------------------|--------------------|
| Number of subjects | 183 | 852 | 31 |
| Number of women, % | 95, 51.9 | 412, 48.4 | 10, 32.3 |
| Age (years)* | 73.3 \pm 5.5 | 75.5 \pm 6.4 | 78.4 \pm 6.8 |
| Follow-up time (years) | 3.5 \pm 1.3 | 3.5 \pm 1.4 | 2.9 \pm 0.9 |
| Number of APOE ϵ 4 carriers, % | 39, 21.3 | 225, 26.4 | 13, 41.9 |
| Intracranial volume (ml)* | 1522.8 \pm 156.6 | 1478.9 \pm 156 | 1419.4 \pm 126.9 |
| Total brain volume (ml)* | 1038.7 \pm 100.7 | 998.4 \pm 98.3 | 926.6 \pm 91.4 |
| Ventricle volume (ml)* | 39.7 \pm 20.2 | 41.1 \pm 21.6 | 44.9 \pm 17 |
| Hippocampus volume (ml)* | 7.8 \pm 0.8 | 7.4 \pm 0.8 | 6.7 \pm 0.9 |
| Precuneus volume (ml)* | 18.0 \pm 2.0 | 17.5 \pm 1.9 | 16.5 \pm 1.8 |
| Middle temporal gyrus volume (ml)* | 20.3 \pm 2.6 | 19.5 \pm 2.4 | 17.5 \pm 2.4 |
| Fusiform gyrus volume (ml)* | 17.5 \pm 2.1 | 17.0 \pm 2.1 | 15.6 \pm 2.1 |
| Entorhinal cortex volume (ml)* | 3.7 \pm 0.7 | 3.6 \pm 0.7 | 3.0 \pm 0.6 |
| Time before AD diagnosis (years)* | | | 2.4 \pm 1.8 |

(c)

Table 9.1: Characteristics of the ADNI dataset (a), the generalizability set of the RS dataset (b), and the prediction set of the RS dataset (c). * indicates values at last scan.

| Reference group | Cases | No. of Subjects in ADNI (Reference group, Cases) | No. of Subjects in RS (Reference group, Cases) | AUC in ADNI | AUC in RS |
|---|--------------|---|---|---------------------|---------------------|
| <i>APOE $\epsilon 4$ non-carriers</i> | | | | | |
| CN | prevalent-AD | 243, 72 | 743, 14 | 0.86 (0.81-0.91) | 0.85 (0.71-0.98) |
| CN + non-AD | prevalent-AD | 570, 72 | 2708, 14 | 0.83 (0.78-0.88) | 0.83 (0.70-0.97) |
| CN | incident-AD | 243, 46 | 743, 52 | 0.83 (0.77-0.90) | 0.70 (0.62-0.78) |
| CN + non-AD | incident-AD | 570, 46 | 2708, 52 | 0.81 (0.74-0.88) | 0.68 (0.60-0.75) |
| <i>APOE $\epsilon 4$ carriers</i> | | | | | |
| CN | prevalent-AD | 92, 151 | 255, 11 | 0.89 (0.85-0.94) | 0.85 (0.74-0.96) |
| CN + non-AD | prevalent-AD | 330, 151 | 1000, 11 | 0.83 (0.79-0.86) | 0.85 (0.74-0.95) |
| CN | incident-AD | 92, 121 | 255, 45 | 0.87 (0.82-0.92) | 0.63 (0.54-0.72) |
| CN + non-AD | incident-AD | 330, 121 | 1000, 45 | 0.79 (0.74-0.83) | 0.62 (0.54-0.71) |

Table 9.2: Generalizability assessment: The AUCs for distinguishing the different diagnostic classes using the estimated disease stages and their corresponding 95% confidence intervals. The confidence intervals of the AUCs were determined using bootstrap resampling while stratifying the diagnostic classes to maintain their relative proportions.

Experiment 1: *Assessing the generalizability of co-init DEBM from ADNI to RS*

In this experiment, we tested the generalizability of the co-init DEBM model trained on ADNI by evaluating the diagnostic and prognostic value of its predicted disease stages in the RS cohort. First we performed a visual assessment by constructing normalized histograms of the estimated *APOE*-specific disease stages for the different diagnostic classes in ADNI and the generalizability set of the RS cohort.

Complementing this visual analysis, for assessing the diagnostic value we used the estimated disease stages to distinguish prevalent-AD from two different reference groups in ADNI and in the generalizability set of the RS cohort. First, only the CN subjects were included in the reference group. To emulate a reference group of participants more representative of the general aging population than the CN group, we used a combined set of CN and non-AD subjects as the second reference group. We computed the area under the receiver operating curve (AUC) for distinguishing the diagnostic classes, and compared the AUCs obtained in ADNI and RS. The confidence intervals of these AUCs were measured using bootstrap resampling while stratifying the diagnostic classes to maintain their relative proportions.

For assessing the prognostic value, we used the estimated disease stages to distinguish incident-AD from the aforementioned two reference groups in ADNI and in the generalizability set of RS cohort. We computed the AUCs and their confidence intervals for distinguishing these diagnostic classes and compared values obtained in ADNI and RS. Furthermore, we computed the correlation of the estimated disease stages with time to dementia diagnosis for incident-AD subjects in ADNI as well as in RS.

Experiment 2: *Predicting AD based on longitudinal data in the RS cohort*

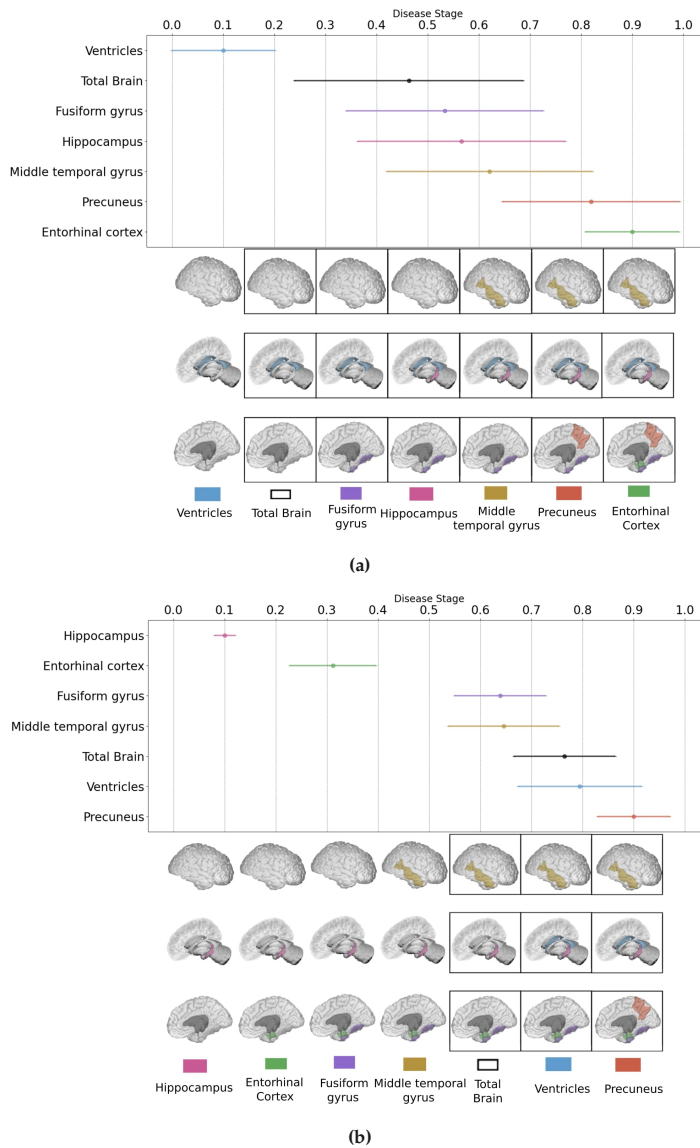


Figure 9.1: Temporal cascades of *APOE* $\epsilon 4$ non-carriers (a) and carriers (b) estimated using co-init DEBM in ADNI. The plot on top of each subfigure shows the event-centers of the different regions and their respective standard deviation estimated from a batch of 100 independent bootstrap samples. The 3D visualization [317] at the bottom of each subfigure highlights the region that becomes abnormal at the corresponding disease stage. The vertical positioning of the biomarkers in the event-center part of each subfigure shows the estimated temporal cascade in the *APOE* genotype, which is different for non-carriers and carriers.

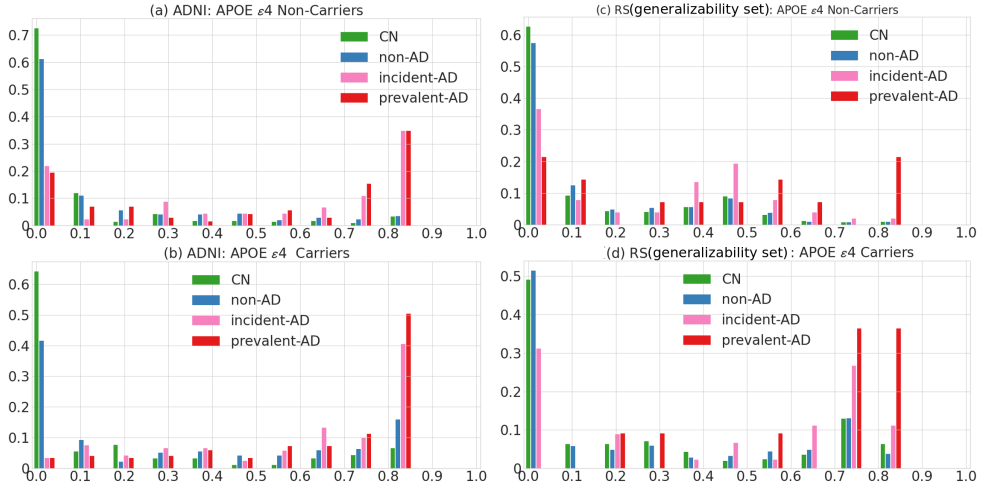


Figure 9.2: Normalized histograms of the estimated $APOE$ -specific disease stages for the different diagnostic classes in ADNI and the generalizability set of RS. The normalized histograms of disease stages are shown for (a) $APOE \epsilon 4$ non-carriers in ADNI, (b) $APOE \epsilon 4$ carriers in ADNI (c) $APOE \epsilon 4$ non-carriers of the generalizability set in RS, and (d) $APOE \epsilon 4$ carriers of the generalizability set in RS. The x-axis represents the disease stage based on the $APOE$ -specific temporal cascades by the co-init DEBM model, and the y-axis represents the relative percentage of subjects in each diagnostic class, meaning that the relative percentages of all disease stages of one diagnostic category add up to one.

In this experiment, we assess if the evolution of the disease stages derived from longitudinal neuroimaging data is predictive of AD in the prediction set of the RS cohort. This experiment is further divided into three parts. In the first part, we build longitudinal trajectories of the disease stages and observe the differences in CN, non-AD and incident-AD subjects. In the second part, we assess the prognostic value of the rate of change of disease stages. Lastly, we assess the marginal utility of the follow-up scans in AD prognostication.

Exp. 2.1: We used the disease stages obtained in the prediction set of the RS cohort for building the trajectories of disease stages in the two $APOE \epsilon 4$ based groups. The trajectories were estimated using linear mixed models with random intercepts and slopes. The time variable in these linear mixed models was follow-up time in years since the first MRI of the subject. To allow different slopes for different diagnostic classes, an interaction between follow-up time and the diagnosis was integrated in the model. Covariates that were accounted for in the model were sex, age at the time of the first MRI, and the interaction of age and follow-up time to allow slope differences for different ages.

Exp. 2.2: We used the rate of change of disease stages (delta disease stage) in the prediction set of the RS cohort to distinguish incident-AD from two different reference groups. As in Experiment 1, the two reference groups selected were CN, and a combined set of CN and non-AD subjects. We computed the AUCs and their confidence intervals for distinguishing these diagnostic classes. For comparison, the AUCs while using the rate of change of the volumetric measures (normalized to their respective intracranial volumes) for distinguishing the same two classes were computed.

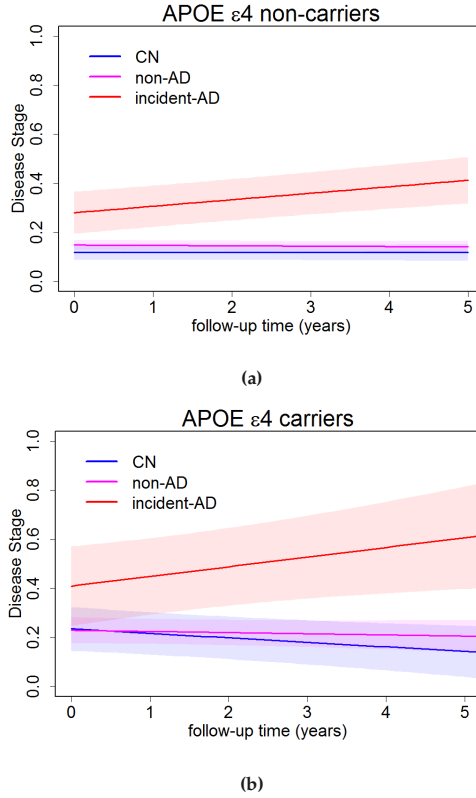


Figure 9.3: Average disease stage trajectories of participants within the prediction set of RS. The trajectories are shown separately for CN, non-AD and incident-AD subjects within the *APOE* $\epsilon 4$ non-carriers group (a) and the *APOE* $\epsilon 4$ carriers group (b). 95% confidence intervals are shown as shaded regions around the trajectories.

Exp. 2.3: Lastly, to evaluate the marginal utility of the follow-up scans for identifying incident-AD subjects, we used the estimated disease stage at the last MRI scan of the subjects in the prediction set of the RS cohort to distinguish incident-AD from the aforementioned two different reference groups. We computed the AUCs and their confidence intervals for distinguishing these diagnostic classes. As a comparison, the AUCs based on participants' age as well as of each individual volumetric imaging biomarker were also computed.

9.3 Results

Figure 9.1 shows the *APOE*-specific temporal cascades constructed for the $\epsilon 4$ non-carriers and carriers in the ADNI dataset. It shows the centers of the biomarker abnormality events along the cascade representing their relative positioning with respect to each other. It can be seen that the temporal cascade of *APOE* $\epsilon 4$ non-carriers and carriers were quite different. Most noticeably, ventricular volume and total brain volume were estimated as early biomarkers for *APOE* $\epsilon 4$ non-carriers, whereas hippocampal

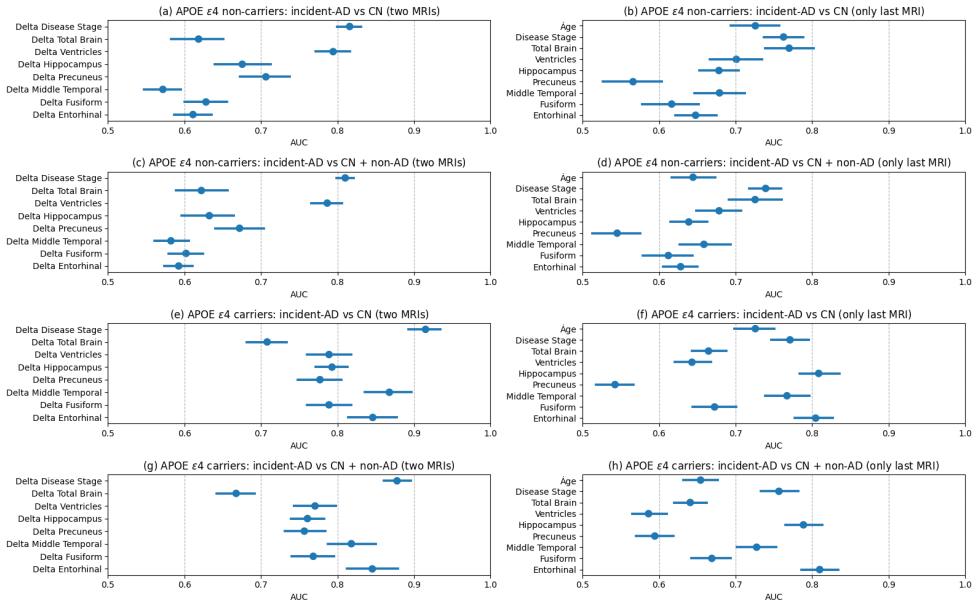


Figure 9.4: Predicting incident-AD subjects in the RS cohort. Figures (a), (c), (e), and (g) show the AUCs for distinguishing incident-AD while using data from two MRI scans based on longitudinal follow-up of the participants. Figures (b), (d), (f), and (h) show the AUCs for distinguishing incident-AD using only the last MRI scan available for each participant. Figures (a-d) are for carriers. Figures (a), (b), (e), and (f) use CN subjects as the reference group, while Figures (c), (d), (g), and (h) uses CN + non-AD as the reference group.

volume and volume of the entorhinal cortex were estimated as early biomarkers for *APOE* $\epsilon 4$ carriers. It can also be seen in Figure 9.1 that the uncertainty estimates in *APOE* $\epsilon 4$ non-carriers were greater than in *APOE* $\epsilon 4$ carriers.

Experiment 1: Assessing the generalizability of co-init DEBM from ADNI to RS

The normalized histograms of the estimated *APOE*-specific disease stages for the different diagnostic classes in ADNI and the generalizability set of RS are shown in Figure 9.2. It can be seen that the distributions of the disease stages of the four diagnostic classes in ADNI were largely similar to those in the generalizability set of RS. The CN and non-AD subjects were positioned towards the left side of the spectrum, whereas the prevalent-AD were positioned predominantly towards the right. It can also be seen that for a proportion of prevalent-AD subjects in the *APOE* $\epsilon 4$ non-carrier group, the model had estimated a low disease stage in both ADNI and RS cohorts. A noticeable difference between ADNI and RS was that a substantial proportion of incident-AD subjects in RS was positioned towards the left side of the histograms in both *APOE* $\epsilon 4$ non-carriers and carriers.

The AUCs for distinguishing the different diagnostic classes using the estimated disease stages are shown in Table 9.2, along with their confidence intervals. It can be observed that the performance of the disease

stages in distinguishing prevalent-AD from the set of CN and non-AD subjects in ADNI (AUC = 0.83 for both *APOE* $\epsilon 4$ non-carriers and carriers) was comparable to that in RS (AUC = 0.83 for *APOE* $\epsilon 4$ non-carriers and AUC = 0.85 for $\epsilon 4$ carriers). It should however be noted that the confidence intervals were larger in the RS cohort. It can also be observed that incident-AD subjects were harder to distinguish than prevalent-AD in the RS cohort (AUC = 0.68 for $\epsilon 4$ non-carriers and AUC = 0.62 for $\epsilon 4$ carriers), but not in ADNI (AUC = 0.81 for $\epsilon 4$ non-carriers and AUC = 0.79 for $\epsilon 4$ carriers).

Lastly, the estimated disease stages showed a significant Pearson correlation with time to diagnosis for *APOE* $\epsilon 4$ carrier incident-AD subjects in both ADNI ($R = 0.31$, $p = 0.0006$) and RS cohorts ($R = 0.29$, $p = 0.04$). However, the correlation was found to be insignificant for *APOE* $\epsilon 4$ non-carrier incident-AD subjects in both ADNI ($R = 0.04$, $p = 0.8$) and RS cohorts ($R = 0.1$, $p = 0.4$).

Experiment 2: Predicting AD based on longitudinal data in the RS cohort

Exp. 2.1: In Figure 9.3, the trajectories of disease stage over time as estimated by linear mixed models are shown for the CN, non-AD and incident-AD groups of the prediction set of RS. The interaction between the incident-AD diagnosis and follow-up time was statistically significant in both *APOE* $\epsilon 4$ non-carriers and carriers (CN vs. incident-AD $p = 0.0032$ and $p = 0.0041$ respectively; non-AD vs. incident-AD $p = 0.0039$ and $p = 0.0032$ respectively), meaning that incident-AD subjects showed a significant increase in disease stage compared to CN and non-AD subjects.

Exp. 2.2: In the left column of Figure 9.4, the AUCs and the corresponding 95% confidence intervals for distinguishing incident-AD using two MRI scans based on longitudinal follow-up of participants are shown for *APOE* $\epsilon 4$ non-carriers and carriers. It can be observed that for distinguishing incident-AD from the reference group, delta disease stage consistently performed the best for both the genotypes. It outperformed the rates of changes of volumetric measures, with respect to the obtained AUC. It can also be observed that distinguishing incident-AD from CN and non-AD subjects in the reference group was harder than distinguishing incident-AD from CN alone, as reflected by the lower AUCs for almost all the measures used.

Exp. 2.3: The right column of Figure 9.4 shows that age was an important predictor for incident-AD. Age distinguished incident-AD well from CN subjects (AUC of 0.73 for both $\epsilon 4$ non-carriers and carriers), but the performance of age as a predictor dropped substantially when distinguishing incident-AD from CN and non-AD subjects (AUC of 0.64 for $\epsilon 4$ non-carriers and 0.65 for $\epsilon 4$ carriers). When only the last MRI scan was used for incident-AD prediction from a reference group of CN and non-AD subjects, volumes of hippocampus and entorhinal cortex were good indicators in *APOE* $\epsilon 4$ carriers (AUC of 0.79 and 0.81 respectively) but not for *APOE* $\epsilon 4$ non-carriers (AUC of 0.64 and 0.63 respectively). Similarly, total brain volume and ventricle volume were good indicators of incident-AD in *APOE* $\epsilon 4$ non-carriers (AUC of 0.73 and 0.68 respectively), but not for $\epsilon 4$ carriers (AUC of 0.64 and 0.59 respectively). Disease stage estimated using the *APOE*-specific temporal cascades performed well consistently in both the *APOE* genotypes (AUC of 0.74 for $\epsilon 4$ non-carriers and 0.76 carriers). The marginal utility of an additional MRI scan can be observed by comparing the left column of Figure 9.4 with the right column of Figure 9.4. It can be seen that delta disease stage was much better for incident-AD prediction from a reference group of CN and non-AD subjects (AUC of 0.81 for $\epsilon 4$ non-carriers and 0.88 for carriers) than any measure obtained using only the last MRI scan.

9.4 Discussion

In this work, we constructed *APOE*-specific temporal cascades in a case-controlled setting and validated their generalizability to a population-based setting. We assessed that progression along these cascades is predictive of AD in the general population. In this section, we discuss the insights we obtained from our results.

9.4.1 Generalizability of the *APOE*-specific temporal cascades

The temporal cascades showed the *APOE* genotype-specific differences in the loss of structural integrity as AD progresses, with ventricular volume and total brain volume being early biomarkers for $\epsilon 4$ non-carriers, and hippocampal volume and volume of the entorhinal cortex being early biomarkers for $\epsilon 4$ carriers. We observed in the normalized histograms that for a proportion of prevalent-AD subjects in the $\epsilon 4$ non-carriers group, the model had estimated a low disease stage. This observation, in combination with the greater uncertainty of the event-centers in that group suggests that there is intra-genotype heterogeneity among the $\epsilon 4$ non-carriers.

We also observed that the normalized histograms of disease stages in the different diagnostic classes were visually largely similar for ADNI and RS. An important difference between the two cohorts was that the model estimated a low disease stage for a substantial proportion of incident-AD subjects in RS, but not in ADNI. Complementing the qualitative analysis, we also observed that the disease stages could distinguish prevalent-AD subjects from CN and non-AD subjects almost equally well in both ADNI and RS cohorts. However, we noticed a lower performance in distinguishing incident-AD from CN and non-AD subjects in RS as compared to ADNI. Three possible explanations for these differences between ADNI and RS are given below.

First, the incident-AD group in ADNI only consisted of prodromal AD subjects with the mean time to AD diagnosis of 1.4 years, whereas the incident-AD group in RS consisted of prodromal and preclinical AD subjects with the mean time to AD diagnosis of 2.8 years. We observed in Experiment 1 that the obtained disease stages of incident-AD subjects correlated with time to AD diagnosis for *APOE* $\epsilon 4$ carriers, making AD harder to detect in the preclinical phase than in the prodromal phase. Hence the difference in the mean time to diagnosis in the two datasets is expected to be a factor contributing to the observed lower performance in the RS cohort.

Secondly, the prodromal AD subjects in ADNI were clinically defined amnesic MCI subjects who have a much higher *a priori* chance of developing AD symptoms than in the general population, making the prediction in the latter cohort a more difficult problem.

Thirdly, a factor contributing to the performance difference could be that ADNI excluded subjects with severe cardiovascular risk factors whereas the RS did not. Hence the probability of co-morbidity of vascular pathology was higher in the RS incident-AD subjects than in the corresponding ADNI set, which could have led to the drop in performance.

Given the high AUCs for all other classification tasks, the comparable disease stage histograms in ADNI and RS, and the possible explanations given above for the specific differences related to incident-AD prediction, we conclude that the *APOE*-specific temporal cascades obtained by co-init DEBM are

generalizable from a case-controlled to a population-based setting.

However, for precise classification of subjects into either diagnostic category, a cut-off point for disease stage needs to be defined. We expect the cut-off point to be different in a case-controlled setting versus a population-based setting. Estimating this cut-off point in a population should ideally be estimated using an independent validation set taking several factors into consideration such as the *a-priori* prevalence of AD in the cohort, and the risks associated with false positives and negatives of this classification.

9.4.2 Predicting pre-symptomatic AD in the RS cohort

We observed that a participant's age distinguished incident-AD well from CN. This is in line with earlier studies that identified age as an important predictor [318, 319]. However, we also observed that the predictive performance of age deteriorated when the reference group was less healthy, *i.e.*, when distinguishing incident-AD from a combined reference group also consisting of subjects with subjective or objective cognitive decline unrelated to AD. This is in line with the expectation that age is poor in distinguishing cognitive decline due to AD and cognitive decline due to other causes.

The predictive performance of the volumetric biomarkers from a single MRI scan depended on the *APOE* $\epsilon 4$ carrier status. We observed that hippocampus and entorhinal cortex were good predictors in *APOE* $\epsilon 4$ carriers. Interestingly, those biomarkers were estimated to be early in the corresponding temporal cascade. Similarly, total brain volume and ventricle volume were good predictors in *APOE* $\epsilon 4$ non-carriers which were also the early biomarkers in its temporal cascade. These results suggest that for predicting pre-symptomatic AD, early biomarkers play an important role and that it is important to understand the genotype-specific differences.

Lastly, we assessed the marginal utility of longitudinal MRI scans in identifying individuals at-risk of developing AD symptoms. We observed that participants with incident-AD showed a significant increase ($p < 0.005$) in disease stage over time as compared to CN and non-AD participants, in both *APOE* $\epsilon 4$ non-carriers and carriers. The rate of change of disease stage distinguished incident-AD subjects better than the disease stage at only the last scan, clearly highlighting the added value of longitudinal MRI scans, particularly in pre-symptomatic subjects. The rate of change of disease stage was also a better predictor of incident-AD than any other volumetric biomarker used in this study. This showed that the progression along the *APOE*-specific temporal cascade can be used to identify subjects in a population at-risk of developing AD.

In this study, we only used imaging biomarkers because cerebrospinal fluid biomarkers in a pre-clinical setting are usually not available. Recent breakthroughs in blood-based biomarkers [22] could help in obtaining fluid biomarkers in the pre-clinical phase of the disease. Previous work on DEBM [93] and co-init DEBM [312] had shown that the model is capable of incorporating biomarkers from multiple modalities for constructing the temporal cascades. We expect that our current approach of predicting pre-symptomatic AD in the general population would be applicable also in the presence of fluid biomarkers, should they become available in the future.

9.5 Conclusion and Future Work

We conclude that *APOE*-specific data-driven temporal cascades estimated by co-init DEBM are generalizable to population-based cohorts and that progression of individuals along such cascades is predictive of incident AD. Although the current study only considered volumetric biomarkers as inputs, it can be extended to fluid-based biomarkers, if these would become available in a population based study. Due to its robustness and explainability, we expect that our model can help identify at-risk individuals from the general population for targeted clinical trials as well as provide biomarker based objective assessment in such trials.

Chapter 10

Subtyping in sporadic Creutzfeldt-Jakob disease with discriminative event-based modeling

This chapter contains the content of the manuscript ‘Discriminative event-based model for subtyping in sporadic Creutzfeldt-Jakob disease. **Vikram Venkatraghavan***, Riccardo Pascuzzo*, Esther E. Bron, Marco Moscatelli, Marina Grisoli, Amy Pickens, Mark L. Cohen, Lawrence B. Schonberger, Pierluigi Gambetti, Brian S. Appleby, Stefan Klein, Alberto Bizzi. *In preparation.*’ An earlier version of this chapter was presented (oral) at the International Society for Magnetic Resonance in Medicine (ISMRM) conference 2020.

* Contributed equally in this work

Vikram Venkatraghavan’s contributions in this chapter are: Conceptualization, methodology, software, validation, and writing.

Abstract

Introduction: Sporadic Creutzfeldt-Jakob disease (sCJD) has multiple subtypes with different disease duration and spatio-temporal cascades of brain lesions. This study aims to diagnose *ante-mortem* the sCJD subtype using a patient-specific, single time-point estimate of the spatio-temporal cascade detected by diffusion-weighted magnetic resonance imaging (DWI).

Methods: We included 488 patients with autopsy-confirmed sCJD subtype diagnosis. For each subtype we used discriminative event-based model (DEBM) to infer the spatio-temporal cascade of DWI hyperintensities, using radiologist's ratings in 12 brain regions. Based on codon 129 prion protein genotype and DEBM cascades, we derived a novel classification algorithm to identify *ante-mortem* the sCJD subtype of each patient.

Results: The proposed algorithm achieved a balanced classification accuracy of 76.5%, with a low rater-dependency of $\pm 1\%$ for two other raters.

Discussion: We developed a novel data-driven approach for *ante-mortem* diagnosis of sCJD subtype, which could be beneficial for patient prognostication and stratification in targeted clinical trials.

Research in Context

Systematic review: The authors performed a literature review encompassing pre-prints and published articles. They found that in sporadic Creutzfeldt-Jakob disease (sCJD), patient stratification based on the different histopathological subtypes may be essential in evaluating the efficacy of therapeutic treatments.

Interpretation: The findings in this study highlight the spatio-temporal cascade of DWI abnormalities in seven molecular sCJD subtypes, confirming that each subtype results in a different cascade of lesion propagation. The proposed method provides an accurate, data-driven yet clinically explainable approach for *ante-mortem* sCJD subtype diagnosis. These features, together with the high inter-rater reliability, make the method suitable for implementation in clinical practice.

Future Directions: The proposed method can be used for stratifying sCJD patients by molecular subtype in targeted clinical trials. Future studies should also consider using other biomarkers (e.g., RT-QuIC) together with DWI, with the aim to increase the accuracy of subtype diagnosis using DEBM.

10.1 Background

Human prion diseases are rare and fatal neurodegenerative conditions caused by misfolding of the prion protein (PrP) into a pathogenic isoform (called PrP^D), which self-replicates, spreads, and accumulates in the brain [320, 321, 322]. Sporadic Creutzfeldt-Jakob disease (sCJD) accounts for about 85% of all human forms of prion diseases. A large phenotypic heterogeneity in clinical signs, disease evolution and duration has led to the identification of several molecular subtypes, mostly determined by two factors: the polymorphism (methionine (M) or valine (V)) at codon 129 of the PrP gene (*PRNP129*) and the type 1 or 2 of proteinase K-resistant PrP^D [323, 324]. The pairing of the *PRNP129* genotype (MM, MV, or VV) with the PrP^D type, forms the basis for a classification system that identifies six molecular pure sCJD subtypes: MM1, MM2, MV1, MV2, VV1, and VV2. MV2 is further subdivided into: MV2-cortical (MV2C), with predominant spongiform degeneration in the cortex, and MV2-kuru (MV2K), characterized by the presence of kuru-type PrP amyloid plaques [323, 324, 325, 326, 327]. Currently, the definite subtype diagnosis of sCJD is established only *post-mortem* by brain tissue examination.

The sCJD subtypes substantially differ in disease duration and clinical characteristics: patients with MM1 and MV1 subtypes present at onset with cognitive decline and myoclonus and show fast progression (median: 3-4 months), patients with VV2 present with initial cerebellar symptoms and intermediate disease duration (median: 6.5 months), whereas MM2, MV2C and MV2K have longer durations (median: between 17 and 21 months) [327, 328]. *Ante-mortem* identification of the sCJD subtype would hence be important for the prognosis and the clinical management of these patients. In addition, the possibility of discriminating subtypes can be helpful in the design of clinical trials. Current evidence indicates *PRNP129* genotype as an important factor in the evaluation of therapeutic treatments [329], and that any candidate therapeutic strategy for prion disease should be tested against multiple prion strains or subtypes [330].

In a recent study [331], we have proposed the first algorithms for the *ante-mortem* diagnosis of sCJD subtype in the form of decision trees using radiologist's ratings of diffusion-weighted imaging (DWI) hyperintensities, collected from a large cohort of patients with autopsy-confirmed sCJD diagnosis [15]. In another recent study involving the same cohort of patients, we used event-based models (EBM) [53, 61] to show that DWI abnormalities started as a focal condition and propagated to the other brain regions following a subtype-specific cascade [332]. We anticipated that such spatio-temporal cascades of DWI signal hyperintensity could enable a more accurate diagnosis of sCJD subtype, but a procedure for the *in vivo* identification of sCJD subtype informed by the cascade of DWI abnormalities had not been developed yet.

In this work, we introduce a novel classification algorithm for patients' sCJD subtype

diagnosis *ante-mortem*. It relies on using a discriminative event-based model (DEBM), a novel data-driven disease progression model [70, 93] to estimate the subtype-specific spatio-temporal cascades of DWI abnormalities from radiologist's ratings in a large cross-sectional dataset. For new patients to be classified, DEBM computes a single-time-point based approximation of the spatio-temporal cascade, and compares this patient-specific cascade to all subtype-specific cascades in order to identify the most probable patient's sCJD subtype. We assess the rater-dependency of the proposed method by estimating the classification accuracy for two other radiologists who were not involved in training the model. Finally, we benchmark the obtained subtype classification performance with that of the recently proposed decision tree classifier [331].

10.2 Methods

10.2.1 Sample and study procedures

Patients were selected from a prior study [15] involving a large cohort of subjects with suspected prion disease, recruited from January 2003 to April 2020 by the National Prion Disease Pathology Surveillance Center (NPDPSC) in Cleveland, Ohio as part of a Magnetic Resonance Imaging (MRI) consultation service program. This cohort has been the object of three recently published studies [15, 331, 332]. For the purposes of this study, we evaluated patients matching the following inclusion criteria: (i) pure subtype diagnosis of sCJD at autopsy; (ii) at least one positive brain DWI examination. Moreover, 50 subjects with clinically suspected prion disease then ruled out at autopsy, were randomly selected from the large cohort of the previous study [15] and included in this study. This research project was approved by the University Hospitals Cleveland Medical Center institutional review board. Informed consent was waived for the purpose of this study on all the subjects as they had deceased.

10.2.2 Biomarker collection and processing

The diagnosis of prion disease was established by histopathological examination of brain tissue, including PrP immunohistochemistry and Western blot analysis [321, 326]. Pure subtype of sCJD was determined by the detection of only one type (1 or 2) of proteinase K-resistant PrP^D from the examination of 3 brain regions and the presence of the corresponding neuropathological phenotype. Codon 129 polymorphism (MM, MV, or VV) of the PrP gene was determined to establish the definite diagnosis of histopathological sCJD subtype.

A senior neuroradiologist (15 years of experience) had prospectively scored all diffusion MRIs in electronic format, blind to the clinical data and preliminary diagnosis, and generated one lesion profile for each MRI study by evaluating the presence of DWI signal hyperintensities

in 12 brain regions on a four-point ordinal scale, as previously described [15]. Briefly, DWI signal hyperintensities were scored from zero (minimum, corresponding to absence of any sCJD-related lesion) to three (maximum, presence of extensive sCJD-related lesions). Five neocortical regions (frontal, parietal, including precuneus, temporal, and occipital lobes), three limbic structures (cingulate, insula, and hippocampus), striatum (caudate and putamen), thalamus and cerebellum were evaluated. For the purposes of this study, MRI examination was considered positive if at least one of these 12 regions was scored two or three.

To assess rater-dependency, two other radiologists (MG, senior radiologist with 15 years of experience, and MM, junior radiologist with 1 year of experience) scored a subgroup of 150 randomly selected patients and the 50 controls, blind to diagnosis and independently from the first radiologist and from each other.

10.2.3 Novel classifier for subtype identification

Spatio-temporal cascades of DWI abnormalities

EBM [53, 61, 332] is a type of data-driven disease progression model that can estimate the temporal cascade of biomarker abnormality events in a disease from a *cross-sectional* dataset. Such an estimation is feasible in a cohort consisting of subjects encompassing a wide spectrum of disease severity, because early biomarkers have a higher prevalence of abnormal biomarker values than the biomarkers that become abnormal later. In this work, we use DEBM introduced by Venkatraghavan *et al.* [70, 93] to estimate the temporal cascade of events in the cohort. DEBM estimates the temporal cascade of events in a two-step process. First, it estimates the degree of abnormality of each biomarker. In this study the regional DWI hyperintensities are treated as biomarkers, and we linearly map the radiologists' scores to probabilities of regional abnormality (0: 0, 1: 0.33, 2: 0.67, 3: 1). Second, DEBM estimates the spatio-temporal cascade of events for each subject, by ordering these probabilities. The mean spatio-temporal cascade for each subtype is estimated as the sequence that minimizes the sum of probabilistic Kendall's Tau distances to the spatio-temporal cascades of all subjects with the same sCJD subtype. Along with the mean cascades, the method also estimates a relative temporal distance between the biomarker abnormality events, resulting in a set of "event-centers (EC)" placed on a disease timeline normalized between 0 and 1 [93]. In order to evaluate if the estimated cascades in the seven subtypes were significantly different from one another, we used permutation testing [333] as detailed in the supplementary material 10A.1.

The linear mapping to transform radiologists' scores into probabilities of regional abnormality is different from the usual approach in event-based models, which is by mixture modeling [53, 61, 332]. In this work we skip the mixture modeling process as we observed a limitation in a recent article using the same dataset [332]. The EBM in that work [332] used a mixture of

a Bernoulli distribution (modelling radiologists' scores 0 and 1) and a uniform distribution (modelling radiologists' scores 0 to 3), which mapped the two highest scores (2 and 3) to a likelihood of 1, thereby losing granularity in the degree of abnormality encoded in the ratings.

Estimating the posterior probability for subtype classification

Once the spatio-temporal cascades of events are estimated for all subtypes, we use them to estimate the probability that a test subject with their *PRNP129* genotype (MM / MV / VV) measured *in vivo* belongs to a particular histopathological subtype. We denote the spatio-temporal cascade of subtype i as S_i , and the spatio-temporal cascade of the test subject j estimated using the radiologist's scores as s_j . The probability of subject j belonging to subtype i ($P_{i,j}$) is calculated by measuring the atypicality of the test subject's cascade when compared to the spatio-temporal cascade of subtype i , using the probabilistic Kendall's Tau distance between the two cascades, $d(s_j, S_i)$:

$$P_{i,j} = \frac{p_{i,j} \times e^{-d(s_j, S_i)}}{\sum_i p_{i,j} \times e^{-d(s_j, S_i)}} \quad (10.1)$$

where $p_{i,j}$ denotes the prevalence of the histopathological subtype i given the *PRNP129* genotype observed in subject j , in the sCJD population. For example, for a test subject with MV *PRNP129* genotype, $p_{i,j}$ for subtypes corresponding to MM and VV are 0, whereas corresponding to subtypes MV1, MV2C, and MV2K are based on their relative prevalence in the sCJD population (supplementary Table 10.ST1). The test subject j is assigned to the subtype with maximum $P_{i,j}$. For subtyping in the absence of *PRNP129* genotype in test subject j , MM1 and its phenocopy MV1 as well as MM2 and its phenocopy MV2C are considered in the same class. Thus $p_{i,j}$ for these merged subtypes are obtained by adding their respective prevalence as specified in supplementary Table 10.ST1.

10.2.4 Validation

We validated our subtype classification method using 10-fold cross-validation, both with and without using *PRNP129* genotype information. We estimated the spatio-temporal cascades of the subtypes in the training set and used the subjects in the test set for classification. We calculated the individual class sensitivity $Sens_i$ for each sCJD subtype as the proportion of correctly identified patients (i.e., true positive TP_i) over the total number N_i of patients with the subtype i :

$$Sens_i = \frac{TP_i}{N_i} \quad (10.2)$$

We computed balanced accuracy for each PRNP genotype ($BA_{MM}, BA_{MV}, BA_{VV}$) as well as for the overall subtype classification problem (BA_{TOT}) for benchmarking the performance of the classifier. Balanced accuracy is defined as the average class sensitivity of the subtypes under each PRNP genotype:

$$BA = \frac{\sum_i Sens_i}{\sum_i 1} \quad (10.3)$$

$$BA_{TOT} = \frac{\sum_i Sens_i}{7} \quad (10.4)$$

We also computed a weighted classification accuracy for each PRNP genotype ($WA_{MM}, WA_{MV}, WA_{VV}$) and for the overall subtype classification problem (WA_{TOT}) for measuring the expected classification accuracy in clinical use considering the prevalence of the subtypes in the sCJD population (Supplementary Table 10.ST1).

$$WA = \frac{\sum_i Sens_i \times p_i}{\sum_i p_i} \quad (10.5)$$

$$WA_{TOT} = \sum_i Sens_i \times p_i \quad (10.6)$$

We compared the balanced and weighted classification accuracies of the proposed method with *PRNP129* information with those of the recently proposed decision tree classifier for the *ante-mortem* sCJD subtype diagnosis [331]. We used the same train-test splits in the 10-folds cross-validation for both methods. The decision tree classifier was retrained and tested in each fold based on the method detailed previously [331]. Furthermore, we analyzed the significance of the difference between the accuracy of the decision tree classifier and that of the proposed classification approach, using the studentized bootstrap [334] approach as detailed in the supplementary material 10A.2.

Assessing rater dependency

To assess the dependency of the proposed method on raters, we estimated the accuracies when using the same rater (rater 1) for training and testing (intra-rater accuracy) and compared it with accuracies obtained when using different raters for testing (inter-rater accuracy). We compared the intra-rater and inter-rater classification accuracies in a repeated train-test split cross-validation experiment. In this experiment, the 150 subjects rated by all three radiologists were split randomly such that 80% of this subset was in the training set and 20% was in the

| Subtype | n (%) | Sex (M/F) | Age at MRI, years | Disease duration, months | Time from onset to MRI, days | Time from MRI to death, days |
|---------|------------|-----------|-------------------|--------------------------|------------------------------|------------------------------|
| MM1 | 216 (44.3) | 114/102 | 66 (59-73) | 2.6 (2.0-3.3) | 47 (30-67) | 28 (18-43) |
| MM2 | 43 (8.8) | 20/23 | 66 (58-73) | 12.7 (5.6-23.1) | 91 (42-235) | 242 (85-343) |
| MV1 | 40 (8.2) | 22/18 | 67 (57-70) | 4.1 (2.6-10.4) | 69 (42-112) | 39 (22-198) |
| MV2C | 37 (7.6) | 16/21 | 65 (61-69) | 16.9 (9.8-24.2) | 109 (54-285) | 345 (156-547) |
| MV2K | 36 (7.4) | 20/16 | 64 (59-69) | 12.2 (7.6-15.8) | 158 (93-317) | 112 (72-211) |
| VV1 | 25 (5.1) | 13/12 | 56 (41-69) | 9.7 (5.7-13.2) | 120 (57-167) | 168 (90-270) |
| VV2 | 91 (18.6) | 43/48 | 65 (59-71) | 5.0 (4.2-6.8) | 99 (61-133) | 61 (33-86) |
| Total | 488 (100) | 248/240 | 65 (59-72) | 4.2 (2.5-8.8) | 65 (37-126) | 46 (23-111) |

Table 10.1: Demographics of sCJD patients according to subtype. Median values (with inter-quartile range) are reported for time interval variables.

test set. The remaining subjects rated only by rater 1 were always used in the training set. Only the scores of rater 1 were used for training. We repeated this 30 times to estimate the mean and standard deviation of the balanced and weighted accuracies. We compared the obtained accuracies in the intra-rater setting with those of the inter-rater setting.

10.3 Results

10.3.1 Sample

A total of 488 sCJD subjects were included (median age at MRI, 65 years [interquartile range, 59-72]; 248 [51%] male) with positive MRI and the following pure seven subtypes: MM1 (n=216), MM2 (n=43), MV1 (n=40), MV2C (n=37), MV2K (n=36), VV1 (n=25), VV2 (n=91). Demographic data of these patients are reported in Table 10.1. For the rater dependency analysis, two other radiologists examined the MRIs of a subgroup of 150 sCJD patients randomly selected, whose data are reported in supplementary Table 10.ST2.

10.3.2 Spatio-temporal propagation of DWI abnormalities is subtype specific

The proposed approach estimated the cascades for each of the seven subtypes. The results of permutation testing for all pairwise comparisons of the subtypes' cascades are shown in supplementary Table 10.ST3. The obtained cascades for the subtypes are reported briefly in this section and have been elaborated further in supplementary material 10B.1.

The two most frequent sCJD subtypes (MM1 and VV2) had opposite cascades ($p < 0.0001$) of lesion propagation (Figure 10.1). In MM1 subjects, the first region to become affected (henceforth referred to as "epicenter") was the parietal cortical ribbon, with $EC = 0.25$, which slightly preceded the involvement of the precuneus ($EC = 0.30$). The caudate was involved

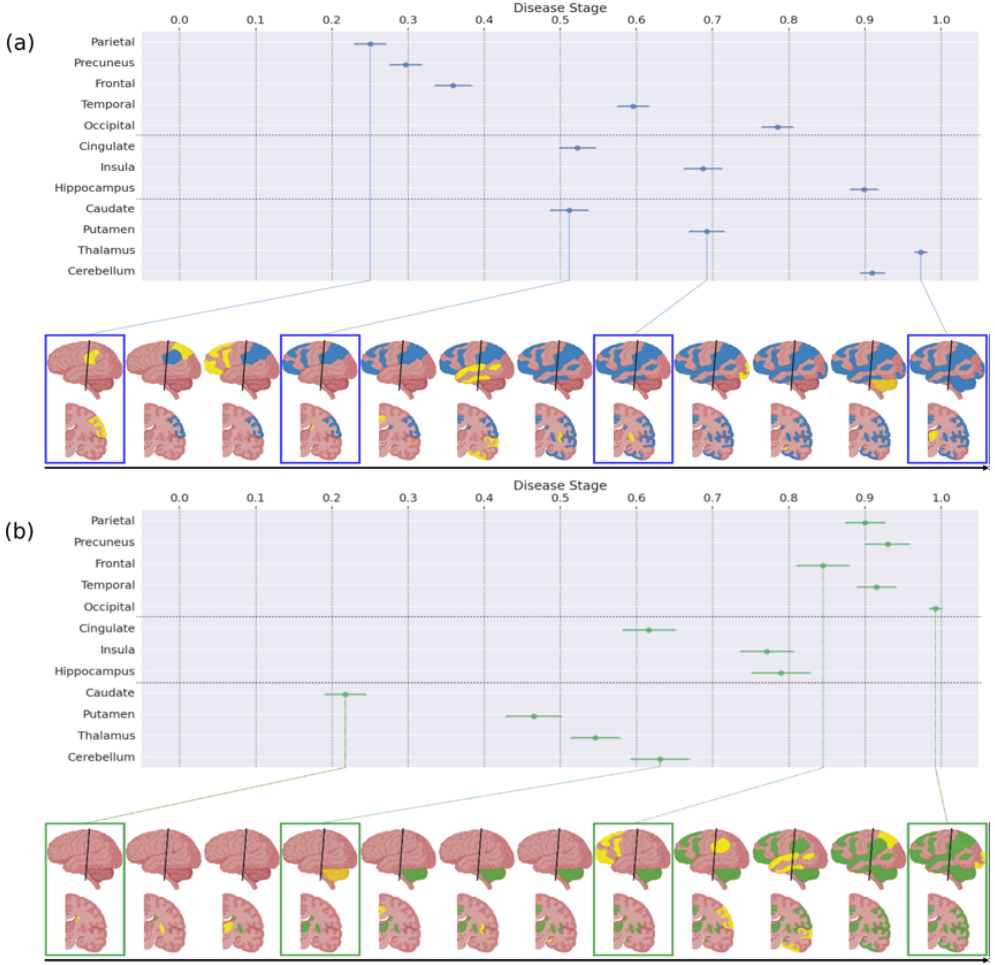


Figure 10.1: Disease progression timeline of MM1 (a) and VV2 (b) subtypes. The plot on top of each subfigure shows the event-centers of the different regions and their respective standard deviation estimated from a batch of 100 independent bootstrap samples. The visualization at the bottom of each subfigure highlights the region that becomes abnormal at the corresponding disease stage in yellow, with all the previously affected region in blue.

quite early in the sequence ($EC = 0.51$). The hippocampus, cerebellum and thalamus were affected later ($EC \geq 0.90$)

In VV2 subjects, we first observed early involvement of the caudate ($EC = 0.21$), followed by putamen ($EC = 0.46$) and thalamus ($EC = 0.55$). Next, cingulate and cerebellum were affected and preceded the other limbic structures. The frontal cortex became abnormal only at a late stage ($EC = 0.85$), preceding the involvement of the cortical ribbon of the other lobes.

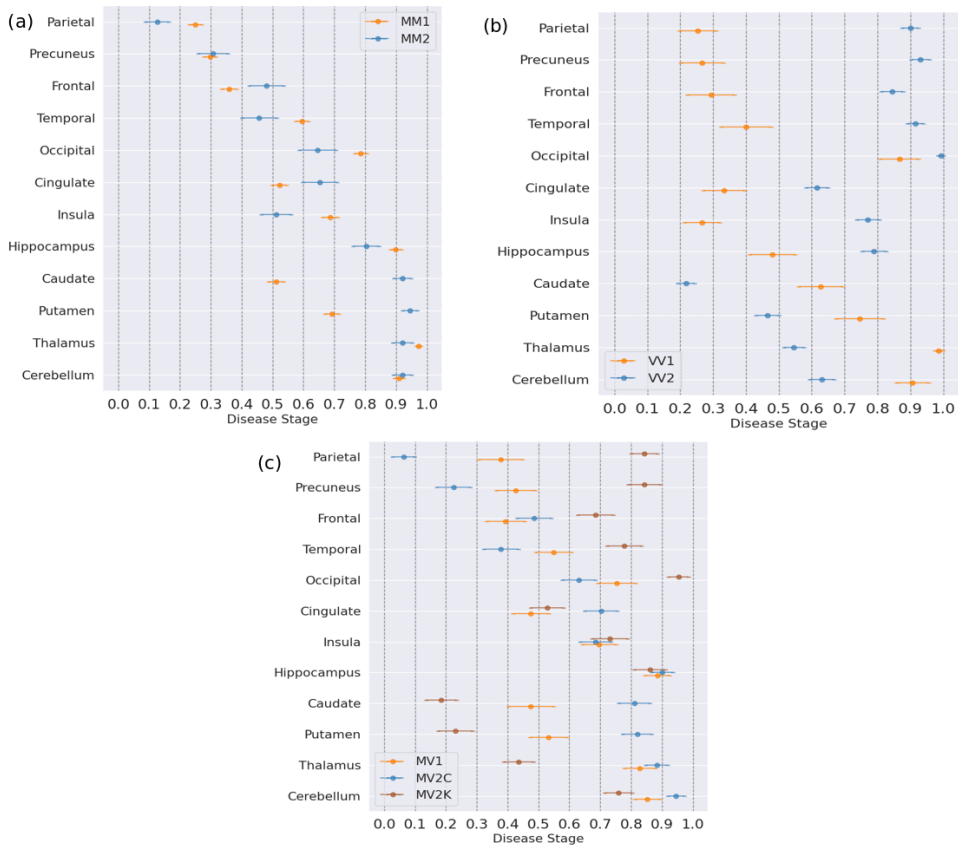


Figure 10.2: Event-center comparisons for the three *PRNP129* genotypes MM (a), VV (b) and MV (c). The standard deviation of the event-centers were estimated from a batch of 100 independent bootstrap samples.

The MM2 cascade was significantly different ($p = 0.01$) from that of MM1, despite sharing the same epicenter in the parietal cortex (Figure 10.2a). The main differences were the striatum (especially the caudate) affected only very late in MM2 ($EC = 0.92$) and the abnormalities in all cortical regions preceding those in the subcortical regions.

The cascades of the two VV subtypes were almost opposite (Figure 10.2b) ($p < 0.0001$). VV1 cascade resembled that of MM1 ($p = 0.26$) with the same epicenter (the parietal cortex, $EC = 0.25$), but with an earlier involvement of the limbic structures (insula: $EC = 0.27$ and cingulate: $EC = 0.33$). Notably, VV1 had the earliest involvement of the hippocampus among all subtypes.

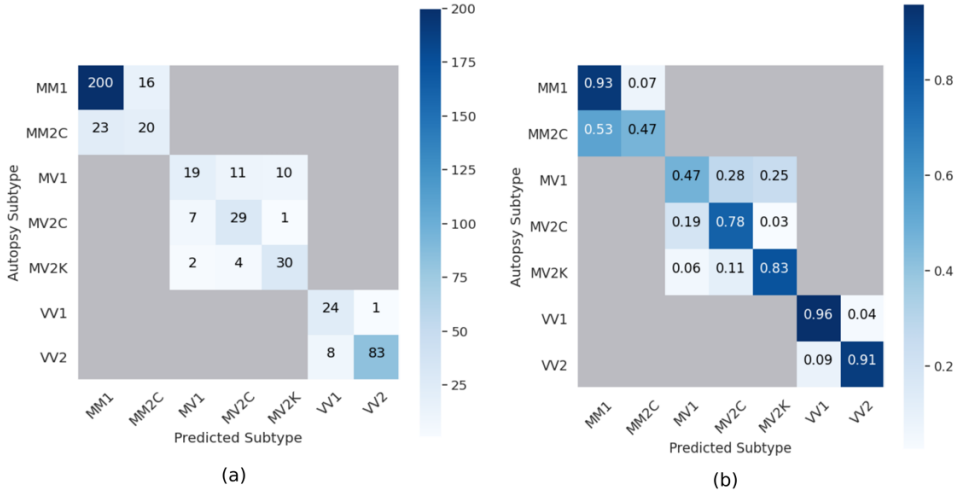


Figure 10.3: Confusion matrix (a) and normalized confusion matrix (b) obtained using the proposed method with 10-fold cross-validation. The diagonal elements in (a) represent the number of correctly classified subjects and the off-diagonal elements represents the incorrectly classified subjects. The diagonal elements in (b) represent the fraction of correctly classified subjects in a specific subtype, while the off-diagonal elements represent the fraction of incorrectly classified subjects. The overall balanced classification accuracy obtained was 76.5% and the weighted classification accuracy obtained was 87.0%.

Considering the MV subtypes, cascades of lesion propagation were opposite for MV2C and MV2K ($p < 0.0001$), while the MV1 cascade was more similar to MV2C ($p = 0.10$) than MV2K ($p < 0.0001$) (Figure 10.2c). Specifically, MV2C had its epicenter located in the parietal cortex and followed the same cascade identified for MM2 ($p = 0.80$), the phenocopy of the MV2C subtype. The cascade of MV2K was reversed, with the epicenter located in the striatum, and traced that of VV2 ($p = 0.32$), but with a later involvement of the cerebellum. Finally, the MV1 cascade largely overlapped with that of MM1 ($p = 0.49$).

10.3.3 Subtype Classification

The results of the proposed classification procedure with *PRNP129* information are shown in Figure 10.3 ($BA_{TOT} = 76.5\%$, $WA_{TOT} = 87\%$) and without *PRNP129* information are shown in supplementary Figure 10.SF1 ($BA_{TOT} = 48.9\%$, $WA_{TOT} = 73.2\%$).

The proposed classification procedure with *PRNP129* information achieved individual sensitivities above 90% for the two most common subtypes MM1 (92.6%) and VV2 (91.2%) (Figure 10.3). It also achieved a sensitivity of 96% for the rare VV1 subtype, with only 1 of 25 VV1 patients incorrectly classified as VV2. However, about half of the times (53.5%) a true MM2 subject was misclassified as MM1, while the reverse occurred quite rarely (7.4%). In the heterozygote MV group, although good performances were obtained for MV2C and

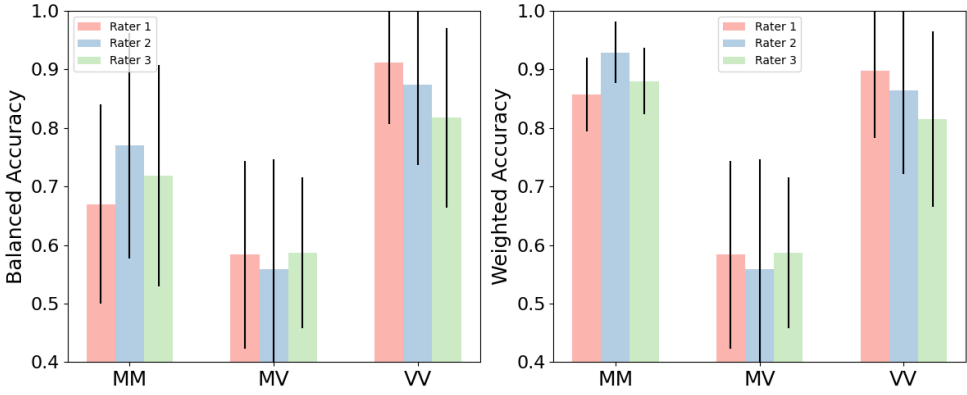


Figure 10.4: Inter-rater analysis comparing (a) the balanced accuracies and (b) the weighted accuracies of the classification for 3 different radiologists. Rater 1's ratings were used for training the model, while the ratings of the other two raters were only used in the test-set. The height of the bar represents the mean weighted accuracy in a specific genotype and the segment represents the standard deviation of the weighted accuracy in 30 iterations of randomized train-test split.

MV2K subtypes (sensitivities of 78.4% and 83.3%, respectively), MV1 was difficult to identify (sensitivity of 47.5%) and it overlapped with the other two MV subtypes. Notably, the misclassifications between MV2C and MV2K were rare (only 1 and 4 cases, respectively).

The comparative classification results of our proposed method and of the decision tree algorithm, while using *PRNP129* information are shown in Table 10.2. The proposed method obtained a significantly higher BA_{TOT} than that of the decision tree (76.5% vs. 67.1%, $p < 0.001$), while the WA_{TOT} was similar (87.0% vs. 85.0%, $p = 0.115$). The accuracies were identical for the MM genotype ($WA_{MM} = 89.8\%$ and $BA_{MM} = 69.6\%$ for both methods). The proposed method had considerably higher BA_{VV} than the decision tree (93.6% vs. 80.7%) but similar WA_{VV} (91.4% for the proposed method and 92.1% for the decision tree classifier). For the MV genotype, the proposed method achieved higher accuracies than the decision tree ($BA_{MV} : 69.7\%$ vs. 56.5%, $WA_{MV} : 70.6\%$ vs. 57.3%). Altogether, these results suggest that the two methods performed similarly for the most common subtypes (MM1 and VV2) and for MM2, whereas the DEBM classifier was more sensitive in identifying the other less frequent subtypes.

In 150 sCJD patients considered for the rater-dependency analysis, the 3 radiologists performed very similarly in terms of BA_{TOT} (rater 1: 69% \pm 11%; rater 2: 70% \pm 10%, rater 3: 69% \pm 9%) as well as in terms of WA_{TOT} (rater 1: 82% \pm 5%; rater 2: 86% \pm 5%, rater 3: 82% \pm 5%). The balanced and weighted accuracies for each genotypic group are shown in Figure 10.4.

| Metric | MM | | MV | | VV | | Total | | |
|-------------------|------|---------------|------|---------------|------|---------------|-------|---------------|---------|
| | DEBM | Decision Tree | DEBM | Decision Tree | DEBM | Decision Tree | DEBM | Decision Tree | p-value |
| Balanced Accuracy | 69.6 | 69.6 | 69.7 | 56.5 | 93.6 | 80.7 | 76.5 | 67.1 | < 0.001 |
| Weighted Accuracy | 89.8 | 89.8 | 70.6 | 57.3 | 91.5 | 92.1 | 87.0 | 85.0 | 0.115 |

Table 10.2: Comparison of the accuracies between the proposed method and a decision tree classifier according to *PRNP129* genotype and overall.

10.4 Discussion

In this study, we developed a classification system for determining *ante-mortem* the sCJD subtype in individual patients, using a cross-sectional estimate of the spatio-temporal cascades of propagation of DWI abnormalities in each subtype, in combination with the *PRNP129* genotype of the patients. We achieved a balanced classification accuracy of 76.5% among all subtypes, and 87.0% accuracy if we weigh each subtype based on its prevalence in the sCJD population. In the absence of *PRNP129* genotype in test subjects, the performance of the proposed algorithm dropped substantially to a weighted accuracy of 73.2%.

Unlike widely used machine learning approaches such as support vector machines or deep learning, the results of the proposed method are clinically explainable since they are based on the differences from the subtype-specific cascades of DWI abnormalities. This is crucial for maintaining accountability, and eventual clinical usability [335, 336]. For its implementation in clinical practice, the classification system needs to be usable by raters other than the one used to train the model. Our rater-dependency analysis showed that the accuracy of the proposed method is not sensitive to the rater. The low rater-dependency together with the explainability of the developed method make it suitable for clinical practice.

Another study recently showed similar propagation cascades in the same population cohort examined by the same neuroradiologist, but using a different EBM approach [332]. In this work, we identified and addressed a drawback in the algorithmic choice of the previously proposed approach. We identified that the two highest scores (scores of 2 and 3) were both mapped to a likelihood of 1 in that approach, thereby losing granularity in the degree of abnormality encoded in the ratings. Although the results of the two approaches turned out to be quite similar, some differences were found, as highlighted in the supplementary material 10B.2. Moreover, the ability to compute mean spatio-temporal cascades in the subtypes as well as approximate estimates of the cascade in each individual is a unique feature of DEBM, which facilitated the development of this novel classifier.

We compared the accuracy of the proposed method with that of the previously proposed

decision tree algorithm [331] and found that DEBM had a significantly higher balanced classification accuracy. Particularly, the proposed approach had higher sensitivity in identifying the rarer subtypes MV1, MV2C, MV2K, and VV1, as compared with the decision tree classifier.

One limitation of this study was that the entire procedure is based on the scores assigned to the DWI abnormalities in 12 brain regions by the neuroradiologists. Although DWI has emerged as a very sensitive diagnostic modality and probably the test of choice for the diagnosis of sCJD at symptoms onset [15, 337, 338, 339, 340, 341], the radiologist's expertise remains a major factor influencing the correct identification. Future work should focus on developing automated methods for estimating the severity of DWI signal abnormalities in the whole brain. Regarding other clinical implications of the proposed method, the involvement of one of the disease epicenters in the MRI of a patient can aid the radiologist to raise the suspect of prion disease at an earlier stage. Furthermore, knowledge of the spatio-temporal cascades of DWI abnormalities can help to corroborate the initial findings by monitoring follow-up MRI examinations, and to provide better prognosis for the patient. The subtype diagnosis of sCJD patients is also essential for optimizing future clinical trial designs, because it allows to identify more homogeneous groups of patients at the beginning of the disease, increasing the time window for therapeutic intervention and, potentially, influencing the selection of the most appropriate treatment [45].

In future research, the use of other laboratory tests such as Real-Time Quaking-Induced Conversion (RT-QuIC), cerebrospinal Tau and 14-3-3 proteins, together with DWI, could also be considered to increase the accuracy of sCJD subtype classification [15, 342].

In conclusion, we proposed a novel and explainable classification method for identifying the sCJD subtype in the individual patient based on the spatio-temporal cascade of abnormalities in diffusion MRI. We achieved state-of-the-art performance in identifying sCJD subtype *ante-mortem*, and the proposed method could eventually aid disease prognosis and block allocation in clinical trials.

Supplementary Material

10A Methods

10A.1 Significance testing of differences in the subtype cascades

We used permutation testing [333] to test the significance of the difference in cascades obtained for the different subtypes under the null hypothesis that the subtypes had the same cascades. We estimated the distribution of the Kendall's Tau distance by generating 10,000 random permutations of the patients' subtypes. We then empirically computed the one-sided p-values for the actual Kendall's Tau distances between each pair of orderings of the seven subtypes, calculated as the proportion of sampled permutations where the distance was greater than or equal to the actual distance.

10A.2 Significance testing of differences in the classification accuracies

To test whether the classification accuracies of the two methods were significantly different, we used the studentized bootstrap approach [334]. We performed 1000 iterations of bootstrap resampling of the predictions of the two methods in the entire cohort, in a stratified manner such that the relative proportions of the different subtypes were preserved after resampling. Based on these bootstrap samples, the standard error for the difference in accuracies between two methods was computed, which was converted to the p-value for the null hypothesis of no difference.

10B Results

10B.1 Cascade of events in the subtypes

This section further elaborates the cascades obtained for the seven subtypes as shown in Figures 10.1 and 10.2.

In MM1, the parietal cortical ribbon was the first region (epicenter) to become affected and slightly preceded the involvement of the precuneus. It was closely followed by abnormality in the frontal cortical ribbon, followed by caudate and cingulate gyrus. Surprisingly, the putamen was not affected right after the caudate, despite both being part of the same structure (i.e. the striatum), but was preceded by the involvement of the temporal and insular cortices. The last regions to become affected were hippocampus, cerebellum and thalamus.

In VV2, as opposite to MM1, the epicenter was located in the caudate, followed by putamen and thalamus. Next, cerebellum and cingulate were affected and preceded the other limbic structures. The cortical ribbon of the frontal lobe became abnormal only at a late stage and preceded that of the other lobes.

MM2 and MM1 shared the same epicenter in the parietal cortex, but then followed different

orderings of propagation. The main difference was the striatum (especially the caudate) affected only very late in MM2. In addition, the temporal cortex became abnormal slightly before the frontal cortex in MM2. Likewise, the event-centres of insula and occipital cortex in MM2 became abnormal earlier than their MM1 counterparts, with the result that abnormalities in all cortical regions preceded the subcortical regions.

VV1 cascade was opposite to that of VV2: it had the epicenter located in the parietal cortex and an earlier involvement of the limbic structures (cingulate insula, and hippocampus). The last regions to become affected were the cerebellum and the thalamus, which was always spared in patients with VV1 subtype.

MV2C cascade was similar to that of MM2, its phenocopy: it had the epicenter located in the parietal cortex, followed by propagation to the cortical ribbon of the temporal and frontal lobes. Eventually, the cortical ribbon of the other lobes was affected and preceded the involvement of the striatum, thalamus and cerebellum, as in MM2. The cascade of MV2K was reversed, with the epicenter located in the striatum, followed by thalamus and cingulate; the cortical ribbon of the frontal lobe was involved later and before all the other lobes, as in VV2. Finally, MV1 cascade was similar to that of MM1: it had the epicenter in the parietal and frontal cortices with an early involvement of the striatum preceding the other cortical regions, while thalamus, cerebellum and hippocampus were affected only at the end.

10B.2 Notable differences in the cascade of events estimated using DEBM and EBM

This section further elaborates on the differences between the cascades of the seven subtypes obtained using DEBM and EBM as noted briefly in the Discussions section.

Our novel approach estimated that the parietal cortex distinctly preceded the precuneus in MM1, as well as in MM2 and MV2C, whereas the reverse was found with EBM16. In addition, our novel approach indicated more clearly the caudate as the epicenter of the disease for VV2, whereas the cerebellum (overlapping with other subcortical regions) was identified as the most probable epicenter according to EBM. It is however hard to conclude which method is more accurate. Future work could focus on performing further experiments on longitudinal data to unambiguously arrive at a conclusion.

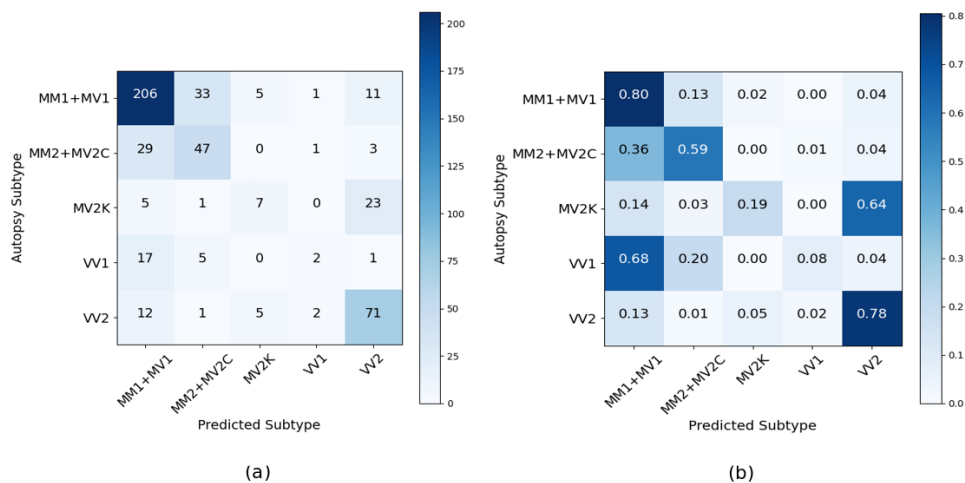


Figure 10.SF1: Confusion matrix (a) and normalized confusion matrix (b) obtained using the proposed method when not using *PRNP129* genotype information, with 10-fold cross-validation. MM1 and its phenocopy MV1, as well as MM2 and its phenocopy MV2 were grouped together for the purposes of this classification. The diagonal elements in (a) represent the number of correctly classified subjects and the off-diagonal elements represents the incorrectly classified subjects. The diagonal elements in (b) represent the fraction of correctly classified subjects in a specific subtype, while the off-diagonal elements represent the fraction of incorrectly classified subjects. The overall balanced classification accuracy obtained was 48.9% and the weighted classification accuracy obtained was 73.2%.

| Subtype | Prevalence (%) |
|---------|----------------|
| MM1 | 61 |
| MM2 | 4 |
| MV1 | 5 |
| MV2C | 5 |
| MV2K | 6 |
| VV1 | 1 |
| VV2 | 18 |

Table 10.ST1: Prevalence of subtypes in the sCJD population as estimated from the literature. Prevalence is retrieved from [327, 328]

| Subtype | n (%) | Sex (M/F) | Age at MRI, years | Disease duration, months | Time from onset to MRI, days | Time from MRI to death, days |
|---------|-----------|-----------|-------------------|--------------------------|------------------------------|------------------------------|
| MM1 | 77 (51.3) | 42/35 | 66 (60-71) | 2.8 (2.1-3.4) | 52 (31-77) | 30 (20-46) |
| MM2 | 12 (8.0) | 5/7 | 68 (54-76) | 11.7 (5.6-17.6) | 104 (60-153) | 245 (69-412) |
| MV1 | 10 (6.7) | 7/3 | 69 (68-71) | 9.0 (2.6-17.0) | 74 (31-122) | 78 (44-276) |
| MV2C | 10 (6.7) | 5/5 | 66 (64-76) | 17.8 (14.4-29.5) | 129 (76-295) | 378 (200-588) |
| MV2K | 8 (5.3) | 6/2 | 67 (59-71) | 10.0 (7.1-12.1) | 79 (23-158) | 178 (84-258) |
| VV1 | 6 (4.0) | 5/1 | 55 (45-66) | 6.1 (4.9-8.8) | 83 (46-148) | 113 (91-157) |
| VV2 | 27 (18.0) | 12/15 | 64 (57-68) | 5.1 (4.2-7.1) | 99 (82-133) | 58 (28-75) |
| Total | 150 (100) | 82/68 | 66 (59-71) | 4.0 (2.7-8.2) | 67 (39-117) | 46 (24-89) |

Table 10.ST2: Demographics of sCJD patients selected for inter-rater analysis, according to subtype. Median values (with inter-quartile range) are reported for time interval variables.

| sCJD subtypes | MM1 | MM2 | MV1 | MV2C | MV2K | VV1 | VV2 |
|---------------|-----|--------------|-------|-------------|--------------|--------------|--------------|
| MM1 | | 0.01* | 0.49 | 0.06 | $< 10^{-4*}$ | 0.26 | $< 10^{-4*}$ |
| MM2 | | | 0.01* | 0.80 | $< 10^{-4*}$ | 0.21 | $< 10^{-4*}$ |
| MV1 | | | | 0.10 | $< 10^{-4*}$ | 0.07 | $< 10^{-4*}$ |
| MV2C | | | | | $< 10^{-4*}$ | 0.11 | $< 10^{-4*}$ |
| MV2K | | | | | | $< 10^{-4*}$ | 0.32 |
| VV1 | | | | | | | $< 10^{-4*}$ |
| VV2 | | | | | | | |

Table 10.ST3: Empirical p-values testing the significance of the differences in orderings of the seven sCJD subtypes from one another, computed using permutation testing.* denotes that the orderings are significantly different ($p < 0.05$). The values in bold represents the comparisons within the same PRNP genotype.

General Discussion & Summary

Chapter 11

General discussion

The cacophony of neurodegeneration is orchestrated by the underlying symphony of pathological processes of the disease. These pathological processes, be it in a fast progressing one like Creutzfeldt-Jakob disease or in a slowly progressing one like Alzheimer's disease, are devastating to the patients and their caregivers in their own way. Understanding the underlying order in the heterogeneous progression of these diseases is crucial for early diagnosis, effective patient care, as well as in identifying novel disease modifying treatment strategies. Disease progression modeling is a class of data-driven approaches that has the potential to aid in these quests. It helps in understanding the pathophysiology of neurodegenerative diseases, in objectively assessing the progression of these diseases, and could potentially help in identifying novel therapeutic targets for such diseases.

In this thesis I focussed on developing novel disease progression models for estimating disease progression timelines of neurodegenerative diseases. This research resulted in several methodological innovations and provided novel insights into the disease processes of several neurodegenerative diseases. In this chapter, in Section 11.1, I discuss the methodological innovations presented in the thesis. This is followed, in Section 11.2, by a discussion on the novel insights into neurodegenerative diseases provided in this thesis. Implications of the findings in this thesis for clinical practice as well as drug trials and other interventions are discussed in Sections 11.3 and 11.4 respectively. Lastly, the roadmap for future methodological innovations in disease progression modeling is discussed in Section 11.5.

11.1 Methodological innovations

Event-based models (EBM) are disease progression models that estimate a temporal sequence of abnormal biomarker events in neurodegeneration based on cross-sectional data. While the concept of estimating a temporal sequence from cross-sectional data intrigued me when I started this thesis in 2016, the existing EBM models available at the time were either too restrictive in assuming that the subjects in the cohort followed a common temporal sequence or did not scale well to more than a few biomarkers (as shown in Chapter 2). Another

limitation in the existing EBM was a result of strict quantization of position in the temporal sequence of events, which could be non-intuitive in terms of inferring actual progression of the disease (as also discussed in Chapter 2).

To address these limitations, I developed a novel framework of EBM called the discriminative EBM (DEBM) (in Chapter 2). In this framework, first a noisy estimate of the subject-specific temporal sequence of biomarker abnormalities is estimated independently for each subject in the cohort, thereby accounting for heterogeneity in the cohort. This is followed by estimating the mean disease progression timeline of the subjects in the cohort. Using this new framework of EBM, one can not only estimate the temporal sequence of abnormality events but also the event centers, referring to the relative temporal distance of these events from one-another. Furthermore, using the estimated disease progression timeline, one can also estimate an objective measure of disease severity in a patient based on the obtained biomarkers.

Apart from these novel steps which are specific to the DEBM framework, a novel mixture modeling approach was developed in this work, which could be an integral part of all the existing EBM frameworks. This mixture modeling approach improved the stability of the obtained model parameters for EBM analysis. The validity and robustness of the different aspects of the developed framework were tested extensively on synthetic datasets simulating the progression of neurodegeneration as well as on a large AD cohort. Furthermore, the generalizability of the disease progression timeline of AD constructed in the large AD cohort to other clinical cohorts was verified in Chapter 3.

These extensive experimentations also led to some novel insights about the developed method and shed light on its limitations. Limitations were related to: i) The stability of the mixture modeling, ii) The importance of the choice of biomarkers in the accuracy of EBMs, iii) The potential of DEBM for personalized prediction. These novel insights led to further development of DEBM in the subsequent chapters, as detailed in the following paragraphs.

Stability of the mixture modeling: The different types of mixture modeling approaches analyzed in Chapter 2 led to the conclusion that the stability of the mixture models is crucial for building an accurate EBM. Hence, while studying the lateral asymmetry in brain in Chapter 4, I developed a novel mixture modeling technique called siamese mixture modeling to further improve the stability of the mixture model. While analyzing the effect of *APOE* on the disease progression timeline of AD in Chapter 8, the mixture models in each of the *APOE* based groups were initialized using data pooled from all the groups. This co-initializing of the mixture models led to a substantial increase in the accuracy of the subsequently estimated temporal sequence in the groups. This novel mixture-modeling approach was further validated in Chapter 9, where it was used to train the DEBM model in a research-based cohort, and validated in a population-based cohort, to investigate its generalizability.

Choice of biomarkers: The simulation experiments in Chapter 2 showed that the accuracy of estimating the disease progression timeline depends on the overlap between the normal and abnormal biomarker distributions used in the model. It implied that when using imaging biomarkers such as volume of a region, the high-dimensional information of the brain regions available in the images are being reduced to scalar biomarkers potentially increasing the overlap between the normal and abnormal distributions. Hence, to address this suboptimality, I developed a novel approach called n-dimensional DEBM (nDEBM) to estimate the disease progression timeline using high-dimensional imaging biomarkers in Chapter 7. In nDEBM semi-supervised support vector machines were used, replacing mixture modeling in the DEBM framework. This was done to address a well-documented scalability problem of mixture modeling [343], known as the curse of dimensionality.

Personalized prediction: Further methodological developments include using the output of DEBM for: i) prediction of clinical status, ventricular volumes, and AD assessment scale (ADAS) cognitive scores, in subjects of a large AD cohort in Chapters 6, ii) *ante-mortem* prediction of molecular subtypes of sporadic Creutzfeldt-Jakob disease (sCJD) in Chapter 10, both further detailed below.

In Chapter 6, DEBM was used in the TADPOLE challenge to assess the stage of the disease in patients and use the rate of change of disease stages to predict the future biomarker values of the subjects as well as to predict their symptom onset precisely. One of the commonly used approaches for modeling and predicting longitudinal biomarker trajectories in AD is to consider biomarker changes in AD as an accelerated form of biomarker changes in healthy aging [51, 52, 58]. In the novel method developed for this challenge, I considered aging and disease progression as independent factors affecting the progression of biomarker values along with interaction between the two. This novelty helped me develop a highly accurate model and the method was ranked first in predicting the future ventricular volumes in subjects, among the various teams that participated in the challenge. In Chapter 9, the rate of change of disease stage was also shown to identify subjects with preclinical or prodromal AD in a population-based cohort.

In Chapter 10, I developed a novel classifier for *ante-mortem* identification of the molecular subtypes of subjects with sCJD, using the distances of patient-specific spatio-temporal sequence of abnormalities from several subtype-specific spatio-temporal sequences. Through this method it was shown that the patient-specific sequence of abnormalities obtained from DEBM could also be clinically useful for patient prognostication.

11.2 Novel insights into neurodegenerative diseases

In collaboration with clinical researchers, I used DEBM to obtain novel insights into the neurodegenerative disease processes in AD, frontotemporal dementia (FTD), CJD, and multiple sclerosis (MS). I will discuss the major findings in the following paragraphs.

The first key insight addresses the effect of the *APOE* gene on the progression timeline of AD. In Chapter 8, cerebrospinal fluid (CSF) biomarker Amyloid β_{42} was identified as the earliest biomarker to become abnormal among *APOE* $\epsilon 4$ carriers and homozygous $\epsilon 3$ carriers, which is consistent with the current understanding of AD. However, it was not an early biomarker among $\epsilon 2$ carriers. Furthermore, we also showed that *APOE*- $\epsilon 2$ carriers, homozygous *APOE*- $\epsilon 3$ carriers, and *APOE*- $\epsilon 4$ carriers have significantly different disease progression timelines in AD.

In Chapter 8, CSF neurofilament light chain (NfL) was consistently observed to be a late biomarker of AD for all the *APOE* based genotypes. This was in contrast to our observation in *GRN* related FTD (in Chapter 4) where serum NfL was one of the earliest biomarker abnormality events. This insight could be useful for differential diagnosis of the two types of dementias in the future. The analyses in Chapter 4 also showed that white matter tracts were also early biomarkers, along with the cognitive biomarker of language and biomarkers from the language networks of the brain. These results suggested that, for *GRN* related FTD axonal degeneration and damage to the language network are two of the earliest disease events. In this chapter, we also observed the presence of intra-phenotype heterogeneity in the behavioural variant *GRN* related FTD whereas the non-fluent variant primary progressive aphasia phenotype was quite homogeneous.

The timeline of biomarker abnormality events in multiple sclerosis (MS) was studied in Chapter 5, using structural, functional, and cognitive biomarkers. Gray matter volumes of thalamus, insula, hippocampus, and cerebellum were observed to be the earliest biomarkers in MS, whereas white matter microstructure related biomarkers were predominantly late biomarkers. These insights could help patient monitoring of MS in the future.

In Chapter 10, we studied the progression timelines of the different molecular subtypes of sCJD where the two most predominant subtypes (MM1 and VV2) were observed to have opposite lesion propagation timelines. In MM1, parietal and frontal cortices were observed to have lesions early in the disease timeline, while thalamus was observed to have lesions only towards the late stages of the disease. VV2, in contrast, had lesions in parietal and frontal cortices very late in the disease. Caudate was the earliest region to have lesions in VV2, with thalamus also being one of the early regions. These insights could be used for block allocation in future clinical trials of sCJD.

11.3 Implications for clinical practice

Clinical diagnosis of neurodegenerative disease requires interpretation of cognitive tests, fluid biomarkers, and imaging biomarkers. Additionally, it could also require inputs from genetics for the presence of any genetic risk factors. Diagnosing a patient based on such multi-disciplinary inputs is quite complex and often subjective [344]. Data-driven approaches such as DEBM and its variants presented in this thesis can elegantly combine multi-disciplinary inputs in order to estimate a patient stage, and could act as an objective summary score for evaluating a disease.

This thesis showed that patient stages quantified by positioning test subjects along the pre-constructed progression timeline have diagnostic value in AD, *GRN*-related FTD, and MS. The thesis also showed that the patient stages have prognostic value in AD. Their prognostic value was not assessed for *GRN*-related FTD and MS in this thesis; this could be explored in the future as more longitudinal datasets become available. In the experimentations on the sCJD dataset, although the identification of molecular subtype in itself had prognostic value, time of survival within a subtype did not correlate with the patient stage obtained for the subtype. My interpretation of this negative result is that a key factor determining the prognostic value of the patient stage obtained from DEBM is the variability in the rate of progression of the disease across patients. When the variability in rate of progression is substantial compared to the entire duration of the disease, the disease stage loses its prognostic value.

Apart from the diagnostic and prognostic value of the patient stages, there are a couple of other important factors that influence the usefulness of data-driven methods in clinical practice. These are: i) Generalizability of the data-driven methods trained in one cohort to other cohorts [309]. ii) Explainability of the decisions obtained using data-driven approaches in a way that can safeguard the accountability of the clinicians [335, 336].

Generalizability: The generalizability of the pre-constructed AD progression timeline to other clinical cohorts was verified in Chapter 3. Their utility in predicting the progression of pre-symptomatic AD in a population-based setting was verified in Chapter 9.

Explainability: The patient stages obtained using DEBM are explainable in terms of how the observed biomarker values were used to assess disease severity of a patient, which is unlike most of the widely used black-box approaches to machine learning. Furthermore, the results of the DEBM based subtype classification method developed in Chapter 10 are also clinically explainable since they are based on the observed differences between the patient-specific and subtype-specific sequences of abnormalities.

The proven diagnostic and prognostic utility of the patient stages, the generalizability of the method, together with the transparency it brings to the decision making process, make DEBM

a strong candidate for aiding in clinical diagnosis of neurodegenerative diseases.

11.4 Implications for drug trials and other interventions

Patients with different subtypes of a neurodegenerative disease are likely to show different characteristics, in terms of the cognitive domains that are affected. Identifying these subtypes can hence be very helpful in patient management in terms of designing targeted supportive therapies or prescribing symptomatic drugs. In Chapter 10, I showed that DEBM can be used to identify subtypes. The developed method can be extended in the future for other neurodegenerative diseases as well.

Subtypes of a neurodegenerative disease are a manifestation of the latent differences in progression pathways of the disease. Identifying these subtypes based on the progression timeline of the diseases could help in identifying candidates with similar disease progression pathways. In the future, disease progression models, such as the ones presented in this thesis, could become an integral part of targeted drug trials. I envision three important roles of such models in targeted drug trials in the future: data-driven participant selection, monitoring the efficacy of the drugs through biomarker-based objective disease severity scores, and identifying novel therapeutic targets using a better understanding of the pathways of progression.

11.5 Roadmap for future methodological innovations

There are several ways to extend the disease progression models presented in this thesis. Although DEBM was designed to estimate disease progression timeline in the presence of disease heterogeneity, staging patients on this timeline only accounts for progression along this timeline and does not account for any heterogeneity. For addressing this limitation, a 2D patient staging space could be explored where the x-axis and y-axis correspond to progression of the patient along and away from the mean progression timeline.

A novel extension of the EBM framework to identify data-driven subtypes of the disease has already been developed by Young *et al.* [86]. Further extensions of this to account for disease progression due to multiple co-existing pathologies such as due to Amyloid- β and TDP-43 [345] or due to Amyloid- β and vascular pathology [346], could be useful to provide further insights into such amalgamation of symphonies.

One of the limitations identified in Chapter 2 was that the accuracy of EBMs decreases as the number of events increases, even when the average quality of the biomarkers remained the same. I suspect that this is an artefact of an old optimization technique [72] used in my implementation of the generalized Mallows model, an integral part of DEBM. To improve

the scalability of DEBM, future work can focus on using more recent optimization techniques for generalized Mallows model, such as [347], which uses pairwise sequences of events to estimate the overall event sequence.

My focus in this thesis was on disease progression modeling using cross-sectional data, to estimate the mean progression timeline in a cohort or a subset of it. Patient-specific variations of these progression sequences were estimated as a by-product that was not validated extensively in this thesis. Future work could focus on in-depth analysis of these patient-specific variations estimated using scalar as well as n-dimensional biomarkers. Use of longitudinal extensions to EBM using hidden Markov model [348] to improve the accuracy of these personalized disease progression patterns could also be an interesting area of research.

While the EBMs estimate the sequence of events as the disease progresses, the earlier events cannot be interpreted to have a causal relationship with the later events. Developing novel methods that integrates causal inference [349] with disease progression modeling could help in achieving even greater understanding of the pathological pathways of neurodegenerative diseases.

Lastly, the use of deep learning in event-based modeling has not yet been explored. A deep-learning-based EBM could be developed by integrating the best of both worlds, combining the scalability and robustness of deep learning methods with the explainability of an EBM. Such a ‘deep event-based model’ may help for example in elucidating the role of the myriad of proteins obtained from CSF proteomics [350] in the pathophysiology of the disease, or in studying the spatio-temporal progression patterns at an unprecedented spatial resolution based on voxel-wise imaging markers. The nDEBM approach presented in Chapter 7 is a first step in this direction.

11.6 Conclusion

In this thesis, I developed a disease progression modeling framework called DEBM to provide novel insights into the neurodegenerative disease progression using cross-sectional data, as well as objectively assess the disease severity of patients with such diseases. This framework was subsequently improved with variants such as nDEBM and co-init DEBM. These were used to provide novel insights into the underlying order in the heterogeneous progression of four neurodegenerative diseases, namely: AD, FTD, CJD, and MS. The spatio-temporal disease progression timeline of structural abnormalities in the brain obtained using DEBM from AD research cohorts was validated to be generalizable to clinical as well as population based cohorts. Hence I conclude that there is great potential for the models developed in this thesis to be useful for clinical diagnosis of neurodegenerative diseases, finding novel ways for effective patient-care, as well as for targeted interventions of these diseases in the future.

Summary

Neurodegenerative diseases such as Alzheimer's disease (AD) are pathologically and phenotypically heterogeneous. This heterogeneous cacophony of neurodegeneration is orchestrated by the underlying symphony of pathological processes of the disease. Understanding the underlying order in the heterogeneous progression of neurodegenerative is crucial for early diagnosis and in identifying novel treatment strategies for drug trials.

In this thesis, I developed novel disease progression modeling methods to estimate the progression timeline of neurodegenerative diseases in a data-driven way from *in-vivo* patient data. In collaboration with clinical researchers, I used the developed approaches to obtain novel insights into the neurodegenerative disease pathways of four such diseases: AD, frontotemporal dementia (FTD), Creutzfeldt-Jakob disease (CJD) and multiple sclerosis (MS). Lastly, to check the utility of the developed approaches for (pre-)clinical use, their generalizability from research cohorts to clinical cohorts as well as their ability to identify preclinical and prodromal stages in a population-based cohort was validated in this thesis. Each of these aspects is summarized below:

Methodological innovations

In Chapter 2, I developed a novel method called discriminative event-based model (DEBM), to estimate the mean progression timeline in a cross-sectional cohort with heterogeneous disease progression patterns. To achieve this in a robust manner, I developed a novel optimization strategy for Gaussian mixture modeling, which is an integral part of event-based models. Furthermore, the novel concept of event-centers was introduced which estimates the relative (temporal) distance between the biomarker abnormality events. Lastly, a novel patient-staging approach was developed to estimate the disease severity in a patient using the disease progression timeline. I validated the utility of each of these innovations in a synthetic dataset simulating the progression of AD, as well in a large AD cohort.

To further improve the robustness of mixture modeling for DEBM analysis, two methodological novelties were introduced in this thesis. First, while studying the brain's lateral asymmetry in progranulin (*GRN*) related familial FTD, in Chapter 4, a technique called siamese mixture modeling was introduced. In this technique, Gaussian parameters were shared between the corresponding regions from the left and right hemispheres of the brain, while also retaining the flexibility of the model to estimate the relative temporal distance between the two counterparts along the progression timeline. Secondly, while studying effect

of *APOE* on the AD progression timeline, in Chapter 8, I proposed and evaluated novel approaches for coupled training of DEBM in several subgroups. This was motivated from the observation that mixture modeling after stratification into smaller disease subgroups, such as based on genetics, would lead to more inaccurate models as compared to fitting the model on the entire dataset. The approach of co-initializing and independently optimizing the DEBM models (co-init DEBM) was found to be the best strategy for coupled training of DEBM.

For accurate estimation of spatio-temporal sequence of disease events, I further extended DEBM to accommodate high-dimensional imaging data (nDEBM) in Chapter 7. Methodological novelty in nDEBM was in using semi-supervised support vector machines which scaled well to high-dimensional biomarkers, addressing a well-documented scalability problem of the mixture modeling used in the DEBM framework. This method was also validated in a synthetic dataset simulating the progression of AD, as well in a large AD cohort. The synthetic dataset used for this validation was generated using a novel deep-learning based simulation framework.

Further methodological developments include using the output of DEBM for patient prognostication in AD and sporadic CJD (sCJD). I participated in a global challenge to predict the future clinical diagnosis of subjects at risk of developing AD, volume of ventricles in the brain of these subjects, as well as their cognitive summary scores. I developed a novel prediction method for this challenge based on DEBM that considered aging and disease progression as independent factors affecting the progression of biomarker values along with interaction between the two. This ended up as the winning entry for the ventricular volume prediction and came second in the overall challenge. The details of the challenge, comparative analysis of all the submitted approaches, as well a brief algorithmic details of our developed approach are provided in Chapter 6.

In Chapter 10, a novel approach was developed using DEBM for *ante-mortem* identification of the molecular subtypes of sCJD. Since the survival times of sCJD patients with different molecular subtypes are substantially different, this *ante-mortem* identification of the molecular subtypes could be clinically used for patient prognostication. The methodological novelty was in developing a nearest mean classifier using the distances of a single time-point approximate estimate of the disease events in a test-subject from the mean spatio-temporal sequence of disease events in different subtypes.

Novel insights into neurodegenerative diseases

In Chapter 4, we obtained novel insights into the progression of *GRN* related FTD, using DEBM. In FTD-*GRN*, degeneration of axons and language deficits were observed to be two of the earliest disease events.

In Chapter 5, we obtained novel insights into the progression of relapse-onset MS using structural, functional and cognitive biomarkers, using DEBM. Gray matter volumes of thalamus, insula, hippocampus and cerebellum were observed to be among the earliest biomarkers to become abnormal in relapse-onset MS.

In Chapter 8, we determined the effect of *APOE* alleles on the disease progression timeline of AD using co-init DEBM and observed that *APOE*- ϵ 2 carriers, homozygous *APOE*- ϵ 3 carriers, and *APOE*- ϵ 4 carriers have significantly different disease progression timelines in AD.

In Chapter 10, we observed that the two most predominant molecular subtypes in sporadic CJD (MM1 and VV2) have spatio-temporally contrasting lesion propagation timelines.

Multi-cohort validation

The methods developed in this thesis were validated for their generalizability to clinical and population-based cohorts that were not used for training the model. In Chapter 3, we validated the generalizability of the sequence of events provided by DEBM as well as the original EBM methods in multiple clinical cohorts for AD.

In Chapter 9, we investigated if the disease progression timelines of AD constructed in a case-controlled setting using co-init DEBM, with subjects stratified according to *APOE* status, are generalizable to a population-based cohort. Our results validated the generalizability of co-init DEBM and that progression along these disease timelines is predictive of AD.

Conclusion

In summary, my research demonstrates the utility of disease progression models in understanding the complex progression pathways in neurodegenerative diseases. The validated diagnostic and prognostic utilities of the models, their generalizability, together with the transparency they bring to the clinical decision making process, make DEBM and its variants, a strong candidate for aiding in clinical diagnosis and prognosis of neurodegenerative diseases. The insights into disease progression pathways obtained using these models, at a group level as well as for each patient, could be invaluable for targeted interventions and drug trials in the future.

Samenvatting*

Neurodegeneratieve ziekten, zoals de ziekte van Alzheimer (AD), zijn heterogeen in pathologie en fenotype. Deze heterogene kakofonie van neurodegeneratie wordt georkestreerd door de onderliggende symfonie van pathologische processen van de ziekte. Het begrijpen van de onderliggende volgorde waarin de heterogene progressie van neurodegeneratie plaatsvindt is cruciaal voor vroege diagnose en voor het identificeren van nieuwe behandelingsstrategieën voor medicatietrials.

In deze thesis heb ik nieuwe datagestuurde methoden ontwikkeld om het ziekteproces te modeleren. Deze methoden kunnen de tijdslijn van veranderingen in het ziekteproces van neurodegeneratieve ziekten afschatten aan de hand van in-vivo gegevens van patiënten. In samenwerking met klinische onderzoekers heb ik deze ontwikkelde methoden gebruikt om nieuwe inzichten te verkrijgen in het ontstaan van neurodegeneratieve ziekten, in het specifiek: AD, frontotemporale demantie (FTD), de ziekte van Creutzfeldt-Jakob (CJD) en multiple sclerosis (MS). Tenslotte heb ik onderzocht in hoeverre de ontwikkelde methoden geschikt zouden zijn voor toepassing in de klinische praktijk. Ik heb gevalideerd of de methoden generaliseerbaar zijn van onderzoekscohorten naar klinische cohorten en of ze gebruikt kunnen worden om individuen te identificeren die zich in een prodromale fase van de ziekte bevinden. De verschillende aspecten van deze thesis zijn hieronder samengevat.

Methodologische innovaties

In Hoofdstuk 2 heb ik een nieuwe methode ontwikkeld, genaamd discriminative event-based model (DEBM), om de gemiddelde ziekteprogressietijdslijn af te schatten in een cross-sectioneel cohort met heterogene ziekteprogressiepatronen. Om dit op een robuuste manier te kunnen doen, heb ik een nieuwe optimalisatiestrategie ontwikkeld voor Gaussian mixture modeling, dat een integraal onderdeel is van een event-based model. Daarnaast heb ik het nieuwe concept van eventcentra geïntroduceerd. Dit is een schatting van de relatieve (temporele) afstand tussen de momenten waarop verschillende biomarkers abnormaal worden. Tenslotte heb ik een nieuwe patiëntstadiërmethode ontwikkeld om op basis van de ziekteprogressietijdslijn de ernst van de ziekte van een patiënt te schatten. Ik heb de bruikbaarheid van elk van deze innovaties gevalideerd in zowel een synthetische dataset waarin de progressie van AD werd gesimuleerd, als in een groot AD-cohort.

*Translation of the summary by Eline J. Vinke.

Om de robuustheid van mixture modeling voor DEBM-analyses verder te optimaliseren zijn twee nieuwe methodes geïntroduceerd in deze thesis. In Hoofdstuk 4 wordt een techniek genaamd siamese mixture modeling geïntroduceerd voor de analyse van de laterale asymmetrie van het brein in progranuline-gerelateerde (GRN) familiale FTD. Met deze techniek worden Gaussian parameters gedeeld tussen corresponderende regio's van de linker en rechter hersenhelft terwijl het model nog steeds de flexibiliteit heeft om de relatieve temporale afstand tussen deze twee corresponderende regio's in de ziekteprogressietijdlijn af te schatten. In Hoofdstuk 8 introduceer en evalueer ik een nieuwe methode voor gekoppelde training van DEBM in verschillende subgroepen. De reden hiervoor was dat we zagen dat mixture modeling na stratificatie in kleinere subgroepen van ziekte, bijvoorbeeld gebaseerd op genetica, leidde tot een lagere nauwkeurigheid dan wanneer de gehele dataset gebruikt werd. De methode van co-initialiseren en onafhankelijke optimalisatie van DEBM-modellen (co-init DEBM) was de beste strategie voor het gekoppeld trainen van DEBM-modellen.

Voor een nauwkeurige afschatting van de spatio-temporele opeenvolging van ziekte-events, heb ik in Hoofdstuk 7 DEBM verder uitgebreid om het gebruik van hoog-dimensionale beelddata mogelijk te maken (nDEBM). De belangrijkste methodologische ontwikkeling in nDEBM was het gebruik van semi-gesuperviseerde support vector machines, die goed schalen naar hoog-dimensionale biomarkers en zo het schaalbaarheidsprobleem van mixture modeling aanpakken binnen het bestaande DEBM-framework. Deze methode is gevalideerd in zowel een synthetische dataset waarin de ziekteprogressie van AD werd gesimuleerd, als in een groot AD cohort. De synthetische dataset die is gebruikt voor deze validatie was gegenereerd met behulp van nieuwe deep-learning gebaseerde simulatiesoftware.

Andere methodologische ontwikkelingen bestaan uit het gebruik van de DEBM-uitkomsten voor het stellen van een prognose voor patiënten met AD en sporadische CJD (sCJD). Ik heb deelgenomen aan een wereldwijde competitie waarin verschillende voorspellingsmethoden werden vergeleken met gegevens van mensen die een risico lopen op het ontwikkelen van AD. De methoden voorspelden de toekomstige klinische diagnoses, ventrikelvolumes van het brein en samenvattende cognitieve scores. Voor deze competitie heb ik een nieuwe DEBM-gebaseerde voorspellingsmethode ontwikkeld. Deze methode houdt rekening met leeftijd en ziekteprogressie en de effecten hiervan op de verandering van biomarkerwaarden. Hierbij nam ik leeftijd en ziekte mee als onafhankelijke factoren, alsook hun interactie. Deze methode heeft de competitie gewonnen in de categorie van ventrikelvolumevoorspelling en heeft de tweede plaats voor de competitie in zijn geheel behaald. Hoofdstuk 6 beschrijft de details van de competitie, de analyses die gebruikt zijn voor de vergelijking van alle ingestuurde methoden en de algoritmische details van onze ontwikkelde methode.

In Hoofdstuk 10 staat een nieuwe DEBM-gebaseerde methode beschreven die is ontwikkeld voor het *ante-mortem* identificeren van moleculaire subtypes van sCJD. Aangezien de

levensverwachting van sCJD patiënten sterk afhangt van de moleculaire subtypes, zou deze *ante-mortem*-subtype-identificatiemethode ingezet kunnen worden in de kliniek, om een prognose van de levensverwachting van een sCJD patiënt te maken. De methodologische vernieuwing van dit model was de afschatting van de tijd tussen de ziekte-events per individu.

Nieuwe inzichten in neurodegeneratieve ziekten

In Hoofdstuk 4 hebben we met behulp van DEBM nieuwe inzichten verkregen in de progressie van *GRN*-gerelateerde FTD. In FTD-*GRN*, werden de degeneratie van axonen en taalachterstanden geobserveerd als de eerste twee ziekte-events.

In Hoofdstuk 5 hebben we met behulp van DEBM nieuwe inzichten verkregen in de progressie van relapse-onset MS door gebruik te maken van structurele, functionele en cognitieve biomarkers. De grijze stof volumes van de thalamus, insula, hippocampus en het cerebellum blijken de eerste biomarkers te zijn die abnormaal worden in relapse-onset MS.

In Hoofdstuk 8 hebben we met behulp van co-init DEBM het effect van *APOE*-allelen bepaald op de ziekteprogressietijdlijn van AD. We hebben geobserveerd dat *APOE-ε2* dragers, homozygote *APOE-ε3* dragers and *APOE-ε4* dragers een significant verschillende ziekteprogressietijdlijn hebben.

In Hoofdstuk 10 hebben we geobserveerd dat de twee meest dominante moleculaire subtypes in sCJD (MM1 en VV2) een contrasterende laesieverspreiding hebben.

Multi-cohort validatie

De nieuwe methoden zijn gevalideerd met betrekking tot hun generaliseerbaarheid naar klinische cohorten en populatiecohorten die niet gebruikt zijn voor de training van de modellen. In Hoofdstuk 3 hebben we de generaliseerbaarheid bepaald van zowel DEBM als EBM voor het bepalen van de ziekteprogressietijdlijn in AD, in meerdere klinische cohorten.

In Hoofdstuk 9 hebben we onderzocht of de ziekteprogressietijdlijn van AD, geconstrueerd aan de hand van data uit een case-control setting, generaliseerbaar is naar een populatiecohort. Onze resultaten laten zien dat co-init DEBM, waarbij er gestratificeerd is op basis van *APOE*, goed generaliseert. Daarnaast hebben we laten zien dat de ziekteprogressie langs de ziekteprogressietijdlijn voorspellend is voor AD.

Conclusie

Samenvattend laat mijn onderzoek de voordelen zien van ziekteprogressiemodellen voor het begrijpen van het ontstaan van complexe neurodegeneratieve ziekten. De validatie voor diagnostiek en prognostiek, de generaliseerbaarheid, samen met de transparantie, maken DEBM en de DEBM-gebaseerde modellen tot een krachtige potentiële tool voor het ondersteunen van klinische diagnostiek en prognoses van neurodegeneratieve ziekten. De verkregen inzichten in de onderliggende pathologie met behulp van deze modellen, zowel op groepsniveau als op patiëntniveau, zouden van onschatbare waarde kunnen zijn voor gerichte interventies en medicatietrials in de toekomst.

Bibliography

- [1] John Hardy and David Allsop. Amyloid deposition as the central event in the aetiology of Alzheimer's disease. *Trends in Pharmacological Sciences*, 12:383 – 388, 1991.
- [2] Lester I. Binder, Angela L. Guillozet-Bongaarts, Francisco Garcia-Sierra, and Robert W. Berry. Tau, tangles, and Alzheimer's disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1739(2):216 – 223, 2005. The Biology and Pathobiology of Tau.
- [3] Randall J Bateman, Paul S Aisen, Bart De Strooper, Nick C Fox, Cynthia A Lemere, John M Ringman, Stephen Salloway, Reisa A Sperling, Manfred Windisch, and Chengjie Xiong. Autosomal-dominant Alzheimer's disease: a review and proposal for the prevention of Alzheimer's disease. *Alzheimer's research & therapy*, 3(1):1, January 2011.
- [4] David M Holtzman, John C Morris, and Alison M Goate. Alzheimer's disease: the challenge of the second century. *Science translational medicine*, 3(77):77sr1, April 2011.
- [5] Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119 – 128, 2010.
- [6] David B Hogan, Nathalie Jetté, Kirsten M Fiest, Jodie I Roberts, Dawn Pearson, Eric E Smith, Pamela Roach, Andrew Kirk, Tamara Pringsheim, and Colleen J Maxwell. The prevalence and incidence of frontotemporal dementia: a systematic review. *The Canadian journal of neurological sciences. Le journal canadien des sciences neurologiques*, 43 Suppl 1:S96—S109, April 2016.
- [7] Imogen Joanna Swift, Aitana Sogorb-Esteve, Carolin Heller, Matthis Synofzik, Markus Otto, Caroline Graff, Daniela Galimberti, Emily Todd, Amanda J Heslegrave, Emma Louise van der Ende, John Cornelis Van Swieten, Henrik Zetterberg, and Jonathan Daniel Rohrer. Fluid biomarkers in frontotemporal dementia: past, present and future. *Journal of Neurology, Neurosurgery & Psychiatry*, 2020.
- [8] Junhong Yu and Tatia M.C. Lee. The longitudinal decline of white matter microstructural integrity in behavioral variant frontotemporal dementia and its association with executive function. *Neurobiology of Aging*, 76:62 – 70, 2019.
- [9] Esther E. Bron, Rebecca M.E. Steketee, Gavin C. Houston, Ruth A. Oliver, Hakim C. Achterberg, Marco Loog, John C. van Swieten, Alexander Hammers, Wiro J. Niessen, Marion Smits, Stefan Klein, and for the Alzheimer's Disease Neuroimaging Initiative. Diagnostic classification of arterial spin labeling and structural MRI in presenile early stage dementia. *Human Brain Mapping*, 35(9):4916–4931, 2014.
- [10] Lize C. Jiskoot, Jessica L. Panman, Lieke H. Meeter, Elise G.P. Dopper, Laura Donker Kaat, Sanne Franzen, Emma L. van der Ende, Rick van Minkelen, Serge A.R.B. Rombouts, Janne M. Papma, and John C. van Swieten. Longitudinal multimodal MRI as prognostic and diagnostic biomarker in presymptomatic familial frontotemporal dementia. *Brain*, 142(1):193–208, 11 2018.
- [11] C. Davatzikos, S.M. Resnick, X. Wu, P. Parmpi, and C.M. Clark. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of mri. *NeuroImage*, 41(4):1220 – 1227, 2008.
- [12] S. Siri, I. Benaglio, A. Frigerio, G. Binetti, and S. F. Cappa. A brief neuropsychological assessment for the differential diagnosis between frontotemporal dementia and Alzheimer's disease. *European Journal of Neurology*, 8(2):125–132, 2001.

Bibliography

- [13] Michael D Geschwind. Prion diseases. *Continuum (Minneapolis, Minn.)*, 21(6 Neuroinfectious Disease):1612–1638, December 2015.
- [14] Se-Eun Bae, Sunghoon Jung, Ha-Yeon Kim, and Hyeon S. Son. Correlation analysis for the incubation period of prion disease. *Prion*, 6(3):276–281, 2012. PMID: 22561168.
- [15] Alberto Bizzi, Riccardo Pascuzzo, Janis Blevins, Marina Grisoli, Raffaele Lodi, Marco E. M. Moscatelli, Gianmarco Castelli, Mark L. Cohen, Lawrence B. Schonberger, Aaron Foutz, Jiri G. Safar, Brian S. Appleby, and Pierluigi Gambetti. Evaluation of a new criterion for detecting prion disease with diffusion magnetic resonance imaging. *JAMA Neurology*, 77(9):1141–1149, 09 2020.
- [16] Daniel S. Reich, Claudia F. Lucchinetti, and Peter A. Calabresi. Multiple sclerosis. *New England Journal of Medicine*, 378(2):169–180, 2018. PMID: 29320652.
- [17] Roel Klaver, Helga E. De Vries, Geert J. Schenk, and Jeroen J.G. Geurts. Grey matter damage in multiple sclerosis. *Prion*, 7(1):66–75, 2013. PMID: 23324595.
- [18] Iris Dekker and Mike P. Wattjes. Brain and spinal cord MR imaging features in multiple sclerosis and variants. *Neuroimaging Clinics of North America*, 27(2):205 – 227, 2017. Advances in Imaging of Multiple Sclerosis.
- [19] Nancy D Chiaravalloti and John DeLuca. Cognitive impairment in multiple sclerosis. *The Lancet Neurology*, 7(12):1139 – 1151, 2008.
- [20] Claudia H Marck, Emily J Hadgkiss, Tracey J Weiland, Dania M van der Meer, Naresh G Pereira, and George A Jelinek. Physical activity and associated levels of disability and quality of life in people with multiple sclerosis: a large international survey. *BMC neurology*, 14:143, July 2014.
- [21] Akinori Nakamura, Naoki Kaneko, Victor L Villemagne, Takashi Kato, James Doecke, Vincent Doré, Chris Fowler, Qiao-Xin Li, Ralph Martins, Christopher Rowe, Taisuke Tomita, Katsumi Matsuzaki, Kenji Ishii, Kazunari Ishii, Yutaka Arahata, Shinichi Iwamoto, Kengo Ito, Koichi Tanaka, Colin L Masters, and Katsuhiko Yanagisawa. High performance plasma amyloid- β biomarkers for Alzheimer’s disease. *Nature*, 554(7691):249–254, February 2018.
- [22] Sebastian Palmqvist, Shorena Janelidze, Yakeel T. Quiroz, Henrik Zetterberg, Francisco Lopera, Erik Stomrud, Yi Su, Yinghua Chen, Geidy E. Serrano, Antoine Leuzy, Niklas Mattsson-Carlsson, Olof Strandberg, Ruben Smith, Andres Villegas, Diego Sepulveda-Falla, Xiyun Chai, Nicholas K. Proctor, Thomas G. Beach, Kaj Blennow, Jeffrey L. Dage, Eric M. Reiman, and Oskar Hansson. Discriminative accuracy of plasma phospho-tau217 for Alzheimer disease vs other neurodegenerative disorders. *JAMA*, 324(8):772–781, 08 2020.
- [23] Kaj Blennow and Harald Hampel. CSF markers for incipient Alzheimer’s disease. *The Lancet Neurology*, 2(10):605 – 613, 2003.
- [24] William E Klunk and Chester A Mathis. The future of amyloid-beta imaging: a tale of radionuclides and tracer proliferation. *Current opinion in neurology*, 21(6):683–687, December 2008.
- [25] Nobuyuki Okamura, Ryuichi Harada, Shozo Furumoto, Hiroyuki Arai, Kazuhiko Yanai, and Yukitsuka Kudo. Tau PET imaging in Alzheimer’s disease. *Current neurology and neuroscience reports*, 14(11):500, November 2014.
- [26] Sandra Weintraub, Alissa H Wicklund, and David P Salmon. The neuropsychological profile of Alzheimer disease. *Cold Spring Harbor perspectives in medicine*, 2(4):a006171, April 2012.
- [27] Caroline van Cauwenberghe, Christine van Broeckhoven, and Kristel Sleegers. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genetics in medicine*, 18(5):421–430, May 2016.
- [28] Petra Steinacker, Peggy Barschke, and Markus Otto. Biomarkers for diseases with TDP-43 pathology. *Molecular and Cellular Neuroscience*, 97:43 – 59, 2019. Disease signatures: biomarkers/indicators of neurodegeneration.

- [29] Nicholas J Ashton, Abdul Hye, Anto P Rajkumar, Antoine Leuzy, Stuart Snowden, Marc Suárez-Calvet, Thomas K Karikari, Michael Schöll, Renaud La Joie, Gil D Rabinovici, Kina Höglund, Clive Ballard, Tibor Hortobágyi, Per Svenningsson, Kaj Blennow, Henrik Zetterberg, and Dag Aarsland. An update on blood-based biomarkers for non-Alzheimer neurodegenerative disorders. *Nature reviews. Neurology*, 16(5):265–284, May 2020.
- [30] Emma L van der Ende, Lieke H Meeter, Jackie M Poos, Jessica L Panman, Lize C Jiskoot, Elise G P Dopper, Janne M Papma, Frank Jan de Jong, Inge M W Verberk, Charlotte Teunissen, Dimitris Rizopoulos, Carolin Heller, Rhian S Convery, Katrina M Moore, Martina Bocchetta, Mollie Neason, David M Cash, Barbara Borroni, Daniela Galimberti, Raquel Sanchez-Valle, Robert Laforce, Fermin Moreno, Matthis Synofzik, Caroline Graff, Mario Masellis, Maria Carmela Tartaglia, James B Rowe, Rik Vandenberghe, Elizabeth Finger, Fabrizio Tagliavini, Alexandre de Mendonça, Isabel Santana, Chris Butler, Simon Ducharme, Alex Gerhard, Adrian Danek, Johannes Levin, Markus Otto, Giovanni B Frisoni, Stefano Cappa, Yolande A L Pijnenburg, Jonathan D Rohrer, and John C van Swieten. Serum neurofilament light chain in genetic frontotemporal dementia: a longitudinal, multicentre cohort study. *The Lancet Neurology*, 18(12):1103 – 1111, 2019.
- [31] Tiziana Casoli, Susy Paolini, Paolo Fabbietti, Patrizia Fattoretti, Lucia Paciaroni, Katia Fabi, Beatrice Gobbi, Roberta Galeazzi, Roberto Rossi, Fabrizia Lattanzio, and Giuseppe Pelliccioni. Cerebrospinal fluid biomarkers and cognitive status in differential diagnosis of frontotemporal dementia and Alzheimer’s disease. *Journal of International Medical Research*, 47(10):4968–4980, 2019. PMID: 31524025.
- [32] Clive Hamlin, Gianfranco Puoti, Sally Berri, Elliott Sting, Carrie Harris, Mark Cohen, Charles Spear, Alberto Bizzi, Sara M. Debanne, and Douglas Y. Rowland. A comparison of tau and 14-3-3 protein in the diagnosis of Creutzfeldt-Jakob disease. *Neurology*, 79(6):547–552, 2012.
- [33] Sanam Soomro and Chandra Mohan. Biomarkers for sporadic Creutzfeldt-Jakob disease. *Annals of Clinical and Translational Neurology*, 3(6):465–472, 2016.
- [34] Marcello Moccia, Serena Ruggieri, Antonio Ianniello, Ahmed Toosy, Carlo Pozzilli, and Olga Ciccarelli. Advances in spinal cord imaging in multiple sclerosis. *Therapeutic advances in neurological disorders*, 12:1756286419840593, 2019.
- [35] Nicola De Stefano, Laura Airas, Nikolaos Grigoriadis, Heinrich P Mattle, Jonathan O’Riordan, Celia Oreja-Guevara, Finn Sellebjerg, Bruno Stankoff, Agata Walczak, Heinz Wiendl, and Bernd C Kieseier. Clinical relevance of brain volume measures in multiple sclerosis. *CNS drugs*, 28(2):147–156, February 2014.
- [36] William J. Housley, David Pitt, and David A. Hafler. Biomarkers in multiple sclerosis. *Clinical Immunology*, 161(1):51 – 58, 2015. New Horizons in Biomarker Research.
- [37] Sandra Meyer-Moock, You-Shan Feng, Mathias Maeurer, Franz-Werner Dippel, and Thomas Kohlmann. Systematic literature review and validity evaluation of the expanded disability status scale (EDSS) and the multiple sclerosis functional composite (MSFC) in patients with multiple sclerosis. *BMC neurology*, 14:58, 2014.
- [38] Daniel Ferreira, Agneta Nordberg, and Eric Westman. Biological subtypes of Alzheimer disease. *Neurology*, 94(10):436–448, 2020.
- [39] Adam L. Boxer, Michael Gold, Howard Feldman, Bradley F. Boeve, Susan L.-J. Dickinson, Howard Fillit, Carole Ho, Robert Paul, Rodney Pearlman, Margaret Sutherland, Ajay Verma, Stephen P. Arneric, Brian M. Alexander, Bradford C. Dickerson, Earl Ray Dorsey, Murray Grossman, Edward D. Huey, Michael C. Irizarry, William J. Marks, Mario Masellis, Frances McFarland, Debra Niehoff, Chiadi U. Onyike, Sabrina Paganoni, Michael A. Panzara, Kenneth Rockwood, Jonathan D. Rohrer, Howard Rosen, Robert N. Chuck, Holly D. Soares, and Nadine Tatton. New directions in clinical trials for frontotemporal lobar degeneration: Methods and outcome measures. *Alzheimer’s & Dementia*, 2019.

Bibliography

- [40] Andrew F. Hill, Susan Joiner, Jonathan D. F. Wadsworth, Katie C. L. Sidle, Jeanne E. Bell, Herbert Budka, James W. Ironside, and John Collinge. Molecular classification of sporadic Creutzfeldt-Jakob disease. *Brain*, 126(6):1333–1346, 06 2003.
- [41] Hans Lassmann, Wolfgang Brück, and Claudia Lucchinetti. Heterogeneity of multiple sclerosis pathogenesis: implications for diagnosis and therapy. *Trends in Molecular Medicine*, 7(3):115 – 121, 2001.
- [42] Alessandro Finkelsztejn. Multiple sclerosis: Overview of disease-modifying agents. *Perspectives in Medicinal Chemistry*, 6:PMC.S13213, 2014. PMID: 25336899.
- [43] Stephen L. Hauser and Bruce A.C. Cree. Treatment of multiple sclerosis: A review. *The American Journal of Medicine*, 2020.
- [44] Jeff Sevigny, Ping Chiao, Thierry Bussière, Paul H Weinreb, Leslie Williams, Marcel Maier, Robert Dunstan, Stephen Salloway, Tianle Chen, Yan Ling, John O’Gorman, Fang Qian, Mahin Arastu, Mingwei Li, Sowmya Chollate, Melanie S Brennan, Omar Quintero-Monzon, Robert H Scannevin, H Moore Arnold, Thomas Engber, Kenneth Rhodes, James Ferrero, Yaming Hang, Alvydas Mikulskis, Jan Grimm, Christoph Hock, Roger M Nitsch, and Alfred Sandrock. The antibody aducanumab reduces A β plaques in Alzheimer’s disease. *Nature*, 537(7618):50—56, September 2016.
- [45] Gianluigi Forloni, Ignazio Roiter, and Fabrizio Tagliavini. Clinical trials of prion disease therapeutics. *Current Opinion in Pharmacology*, 44:53 – 60, 2019. Neurosciences – Prion disease.
- [46] Jiqing Cao, Jianwei Hou, Jing Ping, and Dongming Cai. Advances in developing novel therapeutic strategies for Alzheimer’s disease. *Molecular neurodegeneration*, 13(1):64, December 2018.
- [47] Francesco Panza, Madia Lozupone, Davide Seripa, Antonio Daniele, Mark Watling, Gianluigi Giannelli, and Bruno P Imbimbo. Development of disease-modifying drugs for frontotemporal dementia spectrum disorders. *Nature reviews. Neurology*, 16(4):213—228, April 2020.
- [48] Peter K Panegyres and Elizabeth Armari. Therapies for human prion diseases. *American journal of neurodegenerative disease*, 2(3):176—186, September 2013.
- [49] Andrew L Smith, Jeffrey A Cohen, and Le H Hua. Therapeutic targets for multiple sclerosis: Current treatment goals and future directions. *Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics*, 14(4):952—960, October 2017.
- [50] H. Braak and E. Braak. Neuropathological staging of Alzheimer-related changes. *Acta Neuropathologica*, 82(4):239–259, Sep 1991.
- [51] Bruno M. Jernak, Andrew Lang, Bo Liu, Elyse Katz, Yanwei Zhang, Bradley T. Wyman, David Raunig, C. Pierre Jernak, Brian Caffo, and Jerry L. Prince. A computational neurodegenerative disease progression score: Method and results with the Alzheimer’s disease neuroimaging initiative cohort. *NeuroImage*, 63(3):1478 – 1486, 2012.
- [52] Michael C. Donohue, Hélène Jacqmin-Gadda, Mélanie Le Goff, Ronald G. Thomas, Rema Raman, Anthony C. Gamst, Laurel A. Beckett, Clifford R. Jack, Michael W. Weiner, Jean-François Dartigues, and Paul S. Aisen. Estimating long-term multivariate progression from short-term data. *Alzheimer’s & Dementia*, 10(5, Supplement):S400 – S410, 2014.
- [53] Hubert M. Fonteijn, Marc Modat, Matthew J. Clarkson, Josephine Barnes, Manja Lehmann, Nicola Z. Hobbs, Rachael I. Scahill, Sarah J. Tabrizi, Sebastien Ourselin, Nick C. Fox, and Daniel C. Alexander. An event-based model for disease progression and its application in familial Alzheimer’s disease and Huntington’s disease. *NeuroImage*, 60(3):1880 – 1889, 2012.
- [54] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, and M. W. Weiner. Alzheimer’s disease neuroimaging initiative (ADNI). *Neurology*, 74(3):201–209, 2010.

- [55] John C Morris, Paul S Aisen, Randall J Bateman, Tammie L S Benzinger, Nigel J Cairns, Anne M Fagan, Bernardino Ghetti, Alison M Goate, David M Holtzman, William E Klunk, Eric McDade, Daniel S Marcus, Ralph N Martins, Colin L Masters, Richard Mayeux, Angela Oliver, Kimberly Quaid, John M Ringman, Martin N Rossor, Stephen Salloway, Peter R Schofield, Natalie J Selsor, Reisa A Sperling, Michael W Weiner, Chengjie Xiong, Krista L Moulder, and Virginia D Buckles. Developing an international network for Alzheimer research: The dominantly inherited Alzheimer network. *Clinical investigation*, 2(10):975–984, October 2012.
- [56] Victor L Villemagne, Samantha Burnham, Pierrick Bourgeat, Belinda Brown, Kathryn A Ellis, Olivier Salvado, Cassandra Szoek, S Lance Macaulay, Ralph Martins, Paul Maruff, David Ames, Christopher C Rowe, and Colin L Masters. Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer’s disease: a prospective cohort study. *The Lancet Neurology*, 12(4):357 – 367, 2013.
- [57] Neil P Oxtoby, Alexandra L Young, David M Cash, Tammie L S Benzinger, Anne M Fagan, John C Morris, Randall J Bateman, Nick C Fox, Jonathan M Schott, and Daniel C Alexander. Data-driven models of dominantly-inherited Alzheimer’s disease progression. *Brain*, 141(5):1529–1544, 03 2018.
- [58] Igor Koval, Jean-Baptiste Schiratti, Alexandre Routier, Michael Bacci, Olivier Colliot, Stéphanie Allassonnière, and Stanley Durrleman. Spatiotemporal propagation of the cortical atrophy: Population and individual patterns. *Frontiers in Neurology*, 9:235, 2018.
- [59] Jonathan Huang and Daniel Alexander. Probabilistic event cascades for Alzheimer’s disease. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3095–3103. Curran Associates, Inc., 2012.
- [60] Y. Iturria-Medina, R. C. Sotero, P. J. Toussaint, J. M. Mateos-Pérez, A. C. Evans, and on behalf of the Alzheimer’s Disease Neuroimaging Initiative. Early role of vascular dysregulation on late-onset Alzheimer’s disease based on multifactorial data-driven analysis. *Nature Communications*, 7:11934, 2016.
- [61] Alexandra L. Young, Neil P. Oxtoby, Pankaj Daga, David M. Cash, Nick C. Fox, Sebastien Ourselin, Jonathan M. Schott, and Daniel C. Alexander. A data-driven model of biomarker changes in sporadic Alzheimer’s disease. *Brain*, 137(9):2564–2577, 2014.
- [62] Martin Prince, Anders Wimo, Maelenn Guerchet, Gemma-Claire Ali, Yu-Tzu Wu, and Matthew Prina. World Alzheimer’s report 2015, the global impact of dementia: An analysis of prevalence, incidence, cost and trends. *Alzheimer’s Disease Int’l*, 2015.
- [63] Clifford R. Jack, David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, Enchi Liu, Jose Luis Molinuevo, Thomas Montine, Creighton Phelps, Katherine P. Rankin, Christopher C. Rowe, Philip Scheltens, Eric Siemers, Heather M. Snyder, Reisa Sperling, Cerise Elliott, Eliezer Masliah, Laurie Ryan, and Nina Silverberg. NIA-AA research framework: Toward a biological definition of Alzheimer’s disease. *Alzheimer’s & Dementia*, 14(4):535 – 562, 2018.
- [64] Mert R. Sabuncu, Jorge L. Bernal-Rusiel, Martin Reuter, Douglas N. Greve, and Bruce Fischl. Event time analysis of longitudinal neuroimage data. *NeuroImage*, 97:9 – 18, 2014.
- [65] Alexander Schmidt-Richberg, Christian Ledig, Ricardo Guerrero, Helena Molina-Abril, Alejandro Frangi, Daniel Rueckert, and on behalf of the Alzheimer’s Disease Neuroimaging Initiative. Learning biomarker models for progression estimation of Alzheimer’s disease. *PLoS ONE*, 11(4):1–27, 2016.
- [66] Dan Li, Samuel Iddi, Wesley K Thompson, Michael C Donohue, and for the Alzheimer’s Disease Neuroimaging Initiative. Bayesian latent time joint mixed effect models for multicohort longitudinal data. *Statistical Methods in Medical Research*, 2017.

Bibliography

- [67] J.-B. Schiratti, S. Allassonnière, A. Routier, O. Colliot, and S. Durrleman. A mixed-effects model with time reparametrization for longitudinal univariate manifold-valued data. In Sebastien Ourselin, Daniel C. Alexander, Carl-Fredrik Westin, and M. Jorge Cardoso, editors, *Information Processing in Medical Imaging*, pages 564–575, Cham, 2015. Springer International Publishing.
- [68] Marco Lorenzi, Maurizio Filippone, Giovanni B. Frisoni, Daniel C. Alexander, and Sebastien Ourselin. Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer’s disease. *NeuroImage*, 2017.
- [69] Razvan V Marinescu, Neil P Oxtoby, Alexandra L Young, Esther E Bron, Arthur W Toga, Michael W Weiner, Frederik Barkhof, Nick C Fox, Stefan Klein, Daniel C Alexander, EuroPOND, and ADNI. Tadpole challenge: Prediction of longitudinal evolution in Alzheimer’s disease. *Arxiv*, 2018.
- [70] Vikram Venkatraghavan, Esther E. Bron, Wiro J. Niessen, and Stefan Klein. A discriminative event based model for Alzheimer’s disease progression modeling. In Marc Niethammer, Martin Styner, Stephen Aylward, Hongtu Zhu, Ipek Oguz, Pew-Thian Yap, and Dinggang Shen, editors, *Information Processing in Medical Imaging*, pages 121–133, Cham, 2017. Springer International Publishing.
- [71] Jonathan M. Schott, Jonathan W. Bartlett, Nick C. Fox, Josephine Barnes, and for the Alzheimer’s Disease Neuroimaging Initiative Investigators. Increased brain atrophy rates in cognitively normal older adults with low cerebrospinal fluid $A\beta_{1-42}$. *Annals of Neurology*, 68(6):825–834, 2010.
- [72] Michael A. Fligner and Joseph S. Verducci. Multistage ranking models. *Journal of the American Statistical Association*, 83(403):892–901, 1988.
- [73] Ravi Kumar and Sergei Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web, WWW ’10*, pages 571–580, New York, NY, USA, 2010. ACM.
- [74] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, June 2010.
- [75] Alexander Hammers, Richard Allom, Matthias J. Koepp, Samantha L. Free, Ralph Myers, Louis Lemieux, Tejal N. Mitchell, David J. Brooks, and John S. Duncan. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human Brain Mapping*, 19(4):224–247, 2003.
- [76] Ioannis S. Gousias, Daniel Rueckert, Rolf A. Heckemann, Leigh Dyet, James P. Boardman, A. David Edwards, and Alexander Hammers. Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. *NeuroImage*, 40(2):672–684, 2008.
- [77] John Ashburner and Karl J. Friston. Unified segmentation. *NeuroImage*, 26(3):839 – 851, 2005.
- [78] Leslie M. Shaw, Hugo Vanderstichele, Malgorzata Knapik-Czajka, Christopher M. Clark, Paul S. Aisen, Ronald C. Petersen, Kaj Blennow, Holly Soares, Adam Simon, Piotr Lewczuk, Robert Dean, Eric Siemers, William Potter, Virginia M.-Y. Lee, and John Q. and Trojanowski. Cerebrospinal fluid biomarker signature in Alzheimer’s disease neuroimaging initiative subjects. *Annals of Neurology*, 65(4):403–413, 2009.
- [79] Alexandra L. Young, Neil P. Oxtoby, Sebastien Ourselin, Jonathan M. Schott, and Daniel C. Alexander. A simulation system for biomarker evolution in neurodegenerative disease. *Medical Image Analysis*, 26(1):47 – 56, 2015.
- [80] Neil P. Oxtoby and Daniel C. Alexander. Imaging plus X: multimodal models of neurodegenerative disease. *Current Opinion in Neurology*, 30(4):371–379, 2017.
- [81] Alexandra L. Young, Neil P. Oxtoby, Jonathan Huang, Razvan V. Marinescu, Pankaj Daga, David M. Cash, Nick C. Fox, Sebastien Ourselin, Jonathan M. Schott, and Daniel C. Alexander. *Multiple orderings of events in disease progression*, pages 711–722. Springer International Publishing, Cham, 2015.

- [82] Clifford R Jack Jr., David S Knopman, William J Jagust, Ronald C Petersen, Michael W Weiner, Paul S Aisen, Leslie M Shaw, Prashanthi Vemuri, Heather J Wiste, Stephen D Weigand, Timothy G Lesnick, Vernon S Pankratz, Michael C Donohue, and John Q Trojanowski. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2):207 – 216, 2013.
- [83] Hélène Amieva, Hélène Jacqmin-Gadda, Jean-Marc Orgogozo, Nicolas Le Carret, Catherine Helmer, Luc Letenneur, Pascale Barberger-Gateau, Colette Fabrigoule, and Jean-François Dartigues. The 9 year cognitive decline before dementia of the Alzheimer type: a prospective population-based study. *Brain*, 128(5):1093–1101, 2005.
- [84] Hélène Amieva, Mélanie Le Goff, Xavier Millet, Jean Marc Orgogozo, Karine Pérès, Pascale Barberger-Gateau, Hélène Jacqmin-Gadda, and Jean François Dartigues. Prodromal Alzheimer's disease: Successive emergence of the clinical symptoms. *Annals of Neurology*, 64(5):492–498, 2008.
- [85] Leonardo C. de Souza, Marie Chupin, Foudil Lamari, Claude Jardel, Delphine Leclercq, Olivier Colliot, Stéphane Lehéricy, Bruno Dubois, and Marie Sarazin. CSF tau markers are correlated with hippocampal volume in Alzheimer's disease. *Neurobiology of Aging*, 33(7):1253 – 1257, 2012.
- [86] Alexandra L Young, Razvan-Valentin V Marinescu, Neil P Oxtoby, Martina Bocchetta, Keir Yong, Nicholas Firth, David M Cash, David L Thomas, Katrina M Dick, Jorge Cardoso, John van Swieten, Barbara Borroni, Daniela Galimberti, Mario Masellis, Maria Carmela Tartaglia, James B Rowe, Caroline Graff, Fabrizio Tagliavini, Giovanni Frisoni, Robert Laforce Jr, Elizabeth Finger, Alexandre Medonça, Sandro Sorbi, Jason D Warren, Sebastian Crutch, Nick C Fox, Sebastien Ourselin, Jonathan M Schott, Jonathan D Rohrer, Daniel C Alexander, GENFI, and ADNI. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nature Communications*, 9:4273, 2018.
- [87] Reisa A. Sperling, Paul S. Aisen, Laurel A. Beckett, David A. Bennett, Suzanne Craft, Anne M. Fagan, Takeshi Iwatsubo, Clifford R. Jack, Jeffrey Kaye, Thomas J. Montine, Denise C. Park, Eric M. Reiman, Christopher C. Rowe, Eric Siemers, Yaakov Stern, Kristine Yaffe, Maria C. Carrillo, Bill Thies, Marcelle Morrison-Bogorad, Molly V. Wagster, and Creighton H. Phelps. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):280 – 292, 2011.
- [88] Mara Ten Kate, Alberto Redolfi, Enrico Peira, Isabelle Bos, Stephanie J Vos, Rik Vandenberghe, Silvy Gabel, Jolien Schaevebeke, Philip Scheltens, Olivier Blin, Jill C Richardson, Regis Bordet, Anders Wallin, Carl Eckerstrom, José Luis Molinuevo, Sebastiaan Engelborghs, Christine Van Broeckhoven, Pablo Martinez-Lage, Julius Popp, Magdalini Tsolaki, Frans R J Verhey, Alison L Baird, Cristina Legido-Quigley, Lars Bertram, Valerija Dobricic, Henrik Zetterberg, Simon Lovestone, Johannes Streffer, Silvia Bianchetti, Gerald P Novak, Jerome Revillard, Mark F Gordon, Zhiyong Xie, Viktor Wottschel, Giovanni Frisoni, Pieter Jelle Visser, and Frederik Barkhof. MRI predictors of amyloid pathology: results from the EMIF-AD multimodal biomarker discovery study. *Alzheimer's Research & Therapy*, 10(1):100, 9 2018.
- [89] Mara Ten Kate, Silvia Ingala, Adam J Schwarz, Nick C Fox, Gaël Chételat, Bart N M van Berckel, Michael Ewers, Christopher Foley, Juan Domingo Gispert, Derek Hill, Michael C Irizarry, Adriaan A Lammertsma, José Luis Molinuevo, Craig Ritchie, Philip Scheltens, Mark E Schmidt, Pieter Jelle Visser, Adam Waldman, Joanna Wardlaw, Sven Haller, and Frederik Barkhof. Secondary prevention of Alzheimer's dementia: neuroimaging contributions. *Alzheimer's research & therapy*, 10(1):112, October 2018.
- [90] Marilyn S. Albert, Steven T. DeKosky, Dennis Dickson, Bruno Dubois, Howard H. Feldman, Nick C. Fox, Anthony Gamst, David M. Holtzman, William J. Jagust, Ronald C. Petersen, Peter J. Snyder, Maria C. Carrillo, Bill Thies, and Creighton H. Phelps. The diagnosis of mild cognitive impairment 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*, 7(3):270 – 279, 2011.

Bibliography

- [91] Bruno Dubois, Howard H Feldman, Claudia Jacova, Harald Hampel, José Luis Molinuevo, Kaj Blennow, Steven T DeKosky, Serge Gauthier, Dennis Selkoe, Randall Bateman, Stefano Cappa, Sebastian Crutch, Sebastiaan Engelborghs, Giovanni B Frisoni, Nick C Fox, Douglas Galasko, Marie-Odile Habert, Gregory A Jicha, Agneta Nordberg, Florence Pasquier, Gil Rabinovici, Philippe Robert, Christopher Rowe, Stephen Salloway, Marie Sarazin, Stéphane Epelbaum, Leonardo C de Souza, Bruno Vellas, Pieter J Visser, Lon Schneider, Yaakov Stern, Philip Scheltens, and Jeffrey L Cummings. Advancing research diagnostic criteria for Alzheimer’s disease: the IWG-2 criteria. *The Lancet Neurology*, 13(6):614 – 629, 2014.
- [92] Clifford R. Jack, David A. Bennett, Kaj Blennow, Maria C. Carrillo, Howard H. Feldman, Giovanni B. Frisoni, Harald Hampel, William J. Jagust, Keith A. Johnson, David S. Knopman, Ronald C. Petersen, Philip Scheltens, Reisa A. Sperling, and Bruno Dubois. A/T/N: an unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology*, 87(5):539–547, 2016.
- [93] Vikram Venkatraghavan, Esther E. Bron, Wiro J. Niessen, and Stefan Klein. Disease progression timeline estimation for Alzheimer’s disease using discriminative event based modeling. *NeuroImage*, 186:518 – 532, 2019.
- [94] Arman Eshaghi, Razvan V Marinescu, Alexandra L Young, Nicholas C Firth, Ferran Prados, M Jorge Cardoso, Carmen Tur, Floriana De Angelis, Niamh Cawley, Wallace J Brownlee, Nicola De Stefano, M Laura Stromillo, Marco Battaglini, Serena Ruggieri, Claudio Gasperini, Massimo Filippi, Maria A Rocca, Alex Rovira, Jaume Sastre-Garriga, Jeroen J G Geurts, Hugo Vrenken, Viktor Wottschel, Cyra E Leurs, Bernard Uitdehaag, Lukas Pirpamer, Christian Enzinger, Sebastien Ourselin, Claudia A Gandini Wheeler-Kingshott, Declan Chard, Alan J Thompson, Frederik Barkhof, Daniel C Alexander, and Olga Ciccarelli. Progression of regional grey matter atrophy in multiple sclerosis. *Brain*, 141(6):1665–1677, 05 2018.
- [95] Peter A Wijeratne, Alexandra L Young, Neil P Oxtoby, Razvan V Marinescu, Nicholas C Firth, Eileanoir B Johnson, Amrita Mohan, Cristina Sampaio, Rachael I Scahill, Sarah J Tabrizi, and Daniel C Alexander. An image-based model of brain volume biomarker changes in huntington’s disease. *Annals of clinical and translational neurology*, 5(5):570—582, May 2018.
- [96] Paul S. Aisen, Ronald C. Petersen, Michael C. Donohue, Anthony Gamst, Rema Raman, Ronald G. Thomas, Sarah Walter, John Q. Trojanowski, Leslie M. Shaw, Laurel A. Beckett, Clifford R. Jack, William Jagust, Arthur W. Toga, Andrew J. Saykin, John C. Morris, Robert C. Green, and Michael W. Weiner. Clinical core of the Alzheimer’s disease neuroimaging initiative: Progress and plans. *Alzheimer’s & Dementia*, 6(3):239 – 246, 2010.
- [97] W.M. van der Flier, Y.A.L. Pijnenburg, N. Prins, A.W. Lemstra, F.H. Bouwman, C.E. Teunissen, B.N.M. van Berckel, C.J. Stam, F. Barkhof, P.J. Visser, E. van Egmond, and P. Scheltens. Optimizing patient care and research: The amsterdam dementia cohort. *Journal of Alzheimer’s Disease*, 41(1):313–327, 2014.
- [98] Giovanni B. Frisoni, Annapaola Prestia, Orazio Zanetti, Samantha Galluzzi, Melissa Romano, Maria Cotelli, Massimo Gennarelli, Giuliano Binetti, Luisella Bocchio, Barbara Paghera, Giovanni Amicucci, Matteo Bonetti, Luisa Benussi, Roberta Ghidoni, and Cristina Geroldi. Markers of Alzheimer’s disease in a population attending a memory clinic. *Alzheimer’s & Dementia*, 5(4):307–317, 2009.
- [99] Katharina Brueggen, Michel J. Grothe, Martin Dyrba, Andreas Fellgiebel, Florian Fischer, Massimo Filippi, Federica Agosta, Peter Nestor, Eva Meisenzahl, Janusch Blautzik, Lutz Frölich, Lucrezia Hausner, Arun L.W. Bokde, Giovanni Frisoni, Michela Pievani, Stefan Klöppel, David Prvulovic, Frederik Barkhof, Petra J.W. Pouwels, Johannes Schröder, Harald Hampel, Karlheinz Hauenstein, and Stefan Teipel. The european DTI study on dementia — a multicenter DTI and MRI study on Alzheimer’s disease and mild cognitive impairment. *NeuroImage*, 144:305 – 308, 2017. Data Sharing Part II.
- [100] Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle

- aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, 2007.
- [101] S. Galluzzi, M. Marizzoni, C. Babiloni, D. Albani, L. Antelmi, C. Bagnoli, D. Bartres-Faz, S. Cordone, M. Didic, L. Farotti, U. Fiedler, G. Forloni, N. Girtler, T. Hensch, J. Jovicich, A. Leeuwis, C. Marra, J. L. Molinuevo, F. Nobili, J. Pariente, L. Parnetti, P. Payoux, C. Del Percio, J.-P. Ranjeva, E. Rolandi, P. M. Rossini, P. Schönknecht, A. Soricelli, M. Tsolaki, P. J. Visser, J. Wiltfang, J. C. Richardson, R. Bordet, O. Blin, G. B. Frisoni, and the PharmaCog Consortium. Clinical and biomarker profiling of prodromal Alzheimer’s disease in workpackage 5 of the innovative medicines initiative PharmaCog project: a ‘European ADNI study’. *Journal of Internal Medicine*, 279(6):576–591, 2016.
- [102] P. Fischer, S. Jungwirth, W. Krampla, S. Weissgram, W. Kirchmeyr, W. Schreiber, K. Huber, M. Rainer, P. Bauer, and K. H. Tragl. Vienna transdanube aging “VITA”: study design, recruitment strategies and level of participation. In Kurt A. Jellinger, Reinhold Schmidt, and Manfred Windisch, editors, *Ageing and Dementia Current and Future Concepts*, pages 105–116, Vienna, 2002. Springer Vienna.
- [103] Kaj Blennow, Harald Hampel, Michael Weiner, and Henrik Zetterberg. Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. *Nature Reviews Neurology*, 6:131 – 144, 2010.
- [104] Stephanie Bombois, Alain Duhamel, Julia Salleron, Vincent Deramecourt, Marie-Anne Mackowiak, Valerie Deken, Nicholas Sergeant, Florence Pasquier, Luc Buee, Bernard Sablonniere, and Susanna Schraen-Maschke. A new decision tree combining $A\beta_{1-42}$ and p-Tau levels in Alzheimer’s diagnosis. *Current Alzheimer Research*, 10(4):357–364, 2013.
- [105] Prashanthi Vemuri and Clifford R Jack. Role of structural MRI in Alzheimer’s disease. *Alzheimer’s research & therapy*, 2(4):23, August 2010.
- [106] Giovanni B. Frisoni, Nick C. Fox, Clifford R. Jack, Philip Scheltens, and Paul M. Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010.
- [107] Ju-Hee Kang, Hugo Vanderstichele, John Q. Trojanowski, and Leslie M. Shaw. Simultaneous analysis of cerebrospinal fluid biomarkers using microsphere-based xMAP multiplex technology for early detection of Alzheimer’s disease. *Methods*, 56(4):484 – 493, 2012. Application of multiplexing technologies in biomedicine.
- [108] John E. Butler. Enzyme-linked immunosorbent assay. *Journal of Immunoassay*, 21(2-3):165–209, 2000.
- [109] Shawn D. Gale, Leslie Baxter, Donald J. Connor, Anne Herring, and James Comer. Sex differences on the rey auditory verbal learning test and the brief visuospatial memory test–revised in the elderly: Normative data in 172 participants. *Journal of Clinical and Experimental Neuropsychology*, 29(5):561–567, 2007.
- [110] András Király, Nikolett Szabó, Eszter Tóth, Gergő Csete, Péter Faragó, Krisztián Kocsis, Anita Must, László Vécsei, and Zsigmond Tamás Kincses. Male brain ages faster: the age and gender dependence of subcortical volumes. *Brain Imaging and Behavior*, 10(3):901–910, 2015.
- [111] R C Gur, P D Mozley, S M Resnick, G L Gottlieb, M Kohn, R Zimmerman, G Herman, S Atlas, R Grossman, and D Berretta. Gender differences in age effect on brain atrophy measured by magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 88(7):2845–2849, 1991.
- [112] Elizabeth R. Delong, David M. Delong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.
- [113] TN Tombaugh and NJ McIntyre. The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*, 40(9):922–935, September 1992.
- [114] S. Hoops, S. Nazem, A. D. Siderowf, J. E. Duda, S. X. Xie, M. B. Stern, and D. Weintraub. Validity of the MoCA and MMSE in the detection of MCI and dementia in parkinson disease. *Neurology*, 73(21):1738–1745, 2009.

Bibliography

- [115] Rongjian Li, Wenlu Zhang, Heung-Il Suk, Li Wang, Jiang Li, Dinggang Shen, and Shuiwang Ji. Deep learning based imaging data completion for improved brain disease diagnosis. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, pages 305–312, Cham, 2014. Springer International Publishing.
- [116] H. Braak, E. Braak, and Bohl J. Staging of Alzheimer-related cortical destruction. *European Neurology*, 33:403–408, 1993.
- [117] George S. Bloom. Amyloid- β and Tau: the trigger and bullet in Alzheimer disease pathogenesis. *JAMA Neurology*, 71(4):505–508, 04 2014.
- [118] Armando Estévez-González, Jaime Kulisevsky, Anunciación Boltes, Pilar Otermín, and Carmen García-Sánchez. Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer’s disease: comparison with mild cognitive impairment and normal aging. *International Journal of Geriatric Psychiatry*, 18(11):1021–1028, 2003.
- [119] WG Rosen, RC Mohs, and KL Davis. A new rating scale for Alzheimer’s disease. *The American journal of psychiatry*, 141(11):1356–1364, November 1984.
- [120] E. C. Mormino, J. T. Kluth, C. M. Madison, G. D. Rabinovici, S. L. Baker, B. L. Miller, R. A. Koeppe, C. A. Mathis, M. W. Weiner, W. J. Jagust, and the Alzheimer’s Disease Neuroimaging Initiative. Episodic memory loss is related to hippocampal-mediated β -amyloid deposition in elderly subjects. *Brain*, 132(5):1310–1323, 11 2008.
- [121] Mara Ten Kate, Frederik Barkhof, Pieter Jelle Visser, Charlotte E Teunissen, Philip Scheltens, Wiesje M van der Flier, and Betty M Tijms. Amyloid-independent atrophy patterns predict time to progression to dementia in mild cognitive impairment. *Alzheimer’s research & therapy*, 9(1):73, September 2017.
- [122] Auriel A. Willette, Vince D. Calhoun, Josephine M. Egan, and Dimitrios Kapogiannis. Prognostic classification of mild cognitive impairment and Alzheimer’s disease: MRI independent component analysis. *Psychiatry Research: Neuroimaging*, 224(2):81 – 88, 2014.
- [123] Robert Perneczky, Stefan Wagenpfeil, Katja Komossa, Timo Grimmer, Janine Diehl, and Alexander Kurz. Mapping scores onto stages: mini-mental state examination and clinical dementia rating. *The American Journal of Geriatric Psychiatry*, 14(2):139 – 144, 2006.
- [124] David A. Bennett, Julie A. Schneider, Zoe Arvanitakis, and Robert S. Wilson. Overview and findings from the religious orders study. *Current Alzheimer Research*, 9(6):628–645, 2012.
- [125] David A. Bennett, Julie A. Schneider, Aron S. Buchman, Lisa L. Barnes, Patricia A. Boyle, and Robert S. Wilson. Overview and findings from the rush memory and aging project. *Current Alzheimer Research*, 9(6):646–663, 2012.
- [126] Walter A. Kukull, Roger Higdon, James D. Bowen, Wayne C. McCormick, Linda Teri, Gerard D. Schellenberg, Gerald van Belle, Lance Jolley, and Eric B. Larson. Dementia and Alzheimer disease incidence: A prospective cohort study. *Archives of Neurology*, 59(11):1737–1746, 11 2002.
- [127] Duane L. Beekly, Erin M. Ramos, William W. Lee, Woodrow D. Deitrich, Mary E. Jacka, Joylee Wu, Janene L. Hubbard, Thomas D. Koepsell, John C. Morris, Walter A. Kukull, and NIA Alzheimer’s disease centers. NIA Alzheimer’s disease centers. the national Alzheimer’s coordinating center (NACC) database: the uniform data set. *Alzheimer’s & Dementia*, 21(3):249 – 258, 2007.
- [128] Alberto Redolfi, Paolo Bosco, David Manset, Giovanni B Frisoni, and neuGRID consortium. Brain investigation and brain conceptualization. *Functional neurology*, 28(3):175—190, 2013.
- [129] A. Redolfi, D. Manset, F. Barkhof, L.O. Wahlund, T. Glatard, J.F. Mangin, and G. B. Frisoni. Head-to-head comparison of two popular cortical thickness extraction algorithms: A cross-sectional and longitudinal study. *PLoS ONE*, 10(3), 2015.

- [130] Giovanni B. Frisoni, Alberto Redolfi, David Manset, Marc-Étienne Rousseau, Arthur Toga, and Alan C. Evans. Virtual imaging laboratories for marker discovery in neurodegenerative diseases. *Nature Reviews Neurology*, 7(8):429–438, May 2011.
- [131] Harro Seelaar, Jonathan D Rohrer, Yolande A L Pijnenburg, Nick C Fox, and John C van Swieten. Clinical, genetic and pathological heterogeneity of frontotemporal dementia: a review. *Journal of Neurology, Neurosurgery & Psychiatry*, 82(5):476–486, 2011.
- [132] John C van Swieten and Peter Heutink. Mutations in progranulin (GRN) within the spectrum of clinical and pathological phenotypes of frontotemporal dementia. *The Lancet Neurology*, 7(10):965 – 974, 2008.
- [133] David M.A. Mann and Julie S. Snowden. Frontotemporal lobar degeneration: pathogenesis, pathology and pathways to phenotype. *Brain Pathology*, 27(6):723–736, 2017.
- [134] Ione O. C. Woollacott and Jonathan D. Rohrer. The clinical spectrum of sporadic and familial forms of frontotemporal dementia. *Journal of Neurochemistry*, 138(S1):6–31, 2016.
- [135] Babykumari P Chitramuthu, Hugh P J Bennett, and Andrew Bateman. Progranulin: a new avenue towards the understanding and treatment of neurodegenerative disease. *Brain*, 140(12):3081–3104, 08 2017.
- [136] Lize C. Jiskoot, Jessica L. Panman, Lauren van Asseldonk, Sanne Franzen, Lieke H. H. Meeter, Laura Donker Kaat, Emma L. van der Ende, Elise G. P. Dopper, Reinier Timman, Rick van Minkelen, John C. van Swieten, Esther van den Berg, and Janne M. Papma. Longitudinal cognitive biomarkers predicting symptom onset in presymptomatic frontotemporal dementia. *Journal of Neurology*, 265(6):1381–1392, Jun 2018.
- [137] Lieke H. Meeter, Elise G. Dopper, Lize C. Jiskoot, Raquel Sanchez-Valle, Caroline Graff, Luisa Benussi, Roberta Ghidoni, Yolande A. Pijnenburg, Barbara Borroni, Daniela Galimberti, Robert Jr Laforce, Mario Masellis, Rik Vandenberghe, Isabelle Le Ber, Markus Otto, Rick van Minkelen, Janne M. Papma, Serge A. Rombouts, Mircea Balasa, Linn Öijerstedt, Vesna Jelic, Katrina M. Dick, David M. Cash, Sophie R. Harding, M. Jorge Cardoso, Sebastien Ourselin, Martin N. Rossor, Alessandro Padovani, Elio Scarpini, Chiara Fenoglio, Maria C. Tartaglia, Foudil Lamari, Christian Barro, Jens Kuhle, Jonathan D. Rohrer, Charlotte E. Teunissen, and John C. van Swieten. Neurofilament light chain: a biomarker for genetic frontotemporal dementia. *Annals of Clinical and Translational Neurology*, 3(8):623–636, 2016.
- [138] Jonathan D Rohrer, Jennifer M Nicholas, David M Cash, John van Swieten, Elise Dopper, Lize Jiskoot, Rick van Minkelen, Serge A Rombouts, M Jorge Cardoso, Shona Clegg, Miklos Espak, Simon Mead, David L Thomas, Enrico De Vita, Mario Masellis, Sandra E Black, Morris Freedman, Ron Keren, Bradley J MacIntosh, Ekaterina Rogaeva, David Tang-Wai, Maria Carmela Tartaglia, Robert Laforce, Fabrizio Tagliavini, Pietro Tiraboschi, Veronica Redaelli, Sara Prioni, Marina Grisoli, Barbara Borroni, Alessandro Padovani, Daniela Galimberti, Elio Scarpini, Andrea Arighi, Giorgio Fumagalli, James B Rowe, Ian Coyle-Gilchrist, Caroline Graff, Marie Fallström, Vesna Jelic, Anne Kinhult Ståhlbom, Christin Andersson, Håkan Thonberg, Lena Lilius, Giovanni B Frisoni, Giuliano Binetti, Michela Pievani, Martina Bocchetta, Luisa Benussi, Roberta Ghidoni, Elizabeth Finger, Sandro Sorbi, Benedetta Nacmias, Gemma Lombardi, Cristina Polito, Jason D Warren, Sebastien Ourselin, Nick C Fox, and Martin N Rossor. Presymptomatic cognitive and neuroanatomical changes in genetic frontotemporal dementia in the genetic frontotemporal dementia initiative (GENFI) study: a cross-sectional analysis. *The Lancet Neurology*, 14(3):253 – 262, 2015.
- [139] Katya Rascovsky, John R. Hodges, David Knopman, Mario F. Mendez, Joel H. Kramer, John Neuhaus, John C. van Swieten, Harro Seelaar, Elise G. P. Dopper, Chiadi U. Onyike, Argye E. Hillis, Keith A. Josephs, Bradley F. Boeve, Andrew Kertesz, William W. Seeley, Katherine P. Rankin, Julene K. Johnson, Maria-Luisa Gorno-Tempini, Howard Rosen, Caroline E. Prioleau-Latham, Albert Lee, Christopher M. Kipps, Patricia Lillo, Olivier Piguet, Jonathan D. Rohrer, Martin N. Rossor, Jason D. Warren, Nick C. Fox, Douglas Galasko, David P. Salmon, Sandra E. Black,

Bibliography

- Marsel Mesulam, Sandra Weintraub, Brad C. Dickerson, Janine Diehl-Schmid, Florence Pasquier, Vincent Deramecourt, Florence Lebert, Yolande Pijnenburg, Tiffany W. Chow, Facundo Manes, Jordan Grafman, Stefano F. Cappa, Morris Freedman, Murray Grossman, and Bruce L. Miller. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain*, 134(9):2456–2477, 08 2011.
- [140] M.L. Gorno-Tempini, A.E. Hillis, S. Weintraub, A. Kertesz, M. Mendez, S.F. Cappa, J.M. Ogar, J.D. Rohrer, S. Black, B.F. Boeve, F. Manes, N.F. Dronkers, R. Vandenberghe, K. Rascovsky, K. Patterson, B.L. Miller, D.S. Knopman, J.R. Hodges, M.M. Mesulam, and M. Grossman. Classification of primary progressive aphasia and its variants. *Neurology*, 76(11):1006–1014, 2011.
- [141] Jacy Bezerra Parnera, Roberta Diehl Rodriguez, Adalberto Studart Neto, Ricardo Nitrini, and Sonia Maria Dozzi Brucki. Corticobasal syndrome: A diagnostic conundrum. *Dementia & Neuropsychologia*, 10:267 – 275, 12 2016.
- [142] Jessica L. Panman, Lize C. Jiskoot, Mark J.R.J. Bouts, Lieke H.H. Meeter, Emma L. van der Ende, Jackie M. Poos, Rogier A. Feis, Anneke J.A. Kievit, Rick van Minkelen, Elise G.P. Dopfer, Serge A.R.B. Rombouts, John C. van Swieten, and Janne M. Papma. Gray and white matter changes in presymptomatic genetic frontotemporal dementia: a longitudinal MRI study. *Neurobiology of Aging*, 76:115 – 124, 2019.
- [143] David S. Knopman, Joel H. Kramer, Bradley F. Boeve, Richard J. Caselli, Neill R. Graff-Radford, Mario F. Mendez, Bruce L. Miller, and Nathaniel Mercaldo. Development of methodology for conducting clinical trials in frontotemporal lobar degeneration. *Brain*, 131(11):2957–2968, 10 2008.
- [144] J. L. Cummings, M. Mega, K. Gray, S. Rosenberg-Thompson, D. A. Carusi, and J. Gornbein. The neuropsychiatric inventory: Comprehensive assessment of psychopathology in dementia. *Neurology*, 44(12):2308–2308, 1994.
- [145] E. Mioshi, S. Hsieh, S. Savage, M. Hornberger, and J.R. Hodges. Clinical staging and disease progression in frontotemporal dementia. *Neurology*, 74(20):1591–1597, 2010.
- [146] Myriam Barandiaran, Fermín Moreno, María de Arriba, Begoña Indakoetxea, Irati Boda, Alazne Gabilondo, Mikel Tainta, and Adolfo López de Munain. Longitudinal neuropsychological study of presymptomatic c.709-1 > A progranulin mutation carriers. *Journal of the International Neuropsychological Society*, 25(1):39–47, 2019.
- [147] Lize C. Jiskoot, Martina Bocchetta, Jennifer M. Nicholas, David M. Cash, David Thomas, Marc Modat, Sebastien Ourselin, Serge A.R.B. Rombouts, Elise G.P. Dopfer, Lieke H. Meeter, Jessica L. Panman, Rick van Minkelen, Emma L. van der Ende, Laura Donker Kaat, Yolande A.L. Pijnenburg, Barbara Borroni, Daniela Galimberti, Mario Masellis, Maria Carmela Tartaglia, James Rowe, Caroline Graff, Fabrizio Tagliavini, Giovanni B. Frisoni, Robert Laforce Jr, Elizabeth Finger, Alexandre de Mendonça, Sandro Sorbi, on behalf of the Genetic Frontotemporal dementia Initiative (GENFI), Janne M. Papma, John C. van Swieten, and Jonathan D. Rohrer. Presymptomatic white matter integrity loss in familial frontotemporal dementia in the GENFI cohort: A cross-sectional diffusion tensor imaging study. *Annals of Clinical and Translational Neurology*, 5(9):1025–1036, 2018.
- [148] Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, Mark W. Woolrich, and Stephen M. Smith. FSL. *NeuroImage*, 62(2):782 – 790, 2012. 20 YEARS OF fMRI.
- [149] D. Louis Collins, C. J. Holmes, T. M. Peters, and A. C. Evans. Automatic 3 – d model-based neuroanatomical segmentation. *Human Brain Mapping*, 3(3):190–208, 1995.
- [150] Susumu Mori, S. Wakana, Peter C M van Zijl, and L.M. Nagae-Poetscher. *MRI atlas of human white matter*. Elsevier, 2005.
- [151] Ricarda A. L. Menke, Elizabeth Gray, Ching-Hua Lu, Jens Kuhle, Kevin Talbot, Andrea Malaspina, and Martin R. Turner. CSF neurofilament light chain reflects corticospinal tract degeneration in ALS. *Annals of Clinical and Translational Neurology*, 2(7):748–755, 2015.

- [152] Hidekazu Tomimoto. White matter integrity and cognitive dysfunction: Radiological and neuropsychological correlations. *Geriatrics & Gerontology International*, 15(S1):3–9, 2015.
- [153] Yoshiki Hase, Karen Horsburgh, Masafumi Ihara, and Raj N. Kalaria. White matter degeneration in vascular and other ageing-related dementias. *Journal of Neurochemistry*, 144(5):617–633, 2018.
- [154] Chris J D Hardy, Aisling H Buckley, Laura E Downey, Manja Lehmann, Vitor C Zimmerer, Rosemary A Varley, Sebastian J Crutch, Jonathan D Rohrer, Elizabeth K Warrington, and Jason D Warren. The language profile of behavioral variant frontotemporal dementia. *Journal of Alzheimer's disease : JAD*, 50(2):359–371, 2016.
- [155] Sharon Ash, Naomi Nevler, Jeffrey Phillips, David J. Irwin, Corey T. McMillan, Katya Rascovsky, and Murray Grossman. A longitudinal study of speech production in primary progressive aphasia and behavioral variant frontotemporal dementia. *Brain and Language*, 194:46 – 57, 2019.
- [156] Naomi Nevler, Sharon Ash, Charles Jester, David J. Irwin, Mark Liberman, and Murray Grossman. Automatic measurement of prosody in behavioral variant FTD. *Neurology*, 89(7):650–656, 2017.
- [157] M Marsel Mesulam. Primary progressive aphasia and the left hemisphere language network. *Dementia and neurocognitive disorders*, 15(4):93–102, December 2016.
- [158] Adam M Staffaroni, Peter A Ljubenkov, John Kornak, Yann Cobigo, Samir Datta, Gabe Marx, Samantha M Walters, Kevin Chiang, Nick Olney, Fanny M Elahi, David S Knopman, Bradford C Dickerson, Bradley F Boeve, Maria Luisa Gorno-Tempini, Salvatore Spina, Lea T Grinberg, William W Seeley, Bruce L Miller, Joel H Kramer, Adam L Boxer, and Howard J Rosen. Longitudinal multimodal imaging and clinical endpoints for frontotemporal dementia clinical trials. *Brain*, 142(2):443–459, 01 2019.
- [159] Murray Grossman. The non-fluent/agrammatic variant of primary progressive aphasia. *The Lancet Neurology*, 11(6):545 – 555, 2012.
- [160] Jonathan D Rohrer and Howard J Rosen. Neuroimaging in frontotemporal dementia. *International Review of Psychiatry*, 25(2):221–229, 2013.
- [161] Carole H. Sudre, Martina Bocchetta, David Cash, David L. Thomas, Ione Woollacott, Katrina M. Dick, John van Swieten, Barbara Borroni, Daniela Galimberti, Mario Masellis, Maria Carmela Tartaglia, James B. Rowe, Caroline Graff, Fabrizio Tagliavini, Giovanni Frisoni, Robert Laforce, Elizabeth Finger, Alexandre de Mendonça, Sandro Sorbi, Sébastien Ourselin, M. Jorge Cardoso, Jonathan D. Rohrer, Christin Andersson, Silvana Archetti, Andrea Arighi, Luisa Benussi, Giuliano Binetti, Sandra Black, Maura Cosseddu, Marie Fallström, Carlos Ferreira, Chiara Fenoglio, Nick C. Fox, Morris Freedman, Giorgio Fumagalli, Stefano Gazzina, Roberta Ghidoni, Marina Grisoli, Vesna Jelic, Lize Jiskoot, Ron Keren, Gemma Lombardi, Carolina Maruta, Simon Mead, Lieke Meeter, Rick van Minkelen, Benedetta Nacmias, Linn Öjsterstedt, Alessandro Padovani, Jessica Panman, Michela Pievani, Cristina Polito, Enrico Premi, Sara Prioni, Rosa Rademakers, Veronica Redaelli, Ekaterina Rogaeva, Giacomina Rossi, Martin N. Rossor, Elio Scarpini, David Tang-Wai, Hakan Thonberg, Pietro Tiraboschi, Ana Verdelho, and Jason D. Warren. White matter hyperintensities are seen only in GRN mutation carriers in the GENFI cohort. *NeuroImage: Clinical*, 15:171 – 180, 2017.
- [162] Suzee E. Lee, Ana C. Sias, Eena L. Kosik, Taru M. Flagan, Jersey Deng, Stephanie A. Chu, Jesse A. Brown, Anna A. Vidovszky, Eliana Marisa Ramos, Maria Luisa Gorno-Tempini, Anna M. Karydas, Giovanni Coppola, Daniel H. Geschwind, Rosa Rademakers, Bradley F. Boeve, Adam L. Boxer, Howard J. Rosen, Bruce L. Miller, and William W. Seeley. Thalamo-cortical network hyperconnectivity in preclinical proggranulin mutation carriers. *NeuroImage: Clinical*, 22:101751, 2019.
- [163] Alberto Benussi, Stefano Gazzina, Enrico Premi, Maura Cosseddu, Silvana Archetti, Valentina Dell’Era, Valentina Cantoni, Maria Sofia Cotelli, Antonella Alberici, Anna Micheli, Luisa Benussi, Roberta Ghidoni, Alessandro Padovani, and Barbara Borroni. Clinical and biomarker changes in presymptomatic genetic frontotemporal dementia. *Neurobiology of Aging*, 76:133 – 140, 2019.

Bibliography

- [164] Cristian Bonvicini, Elena Milanese, Andrea Pilotto, Nadia Cattane, Enrico Premi, Silvana Archetti, Alessandro Padovani, Massimo Gennarelli, and Barbara Borroni. Understanding phenotype variability in frontotemporal lobar degeneration due to granulin mutation. *Neurobiology of Aging*, 35(5):1206 – 1211, 2014.
- [165] Howard J. Rosen, Bradley F. Boeve, and Adam L. Boxer. Tracking disease progression in familial and sporadic frontotemporal lobar degeneration: Recent findings from ARTFL and LEFFTDS. *Alzheimer's & Dementia*, 16(1):71–78, 2020.
- [166] Jonathan D. Rohrer, Ione O.C. Woollacott, Katrina M. Dick, Emilie Brotherhood, Elizabeth Gordon, Alexander Fellows, Jamie Toombs, Ronald Druyeh, M. Jorge Cardoso, Sebastien Ourselin, Jennifer M. Nicholas, Niklas Norgren, Simon Mead, Ulf Andreasson, Kaj Blennow, Jonathan M. Schott, Nick C. Fox, Jason D. Warren, and Henrik Zetterberg. Serum neurofilament light chain protein is a measure of disease intensity in frontotemporal dementia. *Neurology*, 87(13):1329–1336, 2016.
- [167] Myriam Barandiaran, Ainara Estanga, Fermín Moreno, Begoña Indakoetxea, Ainhoa Alzualde, Nekane Balluerka, José Félix Martí Massó, and Adolfo López de Munain. Neuropsychological features of asymptomatic c.709-1 > A progranulin mutation carriers. *Journal of the International Neuropsychological Society*, 18(6):1086–1090, 2012.
- [168] Lize C. Jiskoot, Elise G.P. Dopper, Tom den Heijer, Reinier Timman, Rick van Minkelen, John C. van Swieten, and Janne M. Papma. Presymptomatic cognitive decline in familial frontotemporal dementia. *Neurology*, 87(4):384–391, 2016.
- [169] B. Borroni, A. Alberici, E. Premi, S. Archetti, V. Garibotto, C. Agosti, R. Gasparotti, M. Di Luca, D. Perani, and A. Padovani. Brain magnetic resonance imaging structural changes in a pedigree of asymptomatic progranulin mutation carriers. *Rejuvenation Research*, 11(3):585–595, 2008.
- [170] David M. Cash, Martina Bocchetta, David L. Thomas, Katrina M. Dick, John C. van Swieten, Barbara Borroni, Daniela Galimberti, Mario Masellis, Maria Carmela Tartaglia, James B. Rowe, Caroline Graff, Fabrizio Tagliavini, Giovanni B. Frisoni, Robert Laforce, Elizabeth Finger, Alexandre de Mendonça, Sandro Sorbi, Martin N. Rossor, Sebastien Ourselin, Jonathan D. Rohrer, Christin Andersson, Silvana Archetti, Andrea Arighi, Luisa Benussi, Sandra Black, Maura Cosseddu, Marie Fallström, Carlos Ferreira, Chiara Fenoglio, Nick Fox, Morris Freedman, Giorgio Fumagalli, Stefano Gazzina, Roberta Ghidoni, Marina Grisoli, Vesna Jelic, Lize Jiskoot, Ron Keren, Gemma Lombardi, Carolina Maruta, Simon Mead, Lieke Meeter, Rick van Minkelen, Benedetta Nacmias, Linn Öijerstedt, Alessandro Padovani, Jessica Panman, Michela Pievani, Cristina Polito, Enrico Premi, Sara Prioni, Rosa Rademakers, Veronica Redaelli, Ekaterina Rogaeva, Giacomina Rossi, Martin Rossor, Elio Scarpini, David Tang-Wai, Carmela Tartaglia, Hakan Thonberg, Pietro Tiraboschi, Ana Verdelho, and Jason Warren. Patterns of gray matter atrophy in genetic frontotemporal dementia: results from the GENFI study. *Neurobiology of Aging*, 62:191 – 196, 2018.
- [171] Enrico Premi, Valentina Garibotto, Stefano Gazzina, Anna Formenti, Silvana Archetti, Roberto Gasparotti, Alessandro Padovani, and Barbara Borroni. Subcortical and deep cortical atrophy in frontotemporal dementia due to granulin mutations. *Dementia and geriatric cognitive disorders extra*, 4(1):95—102, January 2014.
- [172] Claudia Jacova, Ging-Yuek R. Hsiung, Itthipol Tawankanjanachot, Katie Dinelle, Siobhan McCormick, Marjorie Gonzalez, Hyunsoo Lee, Pheth Sengdy, Phoenix Bouchard-Kerr, Matthew Baker, Rosa Rademakers, Vesna Sossi, A. Jon Stoessl, Howard H. Feldman, and Ian R. Mackenzie. Anterior brain glucose hypometabolism predates dementia in progranulin mutation carriers. *Neurology*, 81(15):1322–1331, 2013.
- [173] Karteek Popuri, Emma Dowds, Mirza Faisal Beg, Rakesh Balachandar, Mahadev Bhalla, Claudia Jacova, Adrienne Buller, Penny Slack, Pheth Sengdy, Rosa Rademakers, Dana Wittenberg, Howard H. Feldman, Ian R. Mackenzie, and Ging-Yuek R. Hsiung. Gray matter changes in asymptomatic C9orf72 and GRN mutation carriers. *NeuroImage: Clinical*, 18:591 – 598, 2018.

- [174] Fermín Moreno, Roser Sala-Llloch, Myriam Barandiaran, Raquel Sánchez-Valle, Ainara Estanga, David Bartrés-Faz, Andone Sistiaga, Ainhoa Alzualde, Esther Fernández, José Félix Martí Massó, Adolfo López de Munain, and Begoña Indakoetxea. Distinctive age-related temporal cortical thinning in asymptomatic granulin gene mutation carriers. *Neurobiology of Aging*, 34(5):1462 – 1468, 2013.
- [175] Elise G.P. Dopfer, Serge A.R.B. Rombouts, Lize C. Jiskoot, Tom den Heijer, J. Roos A. de Graaf, Inge de Koning, Anke R. Hammerschlag, Harro Seelaar, William W. Seeley, Ilya M. Veer, Mark A. van Buchem, Patrizia Rizzu, and John C. van Swieten. Structural and functional brain connectivity in presymptomatic familial frontotemporal dementia. *Neurology*, 83(2):e19–e26, 2014.
- [176] Michela Pievani, Donata Paternicò, Luisa Benussi, Giuliano Binetti, Alberto Orlandini, Milena Cobelli, Silvia Magnaldi, Roberta Ghidoni, and Giovanni B. Frisoni. Pattern of structural and functional brain abnormalities in asymptomatic granulin mutation carriers. *Alzheimer's & Dementia*, 10(5, Supplement):S354 – S363.e1, 2014.
- [177] Paola Caroppo, Marie-Odile Habert, Stanley Durrleman, Aurélie Funkiewiez, Vincent Perlberg, Valérie Hahn, Hugo Bertin, Malo Gaubert, Alexandre Routier, Didier Hannequin, Vincent Deramecourt, Florence Pasquier, Sophie Rivaud-Pechoux, Martine Vercelletto, Geoffrey Edouart, Romain Valabregue, Pascal Lejeune, Mira Didic, Jean-Christophe Corvol, Habib Benali, Stephane Lehericy, Bruno Dubois, Olivier Colliot, Alexis Brice, Isabelle Le Ber, and Predict-PGRN study group. Lateral temporal lobe: An early imaging marker of the presymptomatic GRN disease? *Journal of Alzheimer's Disease*, 47(3):751–759, Mar 2015.
- [178] Christopher A. Olm, Corey T. McMillan, David J. Irwin, Vivianna M. Van Deerlin, Philip A. Cook, James C. Gee, and Murray Grossman. Longitudinal structural gray matter and white matter MRI changes in presymptomatic progranulin mutation carriers. *NeuroImage: Clinical*, 19:497 – 506, 2018.
- [179] Lieke H H Meeter, Holger Patzke, Gordon Loewen, Elise G P Dopfer, Yolande A L Pijnenburg, Rick van Minkelen, and John C van Swieten. Progranulin levels in plasma and cerebrospinal fluid in granulin mutation carriers. *Dementia and geriatric cognitive disorders extra*, 6(2):330—340, 2016.
- [180] Gwenaëlle Douaud, Stephen Smith, Mark Jenkinson, Timothy Behrens, Heidi Johansen-Berg, John Vickers, Susan James, Natalie Voets, Kate Watkins, Paul M. Matthews, and Anthony James. Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia. *Brain*, 130(9):2375–2386, 08 2007.
- [181] Stephen M. Smith, Mark Jenkinson, Mark W. Woolrich, Christian F. Beckmann, Timothy E.J. Behrens, Heidi Johansen-Berg, Peter R. Bannister, Marilena De Luca, Ivana Drobnyak, David E. Flitney, Rami K. Niazy, James Saunders, John Vickers, Yongyue Zhang, Nicola De Stefano, J. Michael Brady, and Paul M. Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23:S208 – S219, 2004. Mathematics in Brain Imaging.
- [182] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, Jan 2001.
- [183] Stephen M. Smith, Mark Jenkinson, Heidi Johansen-Berg, Daniel Rueckert, Thomas E. Nichols, Clare E. Mackay, Kate E. Watkins, Olga Ciccarelli, M. Zaheer Cader, Paul M. Matthews, and Timothy E.J. Behrens. Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage*, 31(4):1487 – 1505, 2006.
- [184] E. Kaplan, H. Goodglass, and S. Weintraub. *Boston Naming Test*. Philadelphia: Lea & Febiger, 1983.
- [185] Louis Leon Thurstone. *Primary mental abilities*, volume 119. University of Chicago Press Chicago, 1938.
- [186] Adjutant General's Office. Army individual test battery. manual of directions and scoring, 1944.

Bibliography

- [187] J Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643, 1935.
- [188] D. Wechsler. *WAIS-III Technische Handleiding*. Harcourt Test Publishers, 2005.
- [189] J Jolles, PJ Houx, MPJ Van Boxtel, and RWHM Ponds. The Maastricht aging study: Determinants of cognitive aging. *Maastricht: Neuropsych Publishers*, 1995.
- [190] Paul Ekman and W.V. Friesen. Pictures of facial affect. *Consulting Psychologists Press*, 1976.
- [191] Aurélie Funkiewiez, Maxime Bertoux, Leonardo Cruz de Souza, Richard Lévy, and Bruno Dubois. The SEA (social cognition and emotional assessment): A clinical neuropsychological tool for early diagnosis of frontal variant of frontotemporal lobar degeneration. *Neuropsychology*, 26(1):81, 2012.
- [192] Francesca Happé, Hiram Brownell, and Ellen Winner. Acquired ‘theory of mind’ impairments following stroke. *Cognition*, 70(3):211 – 240, 1999.
- [193] Alastair Compston and Alasdair Coles. Multiple sclerosis. *The Lancet*, 372(9648):1502 – 1517, 2008.
- [194] Massimo Filippi, Arnaud Charil, Marco Rovaris, Martina Absinta, and Maria Assunta Rocca. *Insights from magnetic resonance imaging*, pages 115–149. *Handbook of Clinical Neurology*. Elsevier B.V., 2014.
- [195] N. Bergsland, D. Horakova, M. G. Dwyer, O. Dolezal, Z. K. Seidl, M. Vaneckova, J. Krasensky, E. Havrdova, and Robert Zivadinov. Subcortical and cortical gray matter atrophy in a large sample of patients with clinically isolated syndrome and early relapsing-remitting multiple sclerosis. *American Journal of Neuroradiology*, 33(8):1573–1578, September 2012.
- [196] Jing Huang, Yaou Liu, Tengda Zhao, Ni Shu, Yunyun Duan, Zhuoqiong Ren, Zheng Sun, Zheng Liu, Hai Chen, Huiqing Dong, and Kuncheng Li. White matter microstructural alterations in clinically isolated syndrome and multiple sclerosis. *Journal of Clinical Neuroscience*, 53:27 – 33, 2018.
- [197] M M Schoonheim, J J G Geurts, O T Wiebenga, J C De Munck, C H Polman, C J Stam, F Barkhof, and A M Wink. Changes in functional network centrality underlie cognitive dysfunction and physical disability in multiple sclerosis. *Multiple Sclerosis Journal*, 20(8):1058–1065, 2014. PMID: 24347184.
- [198] Menno M Schoonheim, Veronica Popescu, Fernanda C Rueda Lopes, Oliver T Wiebenga, Hugo Vrenken, Linda Douw, Chris H Polman, Jeroen J G Geurts, and Frederik Barkhof. Subcortical atrophy and cognition: sex effects in multiple sclerosis. *Neurology*, 79(17):1754—1761, October 2012.
- [199] Chris H. Polman, Stephen C. Reingold, Brenda Banwell, Michel Clanet, Jeffrey A. Cohen, Massimo Filippi, Kazuo Fujihara, Eva Havrdova, Michael Hutchinson, Ludwig Kappos, Fred D. Lublin, Xavier Montalban, Paul O’Connor, Magnhild Sandberg-Wollheim, Alan J. Thompson, Emmanuelle Waubant, Brian Weinshenker, and Jerry S. Wolinsky. Diagnostic criteria for multiple sclerosis: 2010 revisions to the mcdonald criteria. *Annals of Neurology*, 69(2):292–302, 2011.
- [200] JF Kurtzke. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*, 33(11):1444—1452, November 1983.
- [201] SM Rao and the Cognitive Function Study Group of the National Multiple Sclerosis Society. A manual for the brief repeatable battery of neuropsychological tests in multiple sclerosis. *Milwaukee: Medical College of Wisconsin*, 1996, 1990.
- [202] M P Amato, E Portaccio, B Goretti, V Zipoli, L Ricchiuti, M F De Caro, F Patti, R Vecchio, S Sorbi, and M Trojano. The Rao’s brief repeatable battery and Stroop test: normative values with age, education and gender corrections in an Italian population. *Multiple Sclerosis Journal*, 12(6):787–793, 2006. PMID: 17263008.

- [203] Menno M. Schoonheim, Hanneke E. Hulst, Roemer B. Brandt, Myrte Strik, Alle Meije Wink, Bernard M.J. Uitdehaag, Frederik Barkhof, and Jeroen J.G. Geurts. Thalamus structure and function determine severity of cognitive impairment in multiple sclerosis. *Neurology*, 84(8):776–783, 2015.
- [204] F Verhage and JJ Van Der Werff. An analysis of variance based on the Groninger intelligence test scores. *Nederlands tijdschrift voor de psychologie en haar grensgebieden*, 19:497—509, 1964.
- [205] Anand J.C. Eijlers, Kim A. Meijer, Quinten van Geest, Jeroen J.G. Geurts, and Menno M. Schoonheim. Determinants of cognitive impairment in patients with multiple sclerosis with and without atrophy. *Radiology*, 288(2):544–551, 2018.
- [206] Martijn D. Steenwijk, Petra J.W. Pouwels, Marita Daams, Jan Willem van Dalen, Matthan W.A. Caan, Edo Richard, Frederik Barkhof, and Hugo Vrenken. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). *NeuroImage: Clinical*, 3:462 – 469, 2013.
- [207] Ferran Prados, Manuel Jorge Cardoso, Baris Kanber, Olga Ciccarelli, Raju Kapoor, Claudia A.M. Gandini Wheeler-Kingshott, and Sebastien Ourselin. A multi-time-point modality-agnostic patch-based method for lesion filling in multiple sclerosis. *NeuroImage*, 139:376 – 384, 2016.
- [208] M. J. Cardoso, M. Modat, R. Wolz, A. Melbourne, D. Cash, D. Rueckert, and S. Ourselin. Geodesic information flows: Spatially-variant graphs and their application to segmentation and fusion. *IEEE Transactions on Medical Imaging*, 34(9):1976–1988, 2015.
- [209] Arman Eshaghi, Ferran Prados, Wallace J. Brownlee, Daniel R. Altmann, Carmen Tur, M. Jorge Cardoso, Floriana De Angelis, Steven H. van de Pavert, Niamh Cawley, Nicola De Stefano, M. Laura Stromillo, Marco Battaglini, Serena Ruggieri, Claudio Gasperini, Massimo Filippi, Maria A. Rocca, Alex Rovira, Jaume Sastre-Garriga, Hugo Vrenken, Cyra E. Leurs, Joep Killestein, Lukas Pirpamer, Christian Enzinger, Sebastien Ourselin, Claudia A.M. Gandini Wheeler-Kingshott, Declan Chard, Alan J. Thompson, Daniel C. Alexander, Frederik Barkhof, Olga Ciccarelli, and on behalf of the MAGNIMS study group. Deep gray matter volume loss drives disability worsening in multiple sclerosis. *Annals of Neurology*, 83(2):210–222, 2018.
- [210] Matteo Pardini, Carole H Sudre, Ferran Prados, Özgür Yaldizli, Varun Sethi, Nils Muhlert, Rebecca S Samson, Steven H van de Pavert, M Jorge Cardoso, Sebastien Ourselin, Claudia A M Gandini Wheeler-Kingshott, David H Miller, and Declan T Chard. Relationship of grey and white matter abnormalities with distance from the surface of the brain in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(11):1212–1217, 2016.
- [211] Silvia Ingala, Linda Mazzai, Carole H. Sudre, Gemma Salvadó, Anna Brugulat-Serrat, Viktor Wottschel, Carles Falcon, Grégory Operto, Betty Tijms, Juan Domingo Gispert, José Luis Molinuevo, and Frederik Barkhof. The relation between APOE genotype and cerebral microbleeds in cognitively unimpaired middle- and old-aged individuals. *Neurobiology of Aging*, 95:104 – 114, 2020.
- [212] Carole H. Sudre, M. Jorge Cardoso, Sebastien Ourselin, and for the Alzheimer’s Disease Neuroimaging Initiative. Longitudinal segmentation of age-related white matter hyperintensities. *Medical Image Analysis*, 38:50 – 64, 2017.
- [213] Arno Klein and Jason Tourville. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in Neuroscience*, 6:171, 2012.
- [214] C.H. Sudre, B. Gomez Anson, I. Davagnanam, A. Schmitt, A.F. Mendelson, F. Prados, L. Smith, D. Atkinson, A.D. Hughes, N. Chaturvedi, M.J. Cardoso, F. Barkhof, H.R. Jaeger, and S. Ourselin. Bullseye’s representation of cerebral white matter hyperintensities. *Journal of Neuroradiology*, 45(2):114 – 122, 2018. 45e Congrès de la Société française de neuroradiologie 45th Congress of the French society of neuroradiology.
- [215] Benjamin De Leener, Samuel Kadoury, and Julien Cohen-Adad. Robust, accurate and fast automatic segmentation of the spinal cord. *NeuroImage*, 98:528 – 536, 2014.

Bibliography

- [216] C. Lukas, F. Prados, P. Valsasina, K. Parmar, I. Brouwer, B. Bellenberg, A. Rovira, D. Chard, L. Kappos, M. A. Rocca, M. Filippi, O. Ciccarelli, F. Barkhof, and H. Vrenken. Quantification of spinal cord atrophy in MS: which software, which vertebral level, spinal cord or brain MRI? a multi-centric, longitudinal comparison of three different volumetric approaches. *Multiple Sclerosis Journal*, 24:88–90, 10 2018.
- [217] Y Liu, C Lukas, MD Steenwijk, M Daams, A Versteeg, Y Duan, K Li, F Weiler, HK Hahn, MP Wattjes, F Barkhof, and H Vrenken. Multicenter validation of mean upper cervical cord area measurements from head 3D T1-weighted MR imaging in patients with multiple sclerosis. *AJNR. American journal of neuroradiology*, 37(4):749–754, April 2016.
- [218] Anand J.C. Eijlers, Kim A. Meijer, Thomas M. Wassenaar, Martijn D. Steenwijk, Bernard M.J. Uitdehaag, Frederik Barkhof, Alle M. Wink, Jeroen J.G. Geurts, and Menno M. Schoonheim. Increased default-mode network centrality in cognitively impaired multiple sclerosis patients. *Neurology*, 88(10):952–960, 2017.
- [219] Alle Meije Wink, Jan C de Munck, Ysbrand D van der Werf, Odile A van den Heuvel, and Frederik Barkhof. Fast eigenvector centrality mapping of voxel-wise connectivity in functional magnetic resonance imaging: implementation, validation, and interpretation. *Brain connectivity*, 2(5):265–274, 2012.
- [220] K.A. Meijer, Q. van Geest, A.J.C. Eijlers, J.J.G. Geurts, M.M. Schoonheim, and H.E. Hulst. Is impaired information processing speed a matter of structural or functional damage in MS? *NeuroImage: Clinical*, 20:844 – 850, 2018.
- [221] C Finke, J Schlichting, S Papazoglou, M Scheel, A Freing, C Soemmer, LM Pech, A Pajkert, C Pfüller, JT Wuerfel, CJ Ploner, F Paul, and AU Brandt. Altered basal ganglia functional connectivity in multiple sclerosis patients with fatigue. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 21(7):925–934, June 2015.
- [222] Domenico M. Mezzapesa, Maria A. Rocca, Mariaemma Rodegher, Giancarlo Comi, and Massimo Filippi. Functional cortical changes of the sensorimotor network are associated with clinical recovery in multiple sclerosis. *Human Brain Mapping*, 29(5):562–573, 2008.
- [223] Anand J C Eijlers, Quinten van Geest, Iris Dekker, Martijn D Steenwijk, Kim A Meijer, Hanneke E Hulst, Frederik Barkhof, Bernard M J Uitdehaag, Menno M Schoonheim, and Jeroen J G Geurts. Predicting cognitive decline in multiple sclerosis: a 5-year follow-up study. *Brain*, 141(9):2605–2618, 07 2018.
- [224] Katherine A Koenig, Ken E Sakaie, Mark J Lowe, Jian Lin, Lael Stone, Robert A Bermel, Erik B Beall, Stephen M Rao, Bruce D Trapp, and Micheal D Phillips. The relationship between cognitive function and high-resolution diffusion tensor MRI of the cingulum bundle in multiple sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 21(14):1794–1801, December 2015.
- [225] Daniel S. Reich, Seth A. Smith, Kathleen M. Zackowski, Eliza M. Gordon-Lipkin, Craig K. Jones, Jonathan A.D. Farrell, Susumu Mori, Peter C.M. van Zijl, and Peter A. Calabresi. Multiparametric magnetic resonance imaging analysis of the corticospinal tract in multiple sclerosis. *NeuroImage*, 38(2):271 – 279, 2007.
- [226] Katrin Parmar, Christine Stadelmann, Maria A. Rocca, Dawn Langdon, Egidio D’Angelo, Marcus D’Souza, Jessica Burggraaff, Christiane Wegner, Jaume Sastre-Garriga, Alonso Barrantes-Freer, Jonas Dorn, Bernard M.J. Uitdehaag, Xavier Montalban, Jens Wuerfel, Christian Enzinger, Alex Rovira, Mar Tintore, Massimo Filippi, Ludwig Kappos, and Till Sprenger. The role of the cerebellum in multiple sclerosis—150 years after Charcot. *Neuroscience & Biobehavioral Reviews*, 89:85 – 98, 2018.
- [227] Bertrand Audoin, Wafaa Zaaraoui, Françoise Reuter, Audrey Rico, Irina Malikova, Sylviane Confort-Gouny, Patrick J Cozzone, Jean Pelletier, and Jean-Philippe Ranjeva. Atrophy mainly affects the limbic system and the deep grey matter at the first stage of multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(6):690–695, 2010.

- [228] R G Henry, M Shieh, D T Okuda, A Evangelista, M L Gorno-Tempini, and D Pelletier. Regional grey matter atrophy in clinically isolated syndromes at presentation. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(11):1236–1244, 2008.
- [229] Viola Biberacher, Christine C Boucard, Paul Schmidt, Christina Engl, Dorothea Buck, Achim Berthele, Muna-Miriam Hoshi, Claus Zimmer, Bernhard Hemmer, and Mark Mühlau. Atrophy and structural variability of the upper cervical cord in early multiple sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 21(7):875–884, June 2015.
- [230] Kim A Meijer, Nils Muhlert, Mara Cercignani, Varun Sethi, Maria A Ron, Alan J Thompson, David H Miller, Declan Chard, Jeroen Jg Geurts, and Olga Ciccarelli. White matter tract abnormalities are associated with cognitive dysfunction in secondary progressive multiple sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 22(11):1429–1437, October 2016.
- [231] Marita Daams, Florian Weiler, Martijn D Steenwijk, Horst K Hahn, Jeroen Jg Geurts, Hugo Vrenken, Ronald A van Schijndel, Lisanne J Balk, Prejaas K Tewarie, Jan-Mendelt Tillema, Joep Killestein, Bernard Mj Uitdehaag, and Frederik Barkhof. Mean upper cervical cord area (MUCCA) measurement in long-standing multiple sclerosis: relation to brain findings and clinical disability. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 20(14):1860–1865, December 2014.
- [232] Carsten Lukas, Dirk L Knol, Madeleine H Sombekke, Barbara Bellenberg, Horst K Hahn, Veronica Popescu, Katrin Weier, Ernst W Radue, Achim Gass, Ludwig Kappos, Yvonne Naegelin, Bernard M J Uitdehaag, Jeroen J G Geurts, Frederik Barkhof, and Hugo Vrenken. Cervical spinal cord volume loss is related to clinical disability progression in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 86(4):410–418, 2015.
- [233] Prejaas Tewarie, Menno M. Schoonheim, Daphne I. Schouten, Chris H. Polman, Lisanne J. Balk, Bernard M.J. Uitdehaag, Jeroen J.G. Geurts, Arjan Hillebrand, Frederik Barkhof, and Cornelis J. Stam. Functional brain networks: Linking thalamic atrophy to clinical disability in multiple sclerosis, a multimodal fMRI and MEG study. *Human Brain Mapping*, 36(2):603–618, 2015.
- [234] Francesca Rossi, Antonio Giorgio, Marco Battaglini, Maria Laura Stromillo, Emilio Portaccio, Benedetta Goretti, Antonio Federico, Bahia Hakiki, Maria Pia Amato, and Nicola De Stefano. Relevance of brain lesion location to cognition in relapsing multiple sclerosis. *PLOS ONE*, 7(11):1–7, 11 2012.
- [235] Antonio Giorgio, Marco Battaglini, Maria Assunta Rocca, Alessandro De Leucio, Martina Absinta, Ronald van Schijndel, Alex Rovira, Mar Tintore, Declan Chard, Olga Ciccarelli, Christian Enzinger, Claudio Gasperini, Jette Frederiksen, Massimo Filippi, Frederik Barkhof, Nicola De Stefano, and MAGNIMS Study Group. Location of brain lesions predicts conversion of clinically isolated syndromes to multiple sclerosis. *Neurology*, 80(3):234–241, 2013.
- [236] Yaou Liu, Teng Xie, Yong He, Yunyun Duan, Jing Huang, Zhuoqiong Ren, Gaolang Gong, Jun Wang, Jing Ye, Huiqing Dong, Helmut Butzkueven, Fu-Dong Shi, Ni Shu, and Kuncheng Li. Cortical thinning correlates with cognitive change in multiple sclerosis but not in neuromyelitis optica. *European radiology*, 24(9):2334–2343, September 2014.
- [237] Alan J Thompson, Brenda L Banwell, Frederik Barkhof, William M Carroll, Timothy Coetzee, Giancarlo Comi, Jorge Correale, Franz Fazekas, Massimo Filippi, Mark S Freedman, Kazuo Fujihara, Steven L Galetta, Hans Peter Hartung, Ludwig Kappos, Fred D Lublin, Ruth Ann Marrie, Aaron E Miller, David H Miller, Xavier Montalban, Ellen M Mowry, Per Soelberg Sorensen, Mar Tintoré, Anthony L Traboulsee, Maria Trojano, Bernard M J Uitdehaag, Sandra Vukusic, Emmanuelle Waubant, Brian G Weinshenker, Stephen C Reingold, and Jeffrey A Cohen. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, 17(2):162 – 173, 2018.
- [238] Madeleine H. Sombekke, Mike P. Wattjes, Lisanne J. Balk, Jessica M. Nielsen, Hugo Vrenken, Bernard M.J. Uitdehaag, Chris H. Polman, and Frederik Barkhof. Spinal cord lesions in patients with clinically isolated syndrome. *Neurology*, 80(1):69–75, 2013.

Bibliography

- [239] Giulio Disanto, Christian Barro, Pascal Benkert, Yvonne Naegelin, Sabine Schädelin, Antonella Giardiello, Chiara Zecca, Kaj Blennow, Henrik Zetterberg, David Leppert, Ludwig Kappos, Claudio Gobbi, Jens Kuhle, and the Swiss Multiple Sclerosis Cohort Study Group. Serum neurofilament light: A biomarker of neuronal damage in multiple sclerosis. *Annals of Neurology*, 81(6):857–870, 2017.
- [240] Dev Mehta, Robert Jackson, Gaurav Paul, Jiong Shi, and Marwan Sabbagh. Why do trials for Alzheimer’s disease drugs keep failing? A discontinued drug perspective for 2010-2015. *Expert opinion on investigational drugs*, 26(6):735–739, 2017.
- [241] Rachael I Scahill, Jonathan M Schott, John M Stevens, Martin N Rossor, and Nick C Fox. Mapping the evolution of regional atrophy in Alzheimer’s disease: unbiased analysis of fluid-registered serial MRI. *Proceedings of the National Academy of Sciences*, 99(7):4703–4707, 2002.
- [242] Mert R. Sabuncu, Rahul S. Desikan, Jorge Sepulcre, Boon Thye T. Yeo, Hesheng Liu, Nicholas J. Schmansky, Martin Reuter, Michael W. Weiner, Randy L. Buckner, Reisa A. Sperling, Bruce Fischl, and for the Alzheimer’s Disease Neuroimaging Initiative. The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Archives of neurology*, 68(8):1040–1048, 2011.
- [243] Rachelle S Doody, Valory Pavlik, Paul Massman, Susan Rountree, Eveleen Darby, and Wenyaw Chan. Erratum to: Predicting progression of Alzheimer’s disease. *Alzheimer’s research & therapy*, 2(3):14, 2010.
- [244] Stefan Klöppel, Cynthia M Stonnington, Carlton Chu, Bogdan Draganski, Rachael I Scahill, Jonathan D Rohrer, Nick C Fox, Clifford R Jack Jr, John Ashburner, and Richard SJ Frackowiak. Automatic classification of MR scans in Alzheimer’s disease. *Brain*, 131(3):681–689, 2008.
- [245] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, Dinggang Shen, and Alzheimer’s Disease Neuroimaging Initiative. Multimodal classification of Alzheimer’s disease and mild cognitive impairment. *Neuroimage*, 55(3):856–867, 2011.
- [246] Jonathan Young, Marc Modat, Manuel J Cardoso, Alex Mendelson, Dave Cash, Sebastien Ourselin, and Alzheimer’s Disease Neuroimaging Initiative. Accurate multimodal probabilistic prediction of conversion to Alzheimer’s disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2:735–745, 2013.
- [247] Jussi Mattila, Juha Koikkalainen, Arho Virkki, Anja Simonsen, Mark van Gils, Gunhild Waldemar, Hilikka Soininen, Jyrki Lötjönen, and Alzheimer’s Disease Neuroimaging Initiative. A disease state fingerprint for evaluation of Alzheimer’s disease. *Journal of Alzheimer’s Disease*, 27(1):163–176, 2011.
- [248] Jean-Baptiste Schiratti, Stéphanie Allasonnière, Olivier Colliot, and Stanley Durrleman. A bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. *Journal of Machine Learning Research*, 18(133):1–33, 2017.
- [249] Murat Bilgel, Jerry L Prince, Dean F Wong, Susan M Resnick, and Bruno M Jernigan. A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging. *Neuroimage*, 134:658–670, 2016.
- [250] Răzvan V. Marinescu, Arman Eshaghi, Marco Lorenzi, Alexandra L. Young, Neil P. Oxtoby, Sara Garbarino, Sebastian J. Crutch, and Daniel C. Alexander. DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders. *NeuroImage*, 192:166 – 177, 2019.
- [251] Igor Koval, Alexandre Bône, Maxime Louis, Simona Bottani, Arnaud Marcoux, Jorge Samper-Gonzalez, Ninon Burgos, Benjamin CHARLIER, Anne Bertrand, Stéphane Epelbaum, Olivier Colliot, Stéphanie Allasonnière, and Stanley Durrleman. Simulating Alzheimer’s disease progression with person-alised digital brain models. *Inria preprint*, 2018.
- [252] Ashish Raj, Amy Kuceyeski, and Michael Weiner. A network diffusion model of disease progression in dementia. *Neuron*, 73(6):1204–1215, 2012.

- [253] Juan Zhou, Efstathios D Gennatas, Joel H Kramer, Bruce L Miller, and William W Seeley. Predicting regional neurodegeneration from the healthy brain functional connectome. *Neuron*, 73(6):1216–1227, 2012.
- [254] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, Tal Arbel, Hrvoje Bogunović, A.P. Bradley, Aaron Carass, C. Feldmann, Alejandro Frangi, P.M. Full, Berbke van Ginneken, A. Hanbury, Michal Honauer, K. and Kozubek, Bennett Landman, K. März, O. Maier, K. Maier-Hein, B.H. Menze, H. Müller, P.F. Neher, Wiro Niessen, N. Rajpoot, G.C. Sharp, K. Sirinukunwattana, S. Speidel, C. Stock, D. Stoyanov, A.A. Taha, F. van der Sommen, C.-W. Wang, M.-A. Weber, G. Zheng, P. Jannin, and A. Kopp-Schneider. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications*, 9(1):5217, 2018.
- [255] K. Murphy, B. van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G. E. Christensen, V. Garcia, T. Vercauteren, N. Ayache, O. Commowick, G. Malandain, B. Glocker, N. Paragios, N. Navab, V. Gorbunova, J. Sporring, M. de Bruijne, X. Han, M. P. Heinrich, J. A. Schnabel, M. Jenkinson, C. Lorenz, M. Modat, J. R. McClelland, S. Ourselin, S. E. A. Muenzing, M. A. Viergever, D. De Nigris, D. L. Collins, T. Arbel, M. Peroni, R. Li, G. C. Sharp, A. Schmidt-Richberg, J. Ehrhardt, R. Werner, D. Smeets, D. Loeckx, G. Song, N. Tustison, B. Avants, J. C. Gee, M. Staring, S. Klein, B. C. Stoel, M. Urschler, M. Werlberger, J. Vandemeulebroucke, S. Rit, D. Sarrut, and J. P. W. Pluim. Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge. *IEEE Transactions on Medical Imaging*, 30(11):1901–1920, 2011.
- [256] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.
- [257] Klaus H Maier-Hein, Peter F. Neher, Christophe Houde, Marc-Alexandre Côté, Eleftherios Garyfallidis, Jidan Zhong, Maxime Chamberland, Fang-Chen Yeh, Ying-Chia Lin, Qing Ji, Wilburn E. Reddick, John O. Glass, David Qixiang Chen, Yuanjing Feng, Chengfeng Gao, Ye Wu, Jieyan Ma, H. Renjie, Qiang Li, Carl-Fredrik Westin, Samuel Deslauriers-Gauthier, J. Omar Ocegueda González, Michael Paquette, Samuel St-Jean, Gabriel Girard, François Rheault, Jasmeen Sidhu, Chantal M. W. Tax, Fenghua Guo, Hamed Y. Mesri, Szabolcs Dávid, Martijn Froeling, Anneriet M. Heemskerk, Alexander Leemans, Arnaud Boré, Basile Pinsard, Christophe Bedetti, Matthieu Desrosiers, Bram Brambati, Julien Doyon, Alessia Sarica, Roberta Vasta, Antonio Cerasa, Aldo Quattrone, Jason Yeatman, Ali R. Khan, Wes Hodges, Simon Alexander, David Romascano, Muhamed Barakovic, Anna Auría, Oscar Esteban, Alia Lemkaddem, Jean-Philippe Thiran, H. Ertan Cetin Gul, Benjamin L. Odry, Boris Maillhé, Mariappan S. Nadar, Fabrizio M Pizzagalli, Gautam Prasad, Julio E. Villalon-Reina, Justin Galvis, Paul M. Thompson, Francisco De Santiago Requejo, Pedro Luque Laguna, Luis Miguel Lacerda, Rachel Barrett, Flavio Dell’Acqua, Marco Catani, Laurent Petit, Emmanuel Caruyer, Alessandro Daducci, Tim Dyrby, Tim Holland-Letz, Claus C. Hilgetag, Bram Stieltjes, and Maxime Descoteaux. The challenge of mapping the human connectome based on diffusion tractography. *Nature communications*, 8(1):1349, 2017.
- [258] Esther E. Bron, Marion Smits, Wiesje M. van der Flier, Hugo Vrenken, Frederik Barkhof, Philip Scheltens, Janne M. Papma, Rebecca M.E. Steketee, Carolina Méndez Orellana, Rozanna Meijboom, Madalena Pinto, Joana R. Meireles, Carolina Garrett, António J. Bastos-Leite, Ahmed Abdulkadir, Olaf Ronneberger, Nicola Amoroso, Roberto Bellotti, David Cárdenas-Peña, Andrés M. Álvarez Meza, Chester V. Dolph, Khan M. Iftekharuddin, Simon F. Eskildsen, Pierrick Coupé, Vladimir S. Fonov, Katja Franke, Christian Gaser, Christian Ledig, Ricardo Guerrero, Tong Tong, Katherine R. Gray, Elaheh Moradi, Jussi Tohka, Alexandre Routier, Stanley Durrleman, Alessia Sarica, Giuseppe Di Fatta, Francesco Sensi, Andrea Chincarini, Garry M. Smith, Zhivko V. Stoyanov, Lauge Sørensen,

Bibliography

- Mads Nielsen, Sabina Tangaro, Paolo Inglese, Christian Wachinger, Martin Reuter, John C. van Swieten, Wiro J. Niessen, and Stefan Klein. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. *NeuroImage*, 111:562 – 579, 2015.
- [259] Isabella Castiglioni, Christian Salvatore, Javier Ramírez, and Juan Manuel Górriz. Machine-learning neuroimaging challenge for automated diagnosis of mild cognitive impairment: Lessons learnt. *Journal of neuroscience methods*, 302:10, 2018.
- [260] Michael W. Weiner, Dallas P. Veitch, Paul S. Aisen, Laurel A. Beckett, Nigel J. Cairns, Robert C. Green, Danielle Harvey, Clifford R. Jack, William Jagust, John C. Morris, Ronald C. Petersen, Andrew J. Saykin, Leslie M. Shaw, Arthur W. Toga, and John Q. Trojanowski. Recent publications from the Alzheimer’s Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer’s & Dementia*, 13(4):e1–e85, 2017.
- [261] Genevra I. Allen, Nicola Amoroso, Catalina Anghel, Venkat Balagurusamy, Christopher J. Bare, Derek Beaton, Roberto Bellotti, David A. Bennett, Kevin L. Boehme, Paul C. Boutros, Laura Caberlotto, Cristian Caloian, Frederick Campbell, Elias Chaibub Neto, Yu-Chuan Chang, Beibei Chen, Chien-Yu Chen, Ting-Ying Chien, Tim Clark, Sudeshna Das, Christos Davatzikos, Jieyao Deng, Donna Dillenberger, Richard J.B. Dobson, Qilin Dong, Jimit Doshi, Denise Duma, Rosangela Errico, Guray Erus, Evan Everett, David W. Fardo, Stephen H. Friend, Holger Fröhlich, Jessica Gan, Peter St George-Hyslop, Satrajit S. Ghosh, Enrico Glaab, Robert C. Green, Yuanfang Guan, Ming-Yi Hong, Chao Huang, Jinseub Hwang, Joseph Ibrahim, Paolo Inglese, Anandhi Iyappan, Qijia Jiang, Yuriko Katsumata, John S.K. Kauwe, Arno Klein, Dehan Kong, Roland Krause, Emilie Lalonde, Mario Lauria, Eunjee Lee, Xihui Lin, Zhandong Liu, Julie Livingstone, Benjamin A. Logsdon, Simon Lovestone, Tsung-wei Ma, Ashutosh Malhotra, Lara M. Mangravite, Taylor J. Maxwell, Emily Merrill, John Nagorski, Aishwarya Namasivayam, Manjari Narayan, Mufassra Naz, Stephen J. Newhouse, Thea C. Norman, Ramil N. Nurtdinov, Yen-Jen Oyang, Yudi Pawitan, Shengwen Peng, Mette A. Peters, Stephen R. Piccolo, Paurush Praveen, Corrado Priami, Veronica Y. Sabelnykova, Philipp Senger, Xia Shen, Andrew Simmons, Aristeidis Sotiras, Gustavo Stolovitzky, Sabina Tangaro, Andrea Tateo, Yi-An Tung, Nicholas J. Tustison, Erdem Varol, George Vradenburg, Michael W. Weiner, Guanghua Xiao, Lei Xie, Yang Xie, Jia Xu, Hojin Yang, Xiaowei Zhan, Yunyun Zhou, Fan Zhu, Hongtu Zhu, Shanfeng Zhu, and Alzheimer’s Disease Neuroimaging Initiative. Crowdsourced estimation of cognitive decline and resilience in Alzheimer’s disease. *Alzheimer’s & Dementia*, 12(6):645–653, 2016.
- [262] Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, and Alzheimer’s Disease Neuroimaging Initiative. Machine learning framework for early MRI-based Alzheimer’s conversion prediction in MCI subjects. *Neuroimage*, 104:398–412, 2015.
- [263] Xiaojing Long, Lifang Chen, Chunxiang Jiang, Lijuan Zhang, and Alzheimer’s Disease Neuroimaging Initiative. Prediction and classification of Alzheimer disease based on quantification of MRI deformation. *PLoS one*, 12(3):e0173372, 2017.
- [264] Joseph H Grochowalski, Ying Liu, and Karen L Siedlecki. Examining the reliability of ADAS-Cog change scores. *Aging, Neuropsychology, and Cognition*, 23(5):513–529, 2016.
- [265] Thomas G Beach, Sarah E Monsell, Leslie E Phillips, and Walter Kukull. Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer disease centers, 2005–2010. *Journal of neuropathology and experimental neurology*, 71(4):266–273, 2012.
- [266] Xiao Han, Jorge Jovicich, David Salat, Andre van der Kouwe, Brian Quinn, Silvester Czanner, Evelina Busa, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Paul Maguire, Diana Rosas, Nikos Makris, Anders Dale, Bradford Dickerson, and Bruce Fischl. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage*, 32(1):180–194, 2006.
- [267] S Kiebel and AP Holmes. *The general linear model*, volume 8. chapter, 2007.

- [268] Martin Reuter, Nicholas J Schmansky, H Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4):1402–1418, 2012.
- [269] Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.
- [270] Gautam Prasad, Talia M Nir, Arthur W Toga, and Paul M Thompson. Tractography density and network measures in Alzheimer’s disease. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 692–695. IEEE, 2013.
- [271] Kenichi Oishi, Andreia Faria, Hangyi Jiang, Xin Li, Kazi Akhter, Jiangyang Zhang, John T. Hsu, Michael I. Miller, Peter C.M. van Zijl, Marilyn Albert, Constantine G. Lyketsos, Roger Woods, Arthur W. Toga, G. Bruce Pike, Pedro Rosa-Neto, Alan Evans, John Mazziotta, and Susumu Mori. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and Alzheimer’s disease participants. *Neuroimage*, 46(2):486–499, 2009.
- [272] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
- [273] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2):103–134, May 2000.
- [274] Ulf Brefeld, Peter Geibel, and Fritz Wysotzki. Support vector machines with example dependent costs. In *Machine Learning: ECML 2003*, pages 23–34, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.
- [275] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 1999.
- [276] Ulf Brefeld and Tobias Scheffer. Co-EM support vector learning. In *Proceedings of the Twenty-first International Conference on Machine Learning, ICML ’04*, pages 16–, New York, NY, USA, 2004. ACM.
- [277] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014.
- [278] Matthew D Adadelta Zeiler. An adaptive learning rate method. arxiv preprint. *arXiv preprint arXiv:1212.5701*, 2012.
- [279] Global action plan on the public health response to dementia 2017-2025. *World Health Organization*, 2017.
- [280] Alexandra J Weigand, Katherine J Bangen, Kelsey R Thomas, Lisa Delano-Wood, Paul E Gilbert, Adam M Brickman, Mark W Bondi, and Alzheimer’s Disease Neuroimaging Initiative. Is tau in the absence of amyloid on the Alzheimer’s continuum?: A study of discordant PET positivity. *Brain Communications*, 2(1), 12 2019.
- [281] Melissa E Murray, Neill R Graff-Radford, Owen A Ross, Ronald C Petersen, Ranjan Duara, and Dennis W Dickson. Neuropathologically defined subtypes of Alzheimer’s disease with distinct clinical characteristics: a retrospective study. *The Lancet Neurology*, 10(9):785 – 796, 2011.
- [282] Rhoda Au, Ryan J. Piers, and Lee Lancashire. Back to the future: Alzheimer’s disease heterogeneity revisited. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(3):368 – 370, 2015.
- [283] C. Patterson. *World Alzheimer Report 2018*. Alzheimer’s Disease International, 2018.
- [284] Sven J van der Lee, Frank J Wolters, M Kamran Ikram, Albert Hofman, M Arfan Ikram, Najaf Amin, and Cornelia M van Duijn. The effect of APOE and other common genetic variants on the onset of Alzheimer’s disease and dementia: a community-based cohort study. *The Lancet Neurology*, 17(5):434 – 444, 2018.

Bibliography

- [285] A.M. Saunders, W.J. Strittmatter, D. Schmechel, P.H. St. George-Hyslop, M.A. Pericak-Vance, S.H. Joo, B.L. Rosi, J.F. Gusella, D.R. Crapper-Mac Lachlan, M.J. Alberts, C. Hulette, B. Crain, D. Goldgaber, and A.D. Roses. Association of apolipoprotein E allele $\epsilon 4$ with late-onset familial and sporadic Alzheimer's disease. *Neurology*, 43(8):1467–1472, 1993.
- [286] Jungsu Kim, Jacob M. Basak, and David M. Holtzman. The role of apolipoprotein E in Alzheimer's disease. *Neuron*, 63(3):287 – 303, 2009.
- [287] Emmanuelle Genin, Didier Hannequin, David Wallon, Kristel Sleegers, Mikko Hiltunen, Onofre Combarros, Maria Bullido, Sebastiaan Engelborghs, De Paul, Claudine Berr, Florence Pasquier, Bruno Dubois, Gloria Tognoni, Nathalie Fievet, Nathalie Brouwers, Karolien Bettens, Beatrice Arosio, Eliecer Coto, Maria Zompo, and Dominique Campion. APOE and Alzheimer disease: A major gene with semi-dominant inheritance. *Molecular psychiatry*, 16:903–7, 05 2011.
- [288] Sandra Weintraub, Merilee Teylan, Benjamin Rader, Kwun C.G. Chan, Mark Bollenbeck, Walter A. Kukull, Christina Coventry, Emily Rogalski, Eileen Bigio, and M.-Marsel Mesulam. APOE is a correlate of phenotypic heterogeneity in Alzheimer disease in a national cohort. *Neurology*, 2019.
- [289] Clifford R. Jack Jr., Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer L. Whitwell, Chadwick Ward, Anders M. Dale, Joel P. Felmlee, Jeffrey L. Gunter, Derek L.G. Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, Charles S. DeCarli, Gunnar Krueger, Heidi A. Ward, Gregory J. Metzger, Katherine T. Scott, Richard Mallozzi, Daniel Blezek, Joshua Levy, Josef P. Debbins, Adam S. Fleisher, Marilyn Albert, Robert Green, George Bartzokis, Gary Glover, John Mugler, and Michael W. Weiner. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.
- [290] Clifford R. Jack Jr., Josephine Barnes, Matt A. Bernstein, Bret J. Borowski, James Brewer, Shona Clegg, Anders M. Dale, Owen Carmichael, Christopher Ching, Charles DeCarli, Rahul S. Desikan, Christine Fennema-Notestine, Anders M. Fjell, Evan Fletcher, Nick C. Fox, Jeff Gunter, Boris A. Gutman, Dominic Holland, Xue Hua, Philip Insel, Kejal Kantarci, Ron J. Killiany, Gunnar Krueger, Kelvin K. Leung, Scott Mackin, Pauline Maillard, Ian B. Malone, Niklas Mattsson, Linda McEvoy, Marc Modat, Susanne Mueller, Rachel Nosheny, Sebastien Ourselin, Norbert Schuff, Matthew L. Senjem, Alix Simonson, Paul M. Thompson, Dan Rettmann, Prashanthi Vemuri, Kristine Walhovd, Yansong Zhao, Samantha Zuk, and Michael Weiner. Magnetic resonance imaging in Alzheimer's Disease Neuroimaging Initiative 2. *Alzheimer's & Dementia*, 11(7):740–756, 2015.
- [291] Damiano Archetti, Silvia Ingala, Vikram Venkatraghavan, Viktor Wottschel, Alexandra L. Young, Maura Bellio, Esther E. Bron, Stefan Klein, Frederik Barkhof, Daniel C. Alexander, Neil P. Oxtoby, Giovanni B. Frisoni, and Alberto Redolfi. Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease. *NeuroImage: Clinical*, 24:101954, 2019.
- [292] Annika Thorsell, Maria Bjerke, Johan Gobom, Eva Brunhage, Eugeen Vanmechelen, Niels Andreasen, Oskar Hansson, Lennart Minthon, Henrik Zetterberg, and Kaj Blennow. Neurogranin in cerebrospinal fluid as a marker of synaptic degeneration in Alzheimer's disease. *Brain Research*, 1362:13 – 22, 2010.
- [293] Mei Jin, Li Cao, and Yan-ping Dai. Role of neurofilament light chain as a potential biomarker for Alzheimer's disease: A correlative meta-analysis. *Frontiers in Aging Neuroscience*, 11:254, 2019.
- [294] Frank de Wolf, Mohsen Ghanbari, Silvan Licher, Kevin McRae-McKee, Luuk Gras, Gerrit Jan Weverling, Paulien Wermeling, Sanaz Sedaghat, M Kamran Ikram, Reem Waziry, Wouter Koudstaal, Jaco Klap, Stefan Kostense, Albert Hofman, Roy Anderson, Jaap Goudsmit, and M Arfan Ikram. Plasma tau, neurofilament light chain and amyloid- β levels and risk of dementia: A population-based cohort study. *Brain*, 143(4):1220–1232, 03 2020.

- [295] Annika Olsson, Hugo Vanderstichele, Niels Andreasen, Geert De Meyer, Anders Wallin, Bjorn Holmberg, Lars Rosengren, Eugene Vanmechelen, and Kaj Blennow. Simultaneous measurement of β -Amyloid_{1–42}, total tau, and phosphorylated tau (Thr181) in cerebrospinal fluid by the xMAP technology. *Clinical Chemistry*, 51(2):336–345, 02 2005.
- [296] Niklas Mattsson, Ulf Andreasson, Henrik Zetterberg, Kaj Blennow, and for the Alzheimer’s Disease Neuroimaging Initiative. Association of plasma neurofilament light with neurodegeneration in patients with Alzheimer disease. *JAMA Neurology*, 74(5):557–566, 05 2017.
- [297] Erik Portelius, Henrik Zetterberg, Tobias Skillback, Ulrika Törnqvist, Ulf Andreasson, John Q. Trojanowski, Michael W. Weiner, Leslie M. Shaw, Niklas Mattsson, and for the Alzheimer’s Disease Neuroimaging Initiative Blennow, Kaj. Cerebrospinal fluid neurogranin: relation to cognition and neurodegeneration in Alzheimer’s disease. *Brain*, 138(11):3373–3385, 09 2015.
- [298] Nicholas J Ashton, Antoine Leuzy, Yau Mun Lim, Claire Troakes, Tibor Hortobagyi, Kina Hoglund, Dag Aarsland, Simon Lovestone, Michael Scholl, Kaj Blennow, Henrik Zetterberg, and Abdul Hye. Increased plasma neurofilament light chain concentration correlates with severity of post-mortem neurofibrillary tangle pathology and neurodegeneration. *Acta neuropathologica communications*, 7(1):5, January 2019.
- [299] Santosh Jadhav, Veronika Cubinkova, Ivana Zimova, Veronika Brezovakova, Aladar Madari, Viera Cigankova, and Norbert Zilka. Tau-mediated synaptic damage in Alzheimer’s disease. *Translational neuroscience*, 6(1):214–226, 2015.
- [300] Willem Huijbers, Elizabeth C Mormino, Sarah E Wigman, Andrew M Ward, Patrizia Vannini, Donald G McLaren, J Alex Becker, Aaron P Schultz, Trey Hedden, Keith A Johnson, and Reisa A Sperling. Amyloid deposition is linked to aberrant entorhinal activity among cognitively normal older adults. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 34(15):5200–5210, April 2014.
- [301] Sean M Nestor, Raul Rupsingh, Michael Borrie, Matthew Smith, Vittorio Accomazzi, Jennie L Wells, Jennifer Fogarty, Robert Bartha, and Alzheimer’s Disease Neuroimaging Initiative. Ventricular enlargement as a possible measure of Alzheimer’s disease progression validated using the Alzheimer’s disease neuroimaging initiative database. *Brain*, 131(9):2443–2454, September 2008.
- [302] Daniel Ferreira, Chloë Verhagen, Juan A. Hernández-cabrera, Lena Cavallin, Chun-jie Guo, Urban Ekman, J. Muehlboeck, Andrew Simmons, José Barroso, Lars-olof Wahlund, and Eric Westman. Distinct subtypes of Alzheimer’s disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Scientific Reports*, 7:46263, 04 2017.
- [303] Vikram Venkatraghavan, Florian Dubost, Esther E. Bron, Wiro J. Niessen, Marleen de Bruijne, and Stefan Klein. Event-based modeling with high-dimensional imaging biomarkers for estimating spatial progression of dementia. In *Information Processing in Medical Imaging - 26th International Conference, IPMI 2019, Proceedings*, volume 11492 of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 169–180. Springer, 2019.
- [304] Clive Holmes. Genotype and phenotype in Alzheimer’s disease. *British Journal of Psychiatry*, 180(2):131–134, 2002.
- [305] Joanne Ryan, Peter Fransquet, Jo Wrigglesworth, and Paul Lacaze. Phenotypic heterogeneity in dementia: A challenge for epidemiology and biomarker studies. *Frontiers in Public Health*, 6:181, 2018.
- [306] Gayatri Devi and Philip Scheltens. Heterogeneity of Alzheimer’s disease: consequence for drug trials? *Alzheimer’s research & therapy*, 10(1):122, December 2018.
- [307] Reisa A Sperling, Jason Karlawish, and Keith A Johnson. Preclinical Alzheimer disease: The challenges ahead. *Nature Reviews. Neurology*, 9(1):54–58, January 2013.

- [308] Razvan V. Marinescu, Neil P. Oxtoby, Alexandra L. Young, Esther E. Bron, Arthur W. Toga, Michael W. Weiner, Frederik Barkhof, Nick C. Fox, Arman Eshaghi, Tina Toni, Marcin Salaturski, Veronika Lunina, Manon Ansart, Stanley Durrleman, Pascal Lu, Samuel Iddi, Dan Li, Wesley K. Thompson, Michael C. Donohue, Aviv Nahon, Yarden Levy, Dan Halbersberg, Mariya Cohen, Huiling Liao, Tengfei Li, Kaixian Yu, Hongtu Zhu, Jose G. Tamez-Pena, Aya Ismail, Timothy Wood, Hector Corrada Bravo, Minh Nguyen, Nanbo Sun, Jiashi Feng, B. T. Thomas Yeo, Gang Chen, Ke Qi, Shiyang Chen, Deqiang Qiu, Ionut Buciuman, Alex Kelner, Raluca Pop, Denisa Rimocsa, Mostafa M. Ghazi, Mads Nielsen, Sebastien Ourselin, Lauge Sorensen, Vikram Venkatraghavan, Keli Liu, Christina Rabe, Paul Manser, Steven M. Hill, James Howlett, Zhiyue Huang, Steven Kiddle, Sach Mukherjee, Anaïs Rouanet, Bernd Taschler, Brian D. M. Tom, Simon R. White, Noel Faux, Suman Sedai, Javier de Velasco Oriol, Edgar E. V. Clemente, Karol Estrada, Leon Aksman, Andre Altmann, Cynthia M. Stonnington, Yalin Wang, Jianfeng Wu, Vivek Devadas, Clementine Fourier, Lars Lau Raket, Aristeidis Sotiras, Guray Erus, Jimit Doshi, Christos Davatzikos, Jacob Vogel, Andrew Doyle, Angela Tam, Alex Diaz-Papkovich, Emmanuel Jammeh, Igor Koval, Paul Moore, Terry J. Lyons, John Gallacher, Jussi Tohka, Robert Cizek, Bruno Jedynak, Kruti Pandya, Murat Bilgel, William Engels, Joseph Cole, Polina Golland, Stefan Klein, and Daniel C. Alexander. The Alzheimer’s disease prediction of longitudinal evolution (TADPOLE) challenge: Results after 1 year follow-up, 2020.
- [309] S. Kloeppel, J. Peter, A. Ludl, A. Pilatus, S. Maier, I. Mader, B. Heimbach, L. Frings, K. Egger, J. Dukart, M.I. Schroeter, R. Perneczky, P. Haussermann, W. Vach, H. Urbach, S. Teipel, M. Huell, and A. Abdulkadir. Applying automated MR-based diagnostic methods to the memory clinic: A prospective study. *Journal of Alzheimers Disease*, 47(4):939–954, Aug 2015.
- [310] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR, 2018.
- [311] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable AI systems for the medical domain?, 2017.
- [312] Vikram Venkatraghavan, Stefan Klein, Lana Fani, Leontine S. Ham, Henri Vrooman, M. Kamran Ikram, Wiro J. Niessen, Esther E. Bron, and ADNI. Analyzing the effect of APOE on Alzheimer’s disease progression using an event-based model for stratified populations. *arXiv*, 2020.
- [313] M Arfan Ikram, Guy Brusselle, Mohsen Ghanbari, André Goedegebure, M Kamran Ikram, Maryam Kavousi, Brenda C T Kieboom, Caroline C W Klaver, Robert J de Knegt, Annemarie I Luik, Tamar E C Nijsten, Robin P Peeters, Frank J A van Rooij, Bruno H Stricker, André G Uitterlinden, Meike W Vernooij, and Trudy Voortman. Objectives, design and main findings until 2020 from the Rotterdam Study. *European journal of epidemiology*, 35(5):483517, May 2020.
- [314] Renée F A G de Bruijn, Saloua Akoudad, Lotte G M Cremers, Albert Hofman, Wiro J Niessen, Aad van der Lugt, Peter J Koudstaal, Meike W Vernooij, and M Arfan Ikram. Determinants, MRI correlates, and prognosis of mild cognitive impairment: the Rotterdam Study. *Journal of Alzheimer’s disease : JAD*, 42 Suppl 3:S239—49, 2014.
- [315] M Arfan Ikram, Aad van der Lugt, Wiro J Niessen, Peter J Koudstaal, Gabriel P Krestin, Albert Hofman, Daniel Bos, and Meike W Vernooij. The Rotterdam Scan Study: design update 2016 and main findings. *European journal of epidemiology*, 30(12):1299—1315, December 2015.
- [316] Sander Lamballais, Elisabeth J. Vinke, Meike W. Vernooij, M. Arfan Ikram, and Ryan L. Muetzel. Cortical gyrification in relation to age and cognition in older adults. *NeuroImage*, 212:116637, 2020.
- [317] Razvan Marinescu, Arman Eshaghi, Daniel Alexander, and Polina Golland. Brainpainter: A software for the visualisation of brain structures, biomarkers and associated pathological processes. *arXiv preprint arXiv:1905.08627*, 2019.
- [318] Blossom C M Stephan, Christophe Tzourio, Sophie Auriacombe, Hélène Amieva, Carole Dufouil, Annick Alépérovitch, and Tobias Kurth. Usefulness of data from magnetic resonance imaging to improve prediction of dementia: population based cohort study. *BMJ*, 350, 2015.

- [319] Kyung Mee Park, Ji Min Sung, Woo Jung Kim, Suk Kyoong An, Kee Namkoong, Eun Lee, and Hyuk Jae Chang. Population-based dementia prediction model using Korean public health examination data: A cohort study. *PLoS One*, 14(2), February 2019.
- [320] Stanley B. Prusiner. A unifying role for prions in neurodegenerative diseases. *Science*, 336(6088):1511–1513, 2012.
- [321] Pierluigi Gambetti, Ignazio Cali, Silvio Notari, Qingzhong Kong, Wen-Quan Zou, and Witold K Surewicz. Molecular biology and pathology of prion strains in sporadic human prion diseases. *Acta neuropathologica*, 121(1):79–90, January 2011.
- [322] Gianfranco Puoti, Alberto Bizzi, Gianluigi Forloni, Jiri G Safar, Fabrizio Tagliavini, and Pierluigi Gambetti. Sporadic human prion diseases: molecular insights and diagnosis. *The Lancet Neurology*, 11(7):618 – 628, 2012.
- [323] P Parchi, A Giese, S Capellari, P Brown, W Schulz-Schaeffer, O Windl, I Zerr, H Budka, N Kopp, P Piccardo, S Poser, A Rojiani, N Streichemberger, J Julien, C Vital, B Ghetti, P Gambetti, and H Kretzschmar. Classification of sporadic Creutzfeldt-Jakob disease based on molecular and phenotypic analysis of 300 subjects. *Annals of neurology*, 46(2):224–233, August 1999.
- [324] Pierluigi Gambetti, Qingzhong Kong, Wenquan Zou, Piero Parchi, and Shu G Chen. Sporadic and familial CJD: classification and characterisation. *British Medical Bulletin*, 66(1):213–239, 06 2003.
- [325] Piero Parchi, Rosaria Strammiello, Silvio Notari, Armin Giese, Jan P M Langeveld, Anna Ladogana, Inga Zerr, Federico Roncaroli, Patrich Cras, Bernardino Ghetti, Maurizio Pocchiari, Hans Kretzschmar, and Sabina Capellari. Incidence and spectrum of sporadic Creutzfeldt-Jakob disease variants with mixed phenotype and co-occurrence of PrP^{Sc} types: an updated classification. *Acta neuropathologica*, 118(5):659–671, November 2009.
- [326] Piero Parchi, Laura de Boni, Daniela Saverioni, Mark L Cohen, Isidro Ferrer, Pierluigi Gambetti, Ellen Gelpi, Giorgio Giaccone, Jean-Jacques Hauw, Romana Höftberger, James W Ironside, Casper Jansen, Gabor G Kovacs, Annemieke Rozemuller, Danielle Seilhean, Fabrizio Tagliavini, Armin Giese, and Hans A Kretzschmar. Consensus classification of human prion disease histotypes allows reliable identification of molecular subtypes: an inter-rater study among surveillance centres in Europe and USA. *Acta neuropathologica*, 124(4):517–529, October 2012.
- [327] Inga Zerr and Piero Parchi. Chapter 9 - sporadic Creutzfeldt-Jakob disease. In Maurizio Pocchiari and Jean Manson, editors, *Human Prion Diseases*, volume 153 of *Handbook of Clinical Neurology*, pages 155 – 174. Elsevier, 2018.
- [328] S. J. Collins, P. Sanchez-Juan, C. L. Masters, G. M. Klug, C. van Duijn, A. Poleggi, M. Pocchiari, S. Almonti, N. Cuadrado-Corrales, J. de Pedro-Cuesta, H. Budka, E. Gelpi, M. Glatzel, M. Tolnay, E. Hewer, I. Zerr, U. Heinemann, H. A. Kretzschmar, G. H. Jansen, E. Olsen, E. Mitrova, A. Alperovitch, J.-P. Brandel, J. Mackenzie, K. Murray, and R. G. Will. Determinants of diagnostic investigation sensitivities across the clinical spectrum of sporadic Creutzfeldt-Jakob disease. *Brain*, 129(9):2278–2287, 07 2006.
- [329] Simon Mead, Matthew Burnell, Jessica Lowe, Andrew Thompson, Ana Lukic, Marie-Claire Porter, Christopher Carswell, Diego Kaski, Janna Kenny, Tze How Mok, Nina Bjurstrom, Edit Franko, Michele Gorham, Ronald Druyeh, Jonathan D. F. Wadsworth, Zane Jaunmuktane, Sebastian Brandner, Harpreet Hyare, Peter Rudge, A. Sarah Walker, and John Collinge. Clinical trial simulations based on genetic stratification and the natural history of a functional outcome measure in Creutzfeldt-Jakob disease. *JAMA Neurology*, 73(4):447–455, 04 2016.
- [330] Kenta Teruya and Katsumi Doh-Ura. Insights from therapeutic studies for PrP prion disease. *Cold Spring Harbor perspectives in medicine*, 7(3), March 2017.
- [331] Alberto Bizzi, Riccardo Pascuzzo, Janis Blevins, Marco E.M. Moscatelli, Marina Grisoli, Raffaele Lodi, Fabio M. Doniselli, Gianmarco Castelli, Mark L. Cohen, Aymeric Stamm, Lawrence B. Schonberger, Brian S. Appleby, and Pierluigi Gambetti. Subtype diagnosis of sporadic Creutzfeldt-Jakob disease with diffusion MRI. *Annals of Neurology*, n/a(n/a).

Bibliography

- [332] Riccardo Pascuzzo, Neil P Oxtoby, Alexandra L Young, Janis Blevins, Gianmarco Castelli, Sara Garbarino, Mark L Cohen, Lawrence B Schonberger, Pierluigi Gambetti, Brian S Appleby, Daniel C Alexander, and Alberto Bizzi. Prion propagation estimated from brain diffusion MRI is subtype dependent in sporadic Creutzfeldt-Jakob disease. *Acta neuropathologica*, 140(2):169–181, August 2020.
- [333] Rand R. Wilcox. 8 - comparing two independent groups. In Rand R. Wilcox, editor, *Applying Contemporary Statistical Techniques*, pages 237 – 284. Academic Press, Burlington, 2003.
- [334] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.
- [335] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. Unintended Consequences of Machine Learning in Medicine. *JAMA*, 318(6):517–518, 08 2017.
- [336] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019. PMID: 30943338.
- [337] Geoffrey S. Young, Michael D. Geschwind, Nancy J. Fischbein, Jennifer L. Martindale, Roland G. Henry, Songling Liu, Ying Lu, Stephen Wong, Hong Liu, Bruce L. Miller, and William P. Dillon. Diffusion-weighted and fluid-attenuated inversion recovery imaging in Creutzfeldt-Jakob disease: High sensitivity and specificity for diagnosis. *American Journal of Neuroradiology*, 26(6):1551–1562, 2005.
- [338] Henriette J. Tschampa, Kai Kallenberg, Horst Urbach, Bettina Meissner, Claudia Nicolay, Hans A. Kretzschmar, Michael Knauth, and Inga Zerr. MRI in the diagnosis of sporadic Creutzfeldt-Jakob disease: a study on inter-observer agreement. *Brain*, 128(9):2026–2033, 06 2005.
- [339] B. Meissner, K. Kallenberg, P. Sanchez-Juan, D. Collie, D. M. Summers, S. Almonti, S. J. Collins, P. Smith, P. Cras, G. H. Jansen, J. P. Brandel, M. B. Coulthart, H. Roberts, B. Van Everbroeck, D. Galanaud, V. Mellina, R. G. Will, and I. Zerr. MRI lesion profiles in sporadic Creutzfeldt-Jakob disease. *Neurology*, 72(23):1994–2001, 2009.
- [340] P Vitali, E Maccagnano, E Caverzasi, RG Henry, A Haman, C Torres-Chae, DY Johnson, BL Miller, and MD Geschwind. Diffusion-weighted MRI hyperintensity patterns differentiate CJD from other rapid dementias. *Neurology*, 76(20):1711–1719, May 2011.
- [341] Alberto Bizzi and Katell Peoc’h. Amended diagnostic protocol increases the early diagnosis of sporadic Creutzfeldt-Jakob disease. *Neurology*, 91(4):155–156, 2018.
- [342] Aaron Foutz, Brian S. Appleby, Clive Hamlin, Xiaoqin Liu, Sheng Yang, Yvonne Cohen, Wei Chen, Janis Blevins, Cameron Fausett, Han Wang, Pierluigi Gambetti, Shulin Zhang, Andrew Hughson, Curtis Tatsuoka, Lawrence B. Schonberger, Mark L. Cohen, Byron Caughey, and Jiri G. Safar. Diagnostic and prognostic value of human prion detection in cerebrospinal fluid. *Annals of Neurology*, 81(1):79–92, 2017.
- [343] Y. Zhao, A. K. Shrivastava, and K. L. Tsui. Regularized Gaussian mixture model for high-dimensional clustering. *IEEE Transactions on Cybernetics*, 49(10):3677–3688, 2019.
- [344] Guy M. McKhann, David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack, Claudia H. Kawas, William E. Klunk, Walter J. Koroshetz, Jennifer J. Manly, Richard Mayeux, Richard C. Mohs, John C. Morris, Martin N. Rossor, Philip Scheltens, Maria C. Carrillo, Bill Thies, Sandra Weintraub, and Creighton H. Phelps. 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 and Dementia*, 7(3):263–269, May 2011.
- [345] Bryan D. James, Robert S. Wilson, Patricia A. Boyle, John Q. Trojanowski, David A. Bennett, and Julie A. Schneider. TDP-43 stage, mixed pathologies, and clinical Alzheimer’s-type dementia. *Brain*, 139(11):2983–2993, 09 2016.

- [346] Dietmar Rudolf Thal, Estifanos Ghebremedhin, Mario Orantes, and Otmar D. Wiestler. Vascular pathology in Alzheimer disease: Correlation of cerebral amyloid angiopathy and arteriosclerosis/lipohyalinosis with cognitive decline. *Journal of Neuropathology & Experimental Neurology*, 62(12):1287–1301, 12 2003.
- [347] Tyler Lu and Craig Boutilier. Effective sampling and learning for Mallows models with pairwise-preference data. *The Journal of Machine Learning Research*, 15(1):3783–3829, January 2014.
- [348] Peter A. Wijeratne and Daniel C. Alexander. Learning transition times in event sequences: the event-based hidden markov model of disease progression, 2020.
- [349] Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.
- [350] Betty M Tijms, Johan Gobom, Lianne Reus, Iris Jansen, Shengjun Hong, Valerija Dobricic, Fabian Kilpert, Mara ten Kate, Frederik Barkhof, Magda Tsolaki, Frans R J Verhey, Julius Popp, Pablo Martinez-Lage, Rik Vandenberghe, Alberto Lleó, José Luís Molinuevo, Sebastiaan Engelborghs, Lars Bertram, Simon Lovestone, Johannes Streffer, Stephanie Vos, Isabelle Bos, The Alzheimer’s Disease Neuroimaging Initiative, Kaj Blennow, Philip Scheltens, Charlotte E Teunissen, Henrik Zetterberg, and Pieter Jelle Visser. Pathophysiological subtypes of Alzheimer’s disease based on cerebrospinal fluid proteomics. *Brain*, 11 2020.

Acknowledgements

Acknowledgements

Writing this last page of my thesis, it is finally sinking in that my PhD journey is reaching its conclusion ! As engrossed as I have been in the rigours of research over the last few years, writing these acknowledgements gives me an opportunity to zoom out and reflect on the bigger picture. This, right here, is my humble contribution to science, to increasing our understanding of the cacophony of neurodegeneration. I couldn't have done this work without the invaluable contributions and support of others.

First and foremost, I would like to express my gratitude to my promotor, Wiro Niessen, for building an inclusive and diverse research group. Having never worked outside India before the start of my PhD, I was a bit hesitant about moving to a different country for research. But throughout my time at BIGR, I felt very welcome and comfortable working in the group. Secondly, I feel fortunate to have had the mentorship of not just Wiro, but also of my co-promoters, Stefan Klein and Esther E. Bron. I could always count on Stefan for discussing my ideas, no matter how out-of-the-box they were or how long it took to discuss them. I could also always depend on Esther's pragmatic approach for bringing me back from the dopamine rush of a new idea to do a feasibility check, given the timeline of my PhD.

I also want to thank the members of my thesis committee, Marion Smits, Wiesje M. van der Flier, and Marco Lorenzi, for taking their time to read and approve my thesis. Apart from that, I would also like to thank Wiesje and Betty Tijms for giving me the opportunity to work as a post-doc with them and their vast trove of data, the Amsterdam dementia cohort. During the last year of my PhD, after I expressed my interest in visiting Marco's lab in INRIA, he was kind enough to start looking for ways to fund my research visit to his group. Unfortunately, due to the pandemic, our plans had to change. Nevertheless, I enjoyed our brief discussions before we abandoned the plan, as well as working with him while organizing the disease progression modelling tutorial along with Sara Garbarino, Neil Oxtoby, and Igor Koval in ISBI conference this year. I hope we continue collaborating further in this ever-expanding field.

My research benefited a lot from how approachable Meike Vernooij was to discuss the latest results and provide invaluable feedback, first during the EuroPOND project and later in the BEYOND project. I had a very good time working with Meike, Arfan Ikram, and Eline during my time at the Erasmus MC. I hope I get to collaborate with them in the future as well. I feel richer for having had very insightful discussions with Arfan, Kamran Ikram, and Lana Fani. It was incredible how interested they were in understanding even the minutest details of the DEBM algorithm. Trying (and struggling) to explain it without introducing a lot of

Acknowledgements

mathematical jargon into it enhanced my understanding of it as well. I am also very glad to have Arfan and Betty as my mentors for the Heart-Brain Connection pilot project.

Working together with Eline in the EuroPOND project was also quite enjoyable. I could depend on her R skills for generating beautiful plots and I appreciate the fact she has always been responsive in spite of her very busy schedule, managing multiple projects and taking care of her kid ! I would also like to thank Riwan for being such a good friend. Our research areas during our PhDs did not match for us to work on a project together, but I knew I could always count on him for anything unrelated to work - be it when we worked together in the same room or when we (and the entire world) started working remotely.

I would also like to thank all my collaborators in the EuroPOND consortium: Daniel Alexander, Frederik Barkhof, Stanley Durrleman, Neil, Alberto Bizzi, Alberto Redolfi, Damiano Archetti, Riccardo Pascuzzo, and Viktor Wottschel. It was a pleasure witnessing such a large scale project be executed to perfection. I consider myself fortunate to have the opportunity to work with a majority of this team in a follow-up project as well, while being part of the Amsterdam Alzheimer Center. Furthermore, I would like to thank Alberto Bizzi for making sure that I include our work on sCJD in this thesis, in spite of his unimaginably busy clinical work during the ongoing pandemic.

Having never worked on anything related to neurodegenerative diseases before the start of this PhD, developing DEBM at the initial stages of my PhD felt more like playing with a mathematical toy. The insightful discussions we had and the feedbacks we got from Janne Papma and Rebecca Steketee at the early stages of my PhD were crucial for me to understand the significance of my results and in shaping the direction of my research during my PhD. I don't think I can fit all the helpful things Rebecca has done over the years in one paragraph. But the thing that had the most long lasting impact on me, was her live translations of the discussions (in Dutch) during the multi-disciplinary dementia diagnosis meeting. It helped me overcome the language barrier to understand the intricacies of how clinical diagnosis of dementia is made, and shape my view on how data-driven algorithms could be used for clinical diagnosis.

I also thank Janne, Jessica Panman, Emma van der Ende, Harro Seelaar, and all the members of the GENFI consortium for their collaboration and valuable insights for coming up with clinically relevant research question in FTD that could be solved using the methodologies developed in this thesis. Furthermore, I would also like to extend my thanks to Henri Vrooman. Brainstorming with him early on in my PhD for the Aetionomy project led to the development of Co-init DEBM. Later on, it was also a pleasure working with Henri in the BEYOND project.

I also enjoyed collaborating with Marleen de Bruijne, and the ever-enthusiastic Florian Dubost for our paper on nDEBM. I have shared my working space with Gena, Riwanj, Chaoping, Danilo, Luisa, Bo, Emanoel, Arno, Wyke, Carolyn, and Mark over the years and I thoroughly enjoyed doing and discussing the most mundane things with them.

I would like to extend my thanks to the other faculty members of the BGR group: Theo van Walsum, Jifke Veenland, and Dirk Poot for their insightful comments during the presentations. I would also like to thank the present and past members of the BGR group (a non-exhaustive list): Mart, Gijss, Gokhan, Hakim, Marcel, Hua, Jiahang, Willem, Sebastian, Martijn, Zahra, Wietske, Robin, Thomas, Gerda, Yao, Yuanyuan, Jose, Shuai, Kim, Antonio, Mahlet, Adriaan, Taygun - for organizing / participating in the various social events over the years, as well as for taking turns to switch on the computers in the department after power failures at the most inopportune moments, once everyone started working remotely.

I would also like thank my parents for their support over the years, letting me carve my own path in my life while also making sure that I could count on them as a fall back option. Lastly, and most importantly, I would like to thank my wonderful wife, Smruti, whom I have neglected far too much while finalizing the different chapters of my thesis, over the past year. She is one of the most resilient people I know and has been my inspiration for my pilot project on studying the factors affecting cognitive resilience in the Heart-Brain Connection pilot project.

Publications

Journal Papers

- Jessica L. Panman*, **Vikram Venkatraghavan***, Emma L. van der Ende, Rebecca Steketee, Lize C. Jiskoot, Jackie M. Poos, Elise G.P. Dopper, Lieke H.H. Meeter, Laura Donker Kaat, Serge A.R.B. Rombouts, Meike W. Vernooij, Anneke J.A. Kievit, Enrico Premi, Maura Cosseddu, Elise Bonomi, Jaume Olives, Jonathan D. Rohrer, Raquel Sanchez-Valle, Barbara Borroni, Esther E. Bron, John C. van Swieten, Janne M. Papma, Stefan Klein, GENFI consortium investigators. Modelling the cascade of biomarker changes in GRN related frontotemporal dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, Published Online First, January 2021.
- **Vikram Venkatraghavan**, Stefan Klein, Lana Fani, Leontine S. Ham, Henri Vrooman, M. Kamran Ikram, Wiro J. Niessen, Esther E. Bron, for the Alzheimer's Disease Neuroimaging Initiative. Analyzing the effect of APOE on Alzheimer's disease progression using an event-based model for stratified populations. *NeuroImage*, 227: 117646, 2021.
- Iris Dekker, Menno M. Schoonheim, **Vikram Venkatraghavan**, Anand J.C. Eijlers, Iman Brouwer, Esther E. Bron, Stefan Klein, Mike P. Wattjes, Alle M. Wink, Jeroen J.G. Geurts, Bernard M.J. Uitdehaag, Neil P. Oxtoby, Daniel C. Alexander, Hugo Vrenken, Joep Killestein, Frederik Barkhof, Viktor Wottschel. The sequence of structural, functional and cognitive changes in multiple sclerosis. *NeuroImage Clinical*, 29: 102550, 2019.
- **Vikram Venkatraghavan**, Esther E. Bron, Wiro J. Niessen, Stefan Klein, for the Alzheimer's Disease Neuroimaging Initiative. Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. *NeuroImage*, 186: 518-532, 2019.
- Damiano Archetti, Silvia Ingala, **Vikram Venkatraghavan**, Viktor Wottschel, Alexandra L. Young, Maura Bellio, Esther E. Bron, Stefan Klein, Frederik Barkhof, Daniel C. Alexander, Neil P. Oxtoby, Giovanni B. Frisoni, Alberto Redolfi, Alzheimer's Disease Neuroimaging Initiative, & for EuroPOND Consortium. Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease. *NeuroImage Clinical*, 24: 101954, 2019.
- Elisabeth J. Vinke, Marius de Groot, **Vikram Venkatraghavan**, Stefan Klein, Wiro J. Niessen, M. Arfan Ikram, Meike W. Vernooij. Trajectories of imaging markers in brain aging: the Rotterdam Study. *Neurobiology of Aging*, 71: 32-40, 2018.
- **Vikram Venkatraghavan***, Elisabeth J. Vinke*, Esther E. Bron, Wiro J. Niessen, M. Arfan Ikram, Stefan Klein[†], Meike W. Vernooij[†], for the Alzheimer's Disease Neuroimaging Initiative. Progression along APOE-specific data-driven temporal cascades is predictive of Alzheimer's disease in a population-based cohort. *Under review*.
- Esther E. Bron, Stefan Klein, Janne M. Papma, Lize C. Jiskoot, **Vikram Venkatraghavan**, Jara Linders, Pauline Aalten, Peter P. de Deyn, Geert J. Biessels, Jurgen Claassen, Huub A.M. Middelkoop, Marion Smits, Wiro J. Niessen, John C. van Swieten, Wiesje M. van der Flier, Inez H.G.B. Ramakers, Aad van der Lugt, for the Alzheimer's Disease Neuroimaging Initiative, on behalf of the Parelsnoer Neurodegenerative Diseases study group. Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's

disease. *Under review.*

- **Vikram Venkatraghavan***, Riccardo Pascuzzo*, Esther E. Bron, Marco Moscatelli, Marina Grisoli, Amy Pickens, Mark L. Cohen, Lawrence B. Schonberger, Pierluigi Gambetti, Brian S. Appleby, Stefan Klein, Alberto Bizzi. Discriminative event-based model for subtyping in sporadic Creutzfeldt-Jakob disease. *In preparation.*
- Emma L. van der Ende, Esther E. Bron, Jackie M. Poos, Lize C. Jiskoot, Jessica L. Panman, Janne M. Papma, Lieke H. Meeter, Carlo Wilke, Matthis Synofzik, Carolin Heller, Imogen Swift, Aitana Sogorb-Esteve, Arabella Bouzigues, Barbara Borroni, Roberta Ghidoni, Raquel Sanchez-Valle, Fermin Moreno, Caroline Graff, Robert Laforce Jr, Daniela Galimberti, Mario Masellis, Maria Carmela Tartaglia, Elizabeth Finger, Rik Vandenberghe, James B. Rowe, Alexandre de Mendonca, Fabrizio Tagliavini, Isabel Santana, Simon Ducharme, Christopher R. Butler, Alex Gerhard, Johannes Levin, Adrian Danek, Markus Otto, Yolande Pijnenburg, Giovanni Frisoni, Sandro Sorbi, Wiro J. Niessen, Jonathan D. Rohrer, Stefan Klein, John C. van Swieten, **Vikram Venkatraghavan***, Harro Seelaar*, GENFI consortium investigators. A data-driven disease progression model of fluid biomarkers in genetic frontotemporal dementia. *In preparation.*
- Razvan V. Marinescu, Neil P. Oxtoby, Alexandra L. Young, Esther E. Bron, Arthur W. Toga, Michael W. Weiner, Frederik Barkhof, Nick C. Fox, Arman Eshaghi, Tina Toni, Marcin Salaterski, Veronika Lunina, Manon Ansart, Stanley Durrleman, Pascal Lu, Samuel Iddi, Dan Li, Wesley K. Thompson, Michael C. Donohue, Aviv Nahon, Yarden Levy, Dan Halbersberg, Mariya Cohen, Huiling Liao, Tengfei Li, Kaixian Yu, Hongtu Zhu, Jose G. Tamez-Pena, Aya Ismail, Timothy Wood, Hector Corrada Bravo, Minh Nguyen, Nanbo Sun, Jiashi Feng, B. T. Thomas Yeo, Gang Chen, Ke Qi, Shiyang Chen, Deqiang Qiu, Ionut Buciuman, Alex Kelner, Raluca Pop, Denisa Rimoccea, Mostafa M. Ghazi, Mads Nielsen, Sebastien Ourselin, Lauge Sorensen, **Vikram Venkatraghavan**, Keli Liu, Christina Rabe, Paul Manser, Steven M. Hill, James Howlett, Zhiyue Huang, Steven Kiddle, Sach Mukherjee, Anais Rouanet, Bernd Taschler, Brian D. M. Tom, Simon R. White, Noel Faux, Suman Sedai, Javier de Velasco Oriol, Edgar E. V. Clemente, Karol Estrada, Leon Aksman, Andre Altmann, Cynthia M. Stonnington, Yalin Wang, Jianfeng Wu, Vivek Devadas, Clementine Fourrier, Lars Lau Raket, Aristeidis Sotiras, Guray Erus, Jimit Doshi, Christos Davatzikos, Jacob Vogel, Andrew Doyle, Angela Tam, Alex Diaz-Papkovich, Emmanuel Jammeh, Igor Koval, Paul Moore, Terry J. Lyons, John Gallacher, Jussi Tohka, Robert Cizek, Bruno Jedynak, Kruti Pandya, Murat Bilgel, William Engels, Joseph Cole, Polina Golland, Stefan Klein, Daniel C. Alexander. The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge: Results after 1 Year Follow-up. *In revision.*

Conference Papers

- **Vikram Venkatraghavan***, Florian Dubost*, Esther E. Bron, Wiro J. Niessen, Marleen de Bruijne, Stefan Klein, Alzheimer's Disease Neuroimaging Initiative. Event-based modeling with high-dimensional imaging biomarkers for estimating spatial progression of dementia. *International Conference on Information Processing in Medical Imaging (IPMI)*, 2019, pp. 169 - 180.
- **Vikram Venkatraghavan**, Esther E. Bron, Wiro J. Niessen, Stefan Klein. A discriminative event based model for Alzheimer's disease progression modeling. *International Conference on Information Processing in Medical Imaging (IPMI)*, 2017, pp. 121-133.

Conference Abstracts

- Riccardo Pascuzzo, **Vikram Venkatraghavan**, Marco Moscatelli, Marina Grisoli, Esther E. Bron, Stefan Klein, Janis Blevins, Gianmarco Castelli, Lawrence B. Schonberger, Pierluigi Gambetti, Brian S. Appleby, Alberto Bizzi. Modelling the temporal cascade of abnormalities in diffusion magnetic resonance imaging in sporadic Creutzfeldt-Jakob disease. *International Society for Magnetic Resonance in Medicine (ISMRM) conference*, 2020.
- **Vikram Venkatraghavan**, Jessica L. Panman, Emma L. van der Ende, Rebecca Steketee, Lize C. Jiskoot, Jackie M. Poos, Elise G.P. Dopper, Lieke H.H. Meeter, Laura Donker Kaat, Serge A.R.B. Rombouts, Meike W. Vernooij, Anneke J.A. Kievit, Enrico Premi, Maura Cosseddu, Elise Bonomi, Jaume Olives, Jonathan D. Rohrer, Raquel Sanchez-Valle, Barbara Borroni, Esther E. Bron, John C. van Swieten, Janne M. Papma, Stefan Klein, GENFI consortium investigators. Modelling the cascade of biomarker changes in progranulin related frontotemporal dementia. *Alzheimer's Association International Conference (AAIC)*, 2020.
- **Vikram Venkatraghavan**, Stefan Klein, Lana Fani, Leontine S. Ham, Henri Vrooman, M. Kamran Ikram, Wiro J. Niessen, Esther E. Bron, for the Alzheimer's Disease Neuroimaging Initiative. Analyzing the effect of APOE on Alzheimer's disease progression using an event-based model for stratified populations. *Workshop on Computational approaches for ageing and age-related diseases (CompAge)*, 2020.
- Elisabeth J. Vinke, **Vikram Venkatraghavan**, Esther E. Bron, Wiro J. Niessen, M. Arfan Ikram, Stefan Klein, Meike W. Vernooij, for the Alzheimer's Disease Neuroimaging Initiative. Predicting the incidence of Alzheimer's disease in the general elderly population using event-based modelling. *Workshop on Computational approaches for ageing and age-related diseases (CompAge)*, 2020.
- Iris Dekker, Menno M. Schoonheim, **Vikram Venkatraghavan**, Anand J.C. Eijlers, Iman Brouwer, Esther E. Bron, Stefan Klein, Mike P. Wattjes, Alle M. Wink, Jeroen J.G. Geurts, Bernard M.J. Uitdehaag, Neil P. Oxtoby, Daniel C. Alexander, Hugo Vrenken, Joep Killestein, Frederik Barkhof, Viktor Wottschel. Event-based modelling of multimodal biomarkers in multiple sclerosis. *Workshop on Computational approaches for ageing and age-related diseases (CompAge)*, 2020.
- Esther E. Bron, Stefan Klein, Janne M. Papma, Lize C. Jiskoot, **Vikram Venkatraghavan**, Jara Linders, Pauline Aalten, Peter P. de Deyn, Geert J. Biessels, Jurgen Claassen, Huub A.M. Middelkoop, Marion Smits, Wiro J. Niessen, John C. van Swieten, Wiesje M. van der Flier, Inez H.G.B. Ramakers, Aad van der Lugt. External Validation of MRI-based Machine Learning in Alzheimer's disease: the Parelnoer biobank. *Europe Biobank Week (EBW)*, 2020.
- Esther E. Bron, **Vikram Venkatraghavan**, Jara Linders, Wiro J. Niessen, Stefan Klein. Deep versus Conventional Machine Learning for MRI-based diagnosis and prediction of Alzheimer's disease. *Alzheimer's Association International Conference (AAIC)*, 2020.
- Leontine S. Ham, **Vikram Venkatraghavan**, Alle Meije Wink, Wiesje M. van der Flier, Rebecca M.E. Steketee, Marion Smits, Wiro J. Niessen, Stefan Klein and Esther E. Bron, on behalf of the Heart Brain Connection consortium. Transfer learning for event-based modelling of small and heterogeneous datasets. *Dutch Biomedical Engineering (BME) conference*, 2019.
- Iris Dekker, Menno M. Schoonheim, **Vikram Venkatraghavan**, Anand J.C. Eijlers, Iman Brouwer, Mike P. Wattjes, Alle M. Wink, Jeroen J.G. Geurts, Bernard M.J. Uitdehaag, Neil P. Oxtoby, Hugo Vrenken, Joep Killestein, Frederik Barkhof, Viktor Wottschel. Temporal ordering of multimodal biomarker abnormality in multiple sclerosis. *The European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS) conference*, 2019.
- Jara Linders, **Vikram Venkatraghavan**, Wiro J. Niessen, Esther E. Bron. Deep learning for

classification of Alzheimer's disease: Is MRI pre-processing required? *The European Society for Magnetic Resonance in Medicine and Biology (ESMRMB) conference*, 2019.

- **Vikram Venkatraghavan**, Esther E. Bron, Wiro J. Niessen, Stefan Klein. A discriminative event based model for Alzheimer's disease progression modeling. *Dutch Biomedical Engineering (BME) conference*, 2017.
- Jim Smit, Esther E. Bron, **Vikram Venkatraghavan**, Stefan Klein, Wiro J. Niessen. Computer-aided diagnosis of Alzheimer's disease: A comparison of classification models for discriminating AD patients from healthy controls. *ICT.OPEN*, 2017.

PhD portfolio

PhD period: 2016 – 2020
Department: Radiology & Nuclear Medicine
Research School: Advanced School for Computing and Imaging (ASCI) research school

| In-depth courses | year | ECTS |
|---|-------------|-------------|
| Biostatistical Methods 1: Basic Principles (Part A) | 2016 | 2.0 |
| Front End Vision | 2016 | 4.0 |
| Computer Vision by Learning | 2017 | 4.0 |
| Advanced Pattern Recognition | 2017 | 4.0 |
| Scientific Integrity for PhD students | 2018 | 0.3 |
| Biomedical English Writing and Communication | 2018 | 3.0 |
| Conferences | | |
| Dutch Biomedical Engineering (BME) conference, oral presentation | 2017 | 2.0 |
| International Conference on Information Processing in Medical Imaging (IPMI), oral presentation | 2019 | 2.0 |
| Alzheimer's Association International Conference (AAIC), poster presentation | 2020 | 1.0 |
| Workshops and symposiums | | |
| ECCV pre-conference workshop on deep learning in medical imaging, attendance | 2016 | |
| Medical Imaging Symposium for PhD Students (MISP), attendance | 2016 | |
| Medical Imaging Symposium for PhD Students (MISP), organization | 2018 | 1.0 |
| Workshop by European consortium on Progression Of Neurological Disease (EuroPOND), poster presentation | 2018 | 1.0 |
| Workshop on Computational approaches for ageing and age-related diseases (CompAge), poster presentation | 2020 | 1.0 |
| Workshop on Erasmus MC's ACE Alzheimer's day, poster presentation | 2019 | 1.0 |
| Invited talks | | |
| Erasmus MC's Healthy Sciences research day | 2019 | |
| Centre for Brain Research, Indian Institute of Science | 2020 | |

Awards

| | |
|--|------|
| Winner of the ventricle volume prediction category in The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) challenge | 2019 |
| Best poster award at Erasmus MC's Alzheimer's Day workshop | 2019 |

Research seminar series

| | | |
|--|-------------|-----|
| Seminar for front-line researchers in EuroPOND, oral presentation (4x) | 2016-2019 | 1.0 |
| vph-EuroPOND meetings, oral presentation (5x), organization | 2016-2018 | 2.0 |
| BIGR Seminar, oral presentation (4x) | 2016-2020 | 1.0 |
| IDEA meeting, oral presentation (2x) | 2018-2020 | 1.0 |
| Medical Informatics Research lunch, oral presentation (3x) | 2016 - 2019 | 1.0 |
| Neuro meeting, oral presentation (3x) | 2019 - 2020 | 1.0 |
| Epidemiology's 2020 meeting, oral presentation (1x) | 2019 | 0.5 |

Student project supervision

| | | |
|--|------|-----|
| Modeling healthy aging of the brain. Tarik Luisman, Leiden University | 2017 | 1.0 |
| Computer-aided diagnosis and prognosis of dementia. Jim Smit, TU Delft. | 2018 | 1.0 |
| Siamese neural network for Alzheimer's disease classification. Thomas Michaud, Centrale Marseille. | 2018 | 1.0 |
| Transfer learning for event-based modelling in small and heterogeneous datasets. Leontine Ham, Vrije University Amsterdam. | 2018 | 1.0 |
| Deep learning for classification of Alzheimer's disease: Is prior feature engineering required? Jara Linders, University of Amsterdam. | 2019 | 1.0 |
| Convolutional neural networks for multiclass predictions of Alzheimer's disease progression. Thomas Michelotti, Erasmus University. | 2020 | 1.0 |

Reviewing experience

| | |
|---|------|
| IEEE Journal of Biomedical and Health Informatics (1x) | 2019 |
| Frontiers in Big Data (1x) | 2020 |
| International Conference on Information Processing in Medical Imaging (IPMI) (6x) | 2019 |
| International Symposium on Biomedical Imaging (ISBI) (1x) | 2019 |

Other

| | | |
|---|------|-------------|
| Designing website for the BIGR research group | 2017 | 1.0 |
| Total | | 40.8 |

About the author

Vikram Venkatraghavan was born on June 10th, 1988 in Srirangam, a small town in the southern part of India. He finished high school in 2005 in Pune, India. He studied Electronics and Communications Engineering in the National Institute of Technology, Trichy, India. During his undergraduate studies, he developed interests in image processing, machine learning, and discovered his penchant for using these in medical applications. To further explore his interests, he started his Masters in 2009, at the School of Medical Science and Technology at the Indian Institute of Technology, Kharagpur.

From 2012 to 2016, he worked for Angiometrix, a start-up company based in Bangalore and San Francisco, where he led a small and dedicated team in building an image-processing based system for percutaneous coronary intervention assistance. His work led to the filing of two crucial patents for the company. In April 2016, Vikram started his PhD in neurodegenerative disease progression modeling (DPM) at the Biomedical Imaging Group Rotterdam, Erasmus MC in the Netherlands. Following his passion to develop methodologically novel algorithms that are also clinically useful, he developed the algorithms presented in this thesis. These novel methods led to two oral presentations at the highly selective and prestigious IPMI conference. He participated in an international competition to predict the future clinical diagnosis of subjects, volume of ventricles in the brain of these subjects, as well as their cognitive summary scores, in a large Alzheimer's disease cohort. This ended up as the winning entry for the ventricular volume prediction and came second in the overall challenge. Moreover, he forged close collaborations with clinicians, to increase the clinical relevance of these methods which further led to an oral presentation in the clinically oriented ISMRM conference.

Since February 2021, Vikram has been working as a post-doctoral researcher in the Alzheimer Center of the Amsterdam UMC, where he is investigating approaches for early identification of Alzheimer's disease subtypes. He actively collaborates with the researchers in the DPM community and is organizing and co-chairing tutorials on DPM at the ISBI and MICCAI 2021 conferences. In April 2021, he was awarded an "out-of-the-box" pilot grant from the Dutch Heart-Brain Connection consortium.

