COMPARATIVE ORDINAL LONGITUDINAL DATA ANALYSIS TO PREDICT A DIAGNOSIS OF ALZHEIMER'S DISEASE BY USING MULTIMODAL DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AYCAN ÇAĞRI EMEN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

SEPTEMBER 2024

Approval of the thesis:

COMPARATIVE ORDINAL LONGITUDINAL DATA ANALYSIS TO PREDICT A DIAGNOSIS OF ALZHEIMER'S DISEASE BY USING MULTIMODAL DATA

submitted by **AYCAN ÇAĞRI EMEN** in partial fulfillment of the requirements for the degree of **Master of Science** i**n Statistics, Middle East Technical University** by,

Prof. Dr. Naci Emre Altun
Dean, **Graduate School of Natural and Applied Sciences** _____

Prof. Dr. Ayşen Dener Akkaya
Head of the Department, **Statistics** _____

Prof. Dr. Özlem İlk Dağ
Supervisor, **Statistics, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Fulya Gökalp Yavuz
Statistics, METU _____

Prof. Dr. Özlem İlk Dağ
Statistics, METU _____

Assist. Prof. Dr. Fatma Yerlikaya Özkurt
Industrial Engineering, Atılım University _____

Date: 04.09.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name : Aycan Çağrı Emen

Signature :

ABSTRACT

COMPARATIVE ORDINAL LONGITUDINAL DATA ANALYSIS TO PREDICT A DIAGNOSIS OF ALZHEIMER'S DISEASE BY USING MULTIMODAL DATA

Emen, Aycan Çağrı
Master of Science, Statistics
Supervisor : Prof. Dr. Özlem İlk Dağ

September 2024, 118 pages

Alzheimer's Disease (AD) is one of the most common neurological disorders affecting elderly adults. The disease often progresses significantly before symptoms become apparent, making early diagnosis crucial to slowing its progression. Consequently, it is essential to employ appropriate longitudinal analysis methods for predicting the disease. However, previous studies have often neglected the longitudinal nature of the data. This thesis addresses these shortcomings by applying various statistical and machine learning algorithms to the OASIS-3 dataset. The OASIS-3 dataset, which includes demographics, cognitive assessments, MRI and PET scan outputs, was integrated to form a multimodal perspective, allowing for a comprehensive evaluation of the underlying factors influencing AD. The study utilizes marginal models (Generalized Estimating Equations, GEE), random effect models (both frequentist and Bayesian Generalized Linear Mixed Models, GLMMs), and a machine learning algorithm (Ordinal Mixed Effect Random Forest, OMERF) to classify the severity of the disease and identify influential factors. The findings indicate that the Bayesian approach of the Generalized Mixed Effect Model outperforms other methods in identifying severe AD cases among ordinal data. This

research highlights the importance of using advanced longitudinal approaches to better understand and predict Alzheimer's Disease progression, providing valuable insights for early diagnosis and intervention.

Keywords: Generalized Linear Mixed Effect Models, Marginal Models, Ordinal Response, OASIS-3, Alzheimer's Disease Prediction

ÖZ

MULTİMODAL VERİLERİ KULLANARAK ALZHEİMER HASTALIĞI TANISINI TAHMİN ETMEK İÇİN KARŞILAŞTIRMALI SIRALI BOYLAMSAL VERİ ANALİZİ

Emen, Aycan Çağrı
Yüksek Lisans, İstatistik
Tez Yöneticisi: Prof. Dr. Özlem İlk Dağ

Eylül 2024, 118 sayfa

Alzheimer Hastalığı (AH), yaşlı yetişkinleri etkileyen en yaygın nörolojik hastalıklardan biridir. Hastalık, semptomlar ortaya çıkmadan önce önemli ölçüde ilerleyebilir, bu nedenle erken teşhis, hastalığın ilerlemesini yavaşlatmak için kritik öneme sahiptir. Dolayısıyla, hastalığı tahmin etmek için uygun boylamsal analiz yöntemlerinin kullanılması gereklidir. Ancak, önceki çalışmalar sıklıkla verilerin boylamsal doğasını göz ardı etmiştir. Bu tez, bu eksiklikleri gidermek amacıyla, OASIS-3 veri kümesine çeşitli istatistiksel modeller ve makine öğrenimi algoritması uygulamaktadır. OASIS-3 veri kümesi; demografik bilgiler, bilişsel değerlendirmeler, MRG ve PET tarama sonuçlarını içermektedir ve bu veriler birleştirilerek Alzheimer Hastalığını etkileyen temel faktörlerin kapsamlı bir değerlendirmesi yapılmıştır. Çalışmada, hastalığın şiddetini sınıflandırmak ve etkili faktörleri belirlemek için marjinal modeller (Genelleştirilmiş Tahmin Denklemleri, GEE), rastgele etkili modeller (hem frekansçı hem de Bayesçi Genelleştirilmiş Doğrusal Karışık Modeller, GLMMs) ve bir makine öğrenimi algoritması (Ordinal Mixed Effect Random Forest, OMERF) kullanılmıştır. Bulgular, Bayesçi

Genelleştirilmiş Karışık Modellerinin, şiddetli AH vakalarını ordinal veriler arasında tanımlamada diğer yöntemlerden daha üstün olduğunu göstermektedir. Bu araştırma, Alzheimer Hastalığı'nın ilerleyişini daha iyi anlamak ve tahmin etmek için gelişmiş boylamsal yaklaşımların önemini vurgulamakta olup, erken teşhis ve müdahale için değerli bilgiler sağlamaktadır.

Anahtar Kelimeler: Genelleştirilmiş Doğrusal Karma Modeller, Marjinal Modeller, Sıralı Bağımlı Değişken, OASIS-3, Alzheimer Hastalığı Tahmini

To my family

ACKNOWLEDGMENTS

First of all, I would like to express my deepest gratitude to my advisor Prof. Dr. Özlem İlk Dağ for her never-ending guidance, encouragement, patience and support during this study. It was an honor for me to work with her. I am really lucky to have an opportunity to utilize her experiences and knowledge. I strongly believe that this thesis could not be ended without her stimulating advice and motivational energy.

I would like to thank the examining committee members; Assist. Prof. Dr. Fatma Yerlikaya Özkurt and Assoc. Prof. Dr. Fulya Gökalp Yavuz for their participation, valuable and constructive feedback, suggestions, and comments.

I would especially like to thank my dear friend Ülkem Çalık for her patience, precious support and deep understanding on this long journey of mine.

I want to express my grateful thanks to my dear mother Ayşe Emen, my father Nuri Emen, and my sisters Ayça and Aysun, who have provided me with endless love and unconditional support in my life.

Finally, I would especially like to thank the participants and organizers of the OASIS-3. The data were provided in part by OASIS-3 Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P50AG00561, P30NS09857781, P01AG026276, P01AG003991, R01AG043434, UL1TR000448, R01EB009352.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiv

LIST OF FIGURES xv

LIST OF ABBREVIATIONS xvi

CHAPTERS

1 INTRODUCTION 1

2 LITERATURE REVIEW 5

2.1 Literature Review on Studies of Alzheimer's Disease 5

2.2 Literature Review on Studies of Algorithms Applied in Longitudinal Studies with Ordinal Response 8

3 METHODOLOGY 11

3.1 Data Pre-Processing 12

3.1.1 Session Matching-Up and Merging 12

3.1.2 Feature Selection with Minimum Redundancy and Maximum Relevance (MRMR) 21

3.2 Marginal Models with Ordinal Response 22

3.2.1 Generalized Estimating Equation for Clustered Ordinal Responses 23

3.2.2 Marginal Models for Correlated Ordinal Multinomial Responses ... 25

3.3 Generalized Linear Mixed Effect Models 26

3.3.1 Generalized Linear Mixed Models by L1-Penalized Estimation 27

3.3.2 Cumulative Linear Mixed Models 30
3.4 Bayesian Generalized Linear Mixed Effect Models 32
3.4.1 Monte Carlo Markov Chain for Generalized Linear Mixed Effect Models 34
3.4.2 Bayesian Generalized Linear Multivariate Multilevel Models 36
3.5 Machine Learning Models 38
3.5.1 Ordinal Mixed Effect Random Forest 40
3.6 Model Evaluation Metrics 42
3.6.1 Confusion Matrix 42
3.6.2 Akaike Information Criterion 45
3.6.3 Bayesian Information Criterion 46
4 DATA ANALYSIS 47
4.1 Design of Study 47
4.1.1 Introduction to Open Access Series of Imaging Studies (OASIS) Project 47
4.2 Data Pre-Processing 48
4.3 Data Overview 54
4.4 Implementation of Methods 60
4.4.1 Marginal Models with ORDGEE 64
4.4.2 Marginal Models with ORDLORGEE 66
4.4.3 Generalized Linear Mixed Effect Models with GLMMLASSO 68
4.4.4 Generalized Linear Mixed Effect Models with CLMM 70
4.4.5 Bayesian Generalized Linear Mixed Effect Models with MCMCGLMM 73
4.4.6 Bayesian Mixed Effects Ordinal Regression Models with BRMS ... 80

4.4.7 Ordinal Mixed Effect Random Forest with OMERF 84
4.5 Comparison of Algorithms 89
5 CONCLUSION AND DISCUSSION 93
REFERENCES 97
APPENDICES
A. R Code for Data Cloning (Slightly Modified Version of OMERF) 109

LIST OF TABLES

TABLES

Table 4.1 Normalization of MRI FreeSurfer-derived cortical volumes 49
Table 4.2 Demographic variables of the OASIS-3 dataset..... 55
Table 4.3 Dementia categorization of CDR scores 56
Table 4.4 Distribution of Dementia Status 56
Table 4.5 Summary of Cross-validation..... 61
Table 4.6 Description of the final dataset..... 63
Table 4.7 Summary of the ordgee Random Effect model with “OASIS ID”..... 65
Table 4.8 Correlation Structures of Marginal Models..... 67
Table 4.9 Tukey’s Five-Number Summary of Estimated Intrinsic Parameters 67
Table 4.10 Summary of the ordLORgee Random Effect model with “OASIS ID” 67
Table 4.11 Summary of the glmmLasso Random Effect model with “OASIS ID” 69
Table 4.12 Summary of the clmm Random Effect model with “OASIS ID” 71
Table 4.13 Summary of the threshold or cutpoints 71
Table 4.14 Hyperparameters of MCMCglmm algorithm..... 73
Table 4.15 Summary of the MCMCglmm cutpoints and structures 74
Table 4.16 Summary of the MCMCglmm algorithm..... 75
Table 4.17 Summary of the brms algorithm..... 81
Table 4.18 Summary of the CLMMs using the OMERF algorithm with offset variable 85
Table 4.19 Summary of the threshold or cutpoints via OMERF..... 85
Table 4.20 Test data performance (Overall results) 90
Table 4.21 F1 Scores by Class 91
Table 4.22 Recall by Class 92

LIST OF FIGURES

FIGURES

Figure 1.1 Comparison of normal (left) and severely demented (right) brain tissue (T.U.O.Q., 2018) 1
Figure 3.1 Pillars of OASIS-3 Non-imaging Datasets (Own work) 13
Figure 3.2 Matching Up Session Process by OASIS ID and Days From Entry (Assessment Datasets) 18
Figure 3.3 Matching Up Session Process by OASIS ID and Days From Entry (Scan Datasets) 19
Figure 3.4 An Example of Matching-Up Session Process 20
Figure 3.5 Confusion Matrix examples. (a) Binary Classification problem confusion matrix. (b) Multiclass Classification problem confusion matrix (Markoulidakis et al., 2021) 43
Figure 4.1 Flowchart of utilized feature selection approaches 53
Figure 4.2 Boxplots of MMSE and TVV by response variable 58
Figure 4.3 Boxplots of LH Bankssts Thickness and LH Lingual Thickness by response variable 59
Figure 4.4 Boxplots of Simon Number and LOGIMEM by response variable 59
Figure 4.5 Bar graphs of APOE-e4 Status and Multivitamin Use by response variable 60
Figure 4.6 Multicollinearity plot 62
Figure 4.7 Trace and density plots of the estimate of the intercept and covariates 77
Figure 4.8 Trace and density plots of the estimate of the covariates 78
Figure 4.9 Trace and density plots of the estimate of the random effect and units 79
Figure 4.10 Convergence plots of covariates 83
Figure 4.11 Convergence plots of covariates 83
Figure 4.12 Variable importance plot of increase in node purity 86
Figure 4.13 Variable importance plot of increase of Mean Square Error (MSE) ... 86
Figure 4.14 Partial Dependence Plots of covariates used in OMERF 88

LIST OF ABBREVIATIONS

ABBREVIATIONS

Abbreviations	Explanations
AD	Alzheimer’s Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
ADRC	Alzheimer’s Disease Research Center
AGQ	Adaptive Gaussian Quadrature
CDR	Clinical Dementia Rating
CDRSUM	CDR-Sum of Boxes
CDRTOT	CDR-Total Score
CFS	Correlation-based Feature Selection
CIC	Correlation Inflation Criterion
CLM	Cumulative Link Model
CLMM	Cumulative Link Mixed Model
DIC	Deviance Information Criterion
FN	False Negative
FP	False Positive
GEE	Generalized Estimation Equation
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
glmmLasso	Generalized Linear Mixed Models by L1-penalized Estimation
HMC	Hamilton Monte Carlo
NUTS	No-U-Turn-Sampler
LH	Left Hemisphere
RH	Right Hemisphere
LKJ	Lewandowski-Kurowicka-Joe Correlation Distribution
LOCF	Last Observation Carried Forward
MCMC	Markov Chain Monte Carlo
MIM	Mutual Information Matrix

MLE	Maximum Likelihood Estimation
MRI	Magnetic Resonance Imaging
mRMR	Minimum Redundancy Maximum Relevancy
MSE	Mean Squared Error
OASIS	Open Access Series of Imaging Studies
OMERF	Ordinal Mixed Effect Random Forest
PET	Positron Emission Tomography
QIC	Quasi Information Criterion
QICu	Quasi-likelihood Under the Independence Model Criterion
REML	Restricted Maximum Likelihood
RJC	Rotnitzky–Jewell Information Criterion
TVV	Total Ventricular Volume
VIF	Variance Inflation Factor
WAIC	Watanabe–Akaike Information Criterion
CT	Computed Tomography
MCI	Mild cognitive impairment
WUST	Washington University of Science and Technology

CHAPTER 1

INTRODUCTION

Alzheimer's Disease (AD) is a neurological disorder, predominately affecting older adults. Individuals with AD often struggle with tasks that require multitasking. In the early stages of the disease, impairments in executive functioning can range from subtle to significant, impacting individuals' daily activities.

As the disease progresses, changes in the brain become evident, with tissue loss and shrinkage in certain areas compared to non-demented, healthy individuals. Figure 1.1 illustrates the comparison between normal and severely demented brain tissues by highlighting the reduction in brain volume and enlargement of ventricles in severe AD cases.

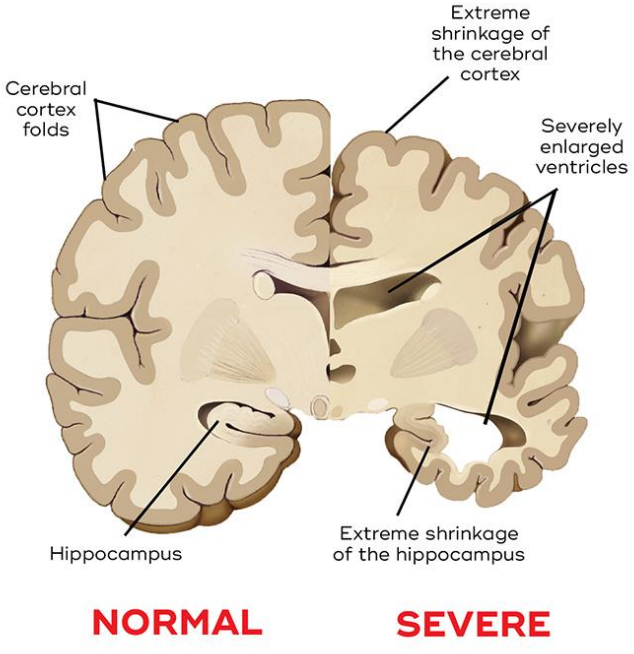


Figure 1.1 Comparison of normal (left) and severely demented (right) brain tissue (T.U.O.Q., 2018)

In studies on AD and its progression, multimodal longitudinal approaches have often been overlooked in predicting or early detecting the disease due to the extensive time and data required. Most studies have focused on evaluating structural Magnetic Resonance Imaging (MRI) and other scans, often relying on single images for disease identification. However, the nature of AD requires monitoring disease progression over time to provide accurate treatment and precautions. Current detection methods through single MRI or PET scans do not adequately account for the progressive nature of the disease.

Hereby, in this study, suitable longitudinal modeling approaches to predict AD and determine influential factors were adopted, considering individuals' relevant information such as family history, cognitive assessments, MRI scans, and more. Classical statistical methods such as linear regressions and specific machine learning algorithms (i.e. Random Forest, Support Vector Machines et cetera.) are inadequate for longitudinal data because of the independence assumption among observations. Therefore, accounting for subject-specific correlations in the models is crucial to make accurate assumptions from longitudinal data.

To address the structure of the longitudinal data, appropriate modeling techniques were employed to capture inter-subject variations. Marginal models account for intraclass correlations, but neglect inter-subject variations. In contrast, random effect models primarily focus on observing inter-subject variations. Moreover, the Ordinal Mixed Effect Random Forest (OMERF) algorithm, a recently proposed novel approach, was developed to estimate both fixed and random effects separately and then combine them to explain the complex structure of longitudinal data.

The principal contributions of this study are as follows:

- Identification of Potential Factors: This research highlights underlying potential factors that may be influential in predicting AD.
- Algorithmic Refinement: This work utilizes advanced statistical and machine learning, including a modified OMERF algorithm, applicable to longitudinal data with ordinal responses.
- Feature Selection Methodology: It provides a comprehensive methodology for selecting optimal features in longitudinal data with ordinal responses, ensuring a robust analysis of the factors affecting dementia severity.

These contributions advance the understanding of AD prediction and offer valuable tools and methodologies for future research in the field.

The rest of the study is structured as follows. Section 2 conducts a literature review that is divided into two aspects. While the first aspect outlines studies regarding AD, the latter briefly discusses applications of longitudinal studies with ordinal responses. Section 3 details the data pre-processing phase to generalize the study's reproducibility, providing the necessary steps taken in this phase. Additionally, Section 3 articulates the classical models and machine learning algorithms explaining their theoretical background and performance metrics. Section 4 delves into a real-world case study, applying the methodologies and algorithms to the Open Access Series of Imaging Studies (OASIS) data. The outcomes of implementing these methods and comparing the algorithms are discussed in the latter part of this section. Finally, Section 5 is covered, which includes the conclusions and discussion. The analyses in this thesis were performed using the R programming language (R version 4.3.1; R Core Team 2023). All R codes for implementing custom functions are available at https://github.com/aycancagriemen.

CHAPTER 2

LITERATURE REVIEW

The detection of AD and its progression are evaluated in this section by examining relevant studies and applications on the longitudinal data. The related articles on the modeling of AD and the dementia status of subjects are reviewed. Whereas, the models and techniques that apply to longitudinal data with ordinal response are discussed in the following section.

2.1 Literature Review on Studies of Alzheimer's Disease

AD is a progressive neurodegenerative disease that has a serious impact on the subject's mental health by impairing memory and cognitive functions (Cass, 2017). Although most of the symptoms of the disease are seen in elderly people, the disease does not occur due to aging (Hardan et al., 2021). Extensive research has been conducted so far to diagnose the disease in its early stages and identify prevention approaches to slow down its progression. Consequently, data analysis on AD diagnosis has been performed considering two types of data, which are cross-sectional and longitudinal (Cui et al., 2018).

Cross-sectional AD diagnosis data analysis has often focused on MRI data, processed via several brain images from individuals, such as those in the ADNI and OASIS datasets. Various methods have been used for cross-sectional datasets, which involve observing individuals at a single point in time. Yet, these methods are limited in capturing the progression of the disease since they only provide a snapshot of each individual. In contrast, longitudinal studies involving multiple time points for the same individuals over time where the study has been carried out. The longitudinal studies enable the discovery of patterns and relationships among different factors

that may be vital for dementia studies. Thus, longitudinal studies provide adequate information on diagnosing the disease in its early stages or capturing changes in the individual's current status. To follow up individual's overall changes and be able to make precise diagnoses, data should be collected for the lifespan of the study. However, due to the nature of the area of dementia studies, dropouts can be observed quite often (Handels, 2020) in longitudinal studies, which may create a bias in results.

In the literature, an in-depth review of AD has been conducted so far in longitudinal studies. Several data sources have been centered on investigating measuring AD markers, such as Alzheimer's Disease Neuroimaging Initiative (ADNI), Baltimore Longitudinal Study of Aging (BLSA), Heinz Nixdorf Recall Study (HNRS), Open Access Series of Imaging Studies (OASIS). Considering multimodal neuroimaging aspects, the OASIS-3 dataset was selected in this thesis. Apart from the OASIS-3, several datasets are available for different research questions in the OASIS Project. For instance, OASIS-1, which was initially presented in 2007 by Marcus et al. as cross-sectional study, has 416 subjects and based on the CDR scoring the dataset can be categorized. The dataset consists of MRI Scan outputs besides demographic information of the subjects. It was followed by OASIS-2 (Marcus et al., 2010), which was published as a longitudinal study of 150 subjects, where imaging sessions of the subjects can be accessed in this dataset. Thirdly, longitudinal multimodal neuroimaging dataset, OASIS-3 (LaMontagne et al., 2019), has been published including MRI and PET Scans outputs along with assessment sessions and demographic information. Lastly, OASIS-4, which is a cohort study of 663 subjects, was evaluated for memory disorder and dementia including several assessments (Koeniget et al., 2020). Detailed summarization regarding OASIS datasets can be found in the Introduction to Open Access Series of Imaging Studies Project section.

Predicting the AD and its progression has involved statistical models, machine learning (ML) and deep learning (DL) algorithms along with image processing approaches by utilizing various data sources previously mentioned. Diwate et al. (2021) presented their work regarding the OASIS-2 dataset, which is the first

longitudinal dataset in OASIS, by investigating different machine learning algorithms for the classification of the dementia status of individuals. Jiang et al. (2020) conducted a study to identify early mild cognitive impairment and normal control classification using Support Vector Machine (SVM) and Deep Learning (DL) in the ADNI dataset. On the OASIS dataset, Naidu et al. (2019) addressed the severity status of AD by implementing ML algorithms (e.g. SVM and Gradient Boosting) and a statistical model (Lasso Regression) and it was concluded that among the suggested model and algorithms, RF performed better compared to others. Further, Zhang et al. (2017) adopted a data-driven approach called landmark-based feature extraction by using longitudinal structural MR images of the ADNI dataset. Their proposed method was compared with several benchmark methods, and results revealed that it outperformed these benchmarks in terms of classification performance.

One of the main limitations of these studies that utilize machine learning models is the disregard of longitudinal data structure and intraclass correlation among clusters. The use of machine learning algorithms has emerged in the field of longitudinal data modeling because they are generally associated with increased complexity. Chen and DuBois (2011) underlined that machine learning models such as SVM is being frequently used in the classification and prediction of high-dimensional data, even though SVMs typically designed for the analysis of cross-sectional data that is collected at one-time point. Furthermore, Cascarano et al. (2023) underlined that most of the ML models disregard the assumption of being independent and identically distributed, thereby this might lead results to be biased. Hence, in this thesis, a machine learning algorithm, which is specifically designed for handling longitudinal data with ordinal responses and effectively managing the intraclass correlation structure, is utilized.

Another important aspect of longitudinal data analysis for AD prediction is that most studies have treated the response variable as binary. However, when taking Clinical Dementia Rating (CDR) scores into account, it can be categorized even further than two categories, which are frequently used as demented and non-demented. In this

study, severity levels of the dementia status were classified as ordinal with three categories based on the CDR scores by considering each visit period.

2.2 Literature Review on Studies of Algorithms Applied in Longitudinal Studies with Ordinal Response

In analyzing longitudinal data with ordinal response, various approaches have been used. These approaches can be categorized as transitional models, random mixed-effect models and marginal models (Diggle et al., 1994; Fahrmeir & Tutz, 1994). Transitional models utilize Markov-type approaches in their model construction to be able to identify the changes among time points that are sequential. Whereas marginal models are used to examine changes in multivariate distributions (response is assumed to be ordinal or multinomial) and random effect models focus on the individuals that are monitored over time.

To handle repeated ordinal measurements, Marginal models or Generalized Estimating Equations approaches have been extended for these cases. For clustered ordinal data, Clayton (1992) extended the GEE approach of Liang and Zeger (1986) by proposing a complementary log-log model. The study by Molenberghs and Lesaffre (1994) addressed the challenges associated with the lack of likelihood-based regression models for multivariate categorical responses. In this context, they proposed a model that utilizes a two-dimensional Plackett distribution, incorporating a global odds ratio for handling time-varying and subject-specific features. Thomas (1996) proposed an intuitive approach by extending mixed effect models for binary responses, based on Conaway's (1990) study, to handle ordinal responses. This method utilizes cumulative probabilities explained through binomial expansions, as opposed to Conaway's approach, which uses linear combinations of marginal probabilities.

An ordinal response mixed-effect regression model was studied to be applicable in both clustered and longitudinal data by allowing multiple random effects in the

model (Hedeker & Gibbons, 1994). The link function is assumed to be a probit or logit function, and the parameter estimation is done by Maximum Marginal Likelihood estimation applying Gauss-Hermite quadrature approximation.

Additionally, hybrid models such as combining GLMMs with penalization approaches, are a notable topic in the literature due to the numerous features available in the data. Most of the algorithms or approaches suggested for this topic primarily focus on binary response data. To extend this to the ordinal response case, Hou and Archer (2015) proposed a novel two-stage approach that integrates the Generalized Monotone Incremental Forward Stagewise (GMIFS) algorithm, developed by Hastie et al. (2007) to determine ordinal response classes.

Bergonzoli et al. (2024) developed an algorithm that employs a random forest algorithm by controlling the nested structure of the longitudinal data. The author's previous work, Generalized Mixed Effect Random Forest (GMERF), provides a strong foundation in the development of overcoming the ordinal classification problems. The approach involves a novel method of estimating model components separately. Fixed effect components are fit via a random forest algorithm (Breiman, 2001) avoiding the grouping structure, while the results from this random forest algorithm are used to estimate the cumulative mixed effect model.

Despite marginal and mixed effect models discussed and investigated frequently from both Bayesian and frequentist perspectives, transitional models, advancing Markov-type techniques, have been studied as well. Lee and Daniels (2007) suggested an ordered logit model for calculating marginal probabilities and a Markov transition model for coping with the temporal structure of the dataset. For parameter estimation, the Fisher scoring approach has been utilized to calculate the Maximum Likelihood Estimation (MLE).

In this thesis, the marginal and random effect models have been discussed along with machine learning algorithms that are applicable to ordinal response. Hence, transitional models that have been utilizing Markov-type approaches are out of the scope of this study.

CHAPTER 3

METHODOLOGY

In this section, the methods suitable for analyzing longitudinal data with ordinal response are explained in detail. The data used in this thesis is a collection from several studies at the Washington University Knight Alzheimer Disease Research Center (LaMontagne et al., 2019). This data includes various types of sessions, such as clinical sessions, neuropsychological sessions and MRI sessions along with demographic information of participants. Therefore, it is necessary to match up sessions of datasets by considering their days from entries, which may occur at different time points.

Firstly, a comprehensive walkthrough of the session merging process was provided. Following this, the feature selection method used in the thesis, Minimum Redundancy Maximum Relevancy (MRMR), was described. This method, which is often utilized in high-throughput genomic data, applies the Mutual Information Matrix (MIM) and correlation coefficients between response and explanatory variables recursively to identify the most relevant features with respect to the response variable.

Then, for analyzing longitudinal ordinal response data, basic theoretical principles were discussed for each model and algorithm, including Generalized Estimating Equations (GEE), Bayesian and Frequentist approaches of Generalized Linear Mixed Effect Models (GLMMs), and Machine Learning algorithm.

As part of the machine learning algorithm, an intuitive approach has been proposed, which extends the existing random forest algorithms, to fit ordinal response mixed effect model cases. Lastly, model evaluation metrics employed to assess the effectiveness of both machine learning algorithm and statistical models are discussed.

Considering the fact that the methodology section sheds light on the analytical frameworks of the study; it plays a crucial role in providing reliable and interpretable results.

3.1 Data Pre-Processing

3.1.1 Session Matching-Up and Merging

3.1.1.1 Introduction

The OASIS-3 (Open Access Series of Imaging Studies-3) dataset is a considerable resource for research on normal aging and AD. It includes longitudinal imaging, clinical, cognitive, and biomarker data for over 1,373 participants, making it an invaluable tool for tracking brain changes over time.

3.1.1.2 Dataset Overview

The dataset comprises MRI scans, PET scans, and various assessment sessions. These sessions provide a wide range of information about the participants, enabling researchers to study the temporal progression of brain changes associated with aging and dementia.

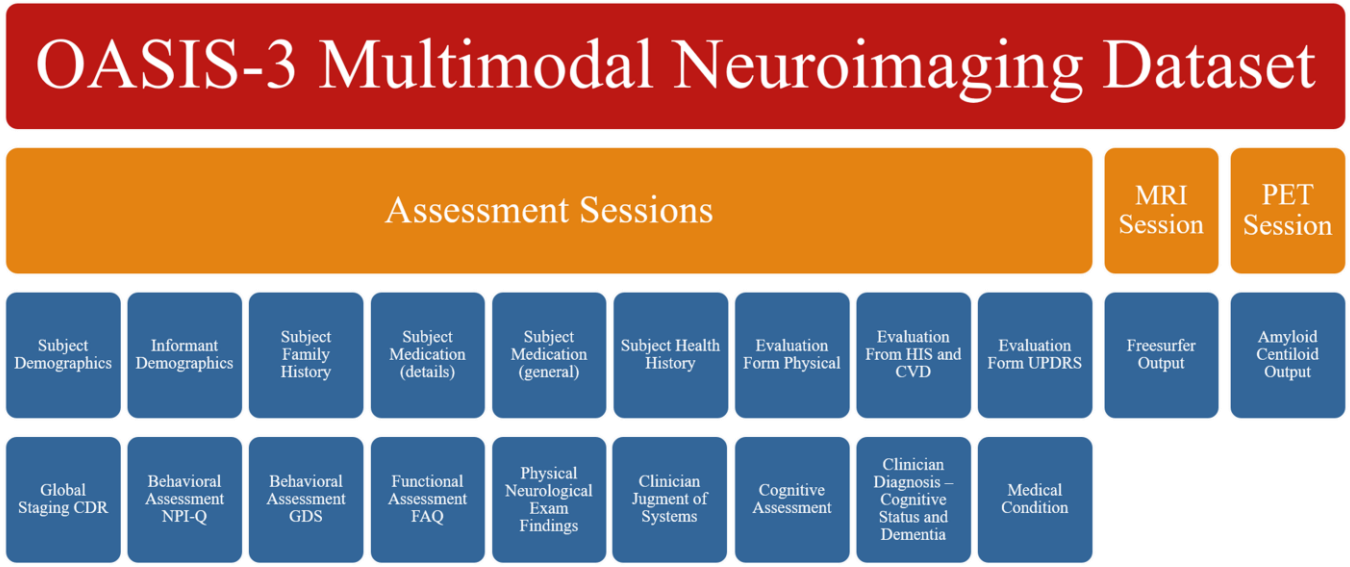


Figure 3.1 Pillars of OASIS-3 Non-imaging Datasets (Own work)

3.1.1.3 Longitudinal Data Collection

The OASIS-3 dataset spans over 30 years, following the same subjects across multiple sessions. However, not all participants are available for every session due to reasons such as loss to follow-up, death, or non-participation. Consequently, sessions do not always occur at the same visit, which poses a challenge for creating a complete dataset.

3.1.1.4 Session Criteria Matching

To create a combined dataset, it is essential to match imaging sessions outputs with corresponding clinical data accurately. Each participant's data is identified by a unique ID in the form "d0000," where "0000" indicates the days since the subject's entry into the study. A "days from entry" value of 0 signifies the subject’s initial visit.

3.1.1.5 Matching Process

The matching process, as highlighted in the OASIS-3 Imaging Data Dictionary, serves as a foundational component of this thesis. The detailed methodology for matching various OASIS-3 sessions has been elaborated in the GitHub repository of XNAT tools and development frameworks (NRG XNAT) under the "OASIS-3 scripts." This thesis extends the provided information by incorporating additional conditional layers to ensure the inclusion of all available records, thereby revisiting the process of matching multiple datasets.

The primary objective is to utilize as much collected data from the OASIS-3 studies as possible to facilitate comprehensive data integration and analysis. The matching process employs a rigorous procedure for merging multiple datasets, including assessment data, MRI scan data, and PET scan data. This process is designed to

uphold data integrity and accuracy through a series of validation checks and merging rules.

Initially, the primary and secondary assessment data undergo a validation step, where each dataset is evaluated for its unique identifier (OASIS ID) and visit day (Days From Entry) variables. If neither identifier is found, the process terminates for that dataset. The OASIS ID is a mandatory requirement for identification. Consequently, the presence of only the OASIS ID or both the OASIS ID and Days From Entry match leads to different conditional checks during the process. If both identifiers are located, the records from both sides are stored for inclusion in the merged data. If only the OASIS ID is present, the records are assessed for their existence within 365 days relative to the primary assessment date. This temporal check ensures the relevance and recency of the data. If there are no available records within the 365 days, primary data records are matched with the nearest records from secondary data. However, primary data records are retained as they are, while all records from secondary data are replaced with NULL before being added to the merged dataset. Additionally, if any OASIS ID is present in only one of the datasets; to prevent data loss, the dataset containing the OASIS ID remains unchanged, while the variables in the other dataset (which do not have a corresponding OASIS ID match) are set to NULL.

Subsequently, when records fall within the acceptable time frame, the next step involves checking for multiple records associated with the identifier. If no multiple records exist, the datasets are merged directly. However, if multiple records are found, further differentiation is made based on their proximity to the primary assessment date. This distinction is crucial for ensuring that the most relevant and recent data points are utilized. An algorithmic rank assignment is employed here to prioritize data based on predefined criteria. For ranking evaluation, if the rank is set to "1," the records are merged. On the other hand, if the rank is greater than "1," the process is terminated.

As a final step, all merged records are combined to form the merged assessment data. It is important to note that, given the multiple assessment datasets (eighteen in total), the matching process for assessment data is an iterative binary match process.

A similar procedure is followed for MRI and PET scan data, starting with the identification of records through OASIS ID, followed by the temporal check within 365 days. The process then branches based on the existence of multiple records, implementing a rank assignment system when necessary to ensure the selection of the most convenient data. During the integration process, only assessment records from individuals with MRI data were considered when merging the MRI and PET scan datasets with the assessment dataset. Once all relevant checks and merges are completed, the integrated assessment and scan data are combined to form the final dataset.

This systematic approach, as depicted in the flowchart, underscores rigorous data validation and consolidation principles. It ensures that only the most accurate and relevant data points are included in the final dataset, thereby enhancing the reliability of subsequent analyses. The multi-step validation and merging process ensures methodological robustness, making it a critical component of the overall data management strategy in research settings. However, a notable drawback is the potential for sparse data resulting from the records that do not satisfy the perfect match criteria (both OASIS ID and Days From Entry). Consequently, careful examination and additional approaches, such as MRMR or variance threshold methods, will be necessary before applying missing data imputation techniques, which will be discussed in further sections.

For matching MRI sessions with ADRC Clinical Data entries, a common criterion is to consider all entries within one year (365 days) before or after the MRI session date as valid. This means that the closest ADRC Clinical Data entry to the MRI session within this time frame is deemed a match.

The depicted flowcharts of the matching-up session data process are highlighted separately in Figure 3.2 and Figure 3.3. Final data is obtained by combining the

assessment and scan datasets, merging through a session matching process. For better clarification, an example scenario is also provided in Figure 3.4.

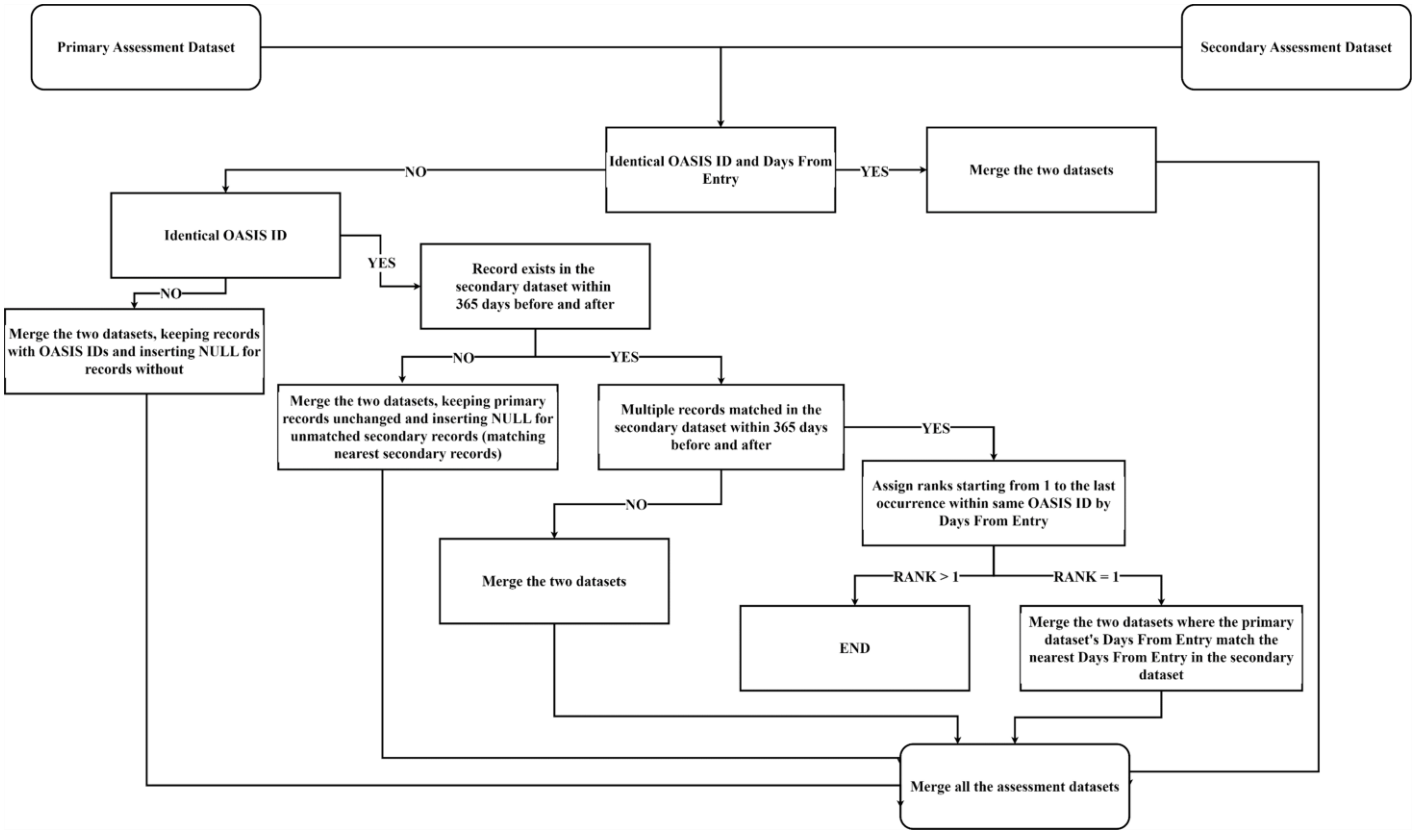


Figure 3.2 Matching Up Session Process by OASIS ID and Days From Entry (Assessment Datasets)

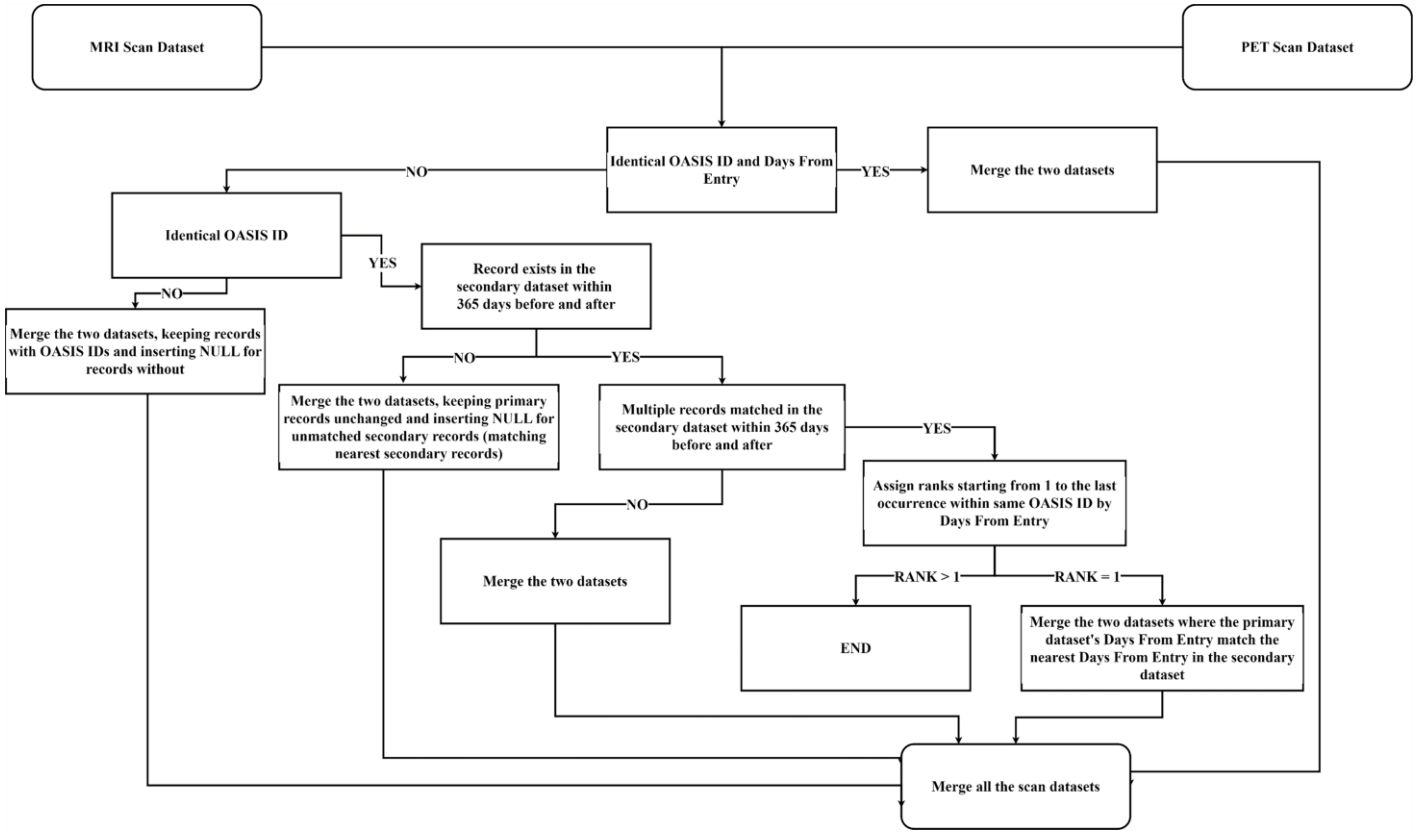
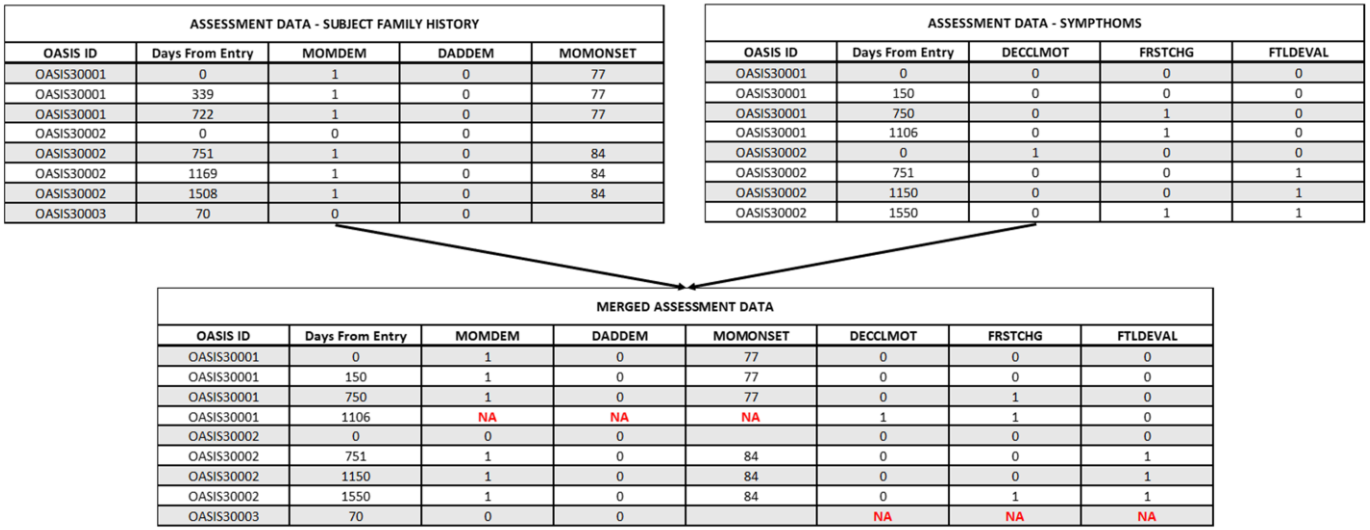


Figure 3.3 Matching Up Session Process by OASIS ID and Days From Entry (Scan Datasets)



ASSESSMENT DATA - SUBJECT FAMILY HISTORY

OASIS ID	Days From Entry	MOMDEM	DADDEM	MOMONSET
OASIS30001	0	1	0	77
OASIS30001	339	1	0	77
OASIS30001	722	1	0	77
OASIS30002	0	0	0	
OASIS30002	751	1	0	84
OASIS30002	1169	1	0	84
OASIS30002	1508	1	0	84
OASIS30003	70	0	0	

ASSESSMENT DATA - SYMPTHOMS

OASIS ID	Days From Entry	DECCLMOT	FRSTCHG	FTLDEVAL
OASIS30001	0	0	0	0
OASIS30001	150	0	0	0
OASIS30001	750	0	1	0
OASIS30001	1106	0	1	0
OASIS30002	0	1	0	0
OASIS30002	751	0	0	1
OASIS30002	1150	0	0	1
OASIS30002	1550	0	1	1

MERGED ASSESSMENT DATA

OASIS ID	Days From Entry	MOMDEM	DADDEM	MOMONSET	DECCLMOT	FRSTCHG	FTLDEVAL
OASIS30001	0	1	0	77	0	0	0
OASIS30001	150	1	0	77	0	0	0
OASIS30001	750	1	0	77	0	1	0
OASIS30001	1106	NA	NA	NA	1	1	0
OASIS30002	0	0	0		0	0	0
OASIS30002	751	1	0	84	0	0	1
OASIS30002	1150	1	0	84	0	0	1
OASIS30002	1550	1	0	84	0	1	1
OASIS30003	70	0	0		NA	NA	NA

Figure 3.4 An Example of Matching-Up Session Process

Data pre-processing is a critical component of data analysis. In the context of the OASIS-3 dataset, session matching is one of the most important steps in longitudinal analysis. By accurately aligning scan session data with assessment data, the progression of aging and dementia can be highlighted more effectively. This precision in data alignment paves the way for early detection and intervention strategies, contributing to better outcomes in dementia studies.

3.1.2 Feature Selection with Minimum Redundancy and Maximum Relevance (MRMR)

Feature selection is a crucial dimension reduction technique that identifies the most relevant features for a given task. By eliminating redundant variables, not only enhances the predictive performance of statistical models or machine learning algorithms but also expedites the model-building process. Therefore, careful attention is required in the feature selection process to ensure that the selected variables effectively serve the study's objectives. Although the literature offers numerous feature selection methods, identifying the optimal approach can be particularly challenging with longitudinal or repeated measures data, which often exhibit intra-class correlation structures.

In this thesis, the primary feature selection method employed is the parallelized Minimum Redundancy and Maximum Relevance (mRMR) method. mRMR was introduced by De Jay et al. in 2013 and it is a heuristic approach designed to identify optimal sets of relevant features in high-dimensional data. This method utilizes the Mutual Information Matrix (MIM) and the correlation coefficient between variables x and y.

$$I(x,y) = -\frac{1}{2}\ln(1-\rho(x,y)^2)$$

where I and ρ denote the MI and correlation coefficient between variables x and y, respectively. The choice of correlation coefficient depends on the type of variables:

Pearson's or Spearman's estimators for continuous variables, Cramer's V for discrete variables, and Somers' Dxy index for the correlation between continuous variables and survival data.

The method is considered y as an output variable and defines features as follows:

$$X = (X_1, \ldots, X_n)$$

The method ranks X by maximizing the MI with y and minimizing the average MI with all the previously selected variables. Covariates that have the highest MI are determined via the following equation:

$$x_i = arg \max_{x_i \epsilon X} I(x_i, y)$$

The set of selected features, denoted by S, is then initialized with x_i. Next, another feature is added to S by choosing the feature having the highest relevance with the output variable and the lowest redundancy with the previously selected features, thus maximizing the score q at step j

$$q_j = I(x_j, y) - \frac{1}{|S|} \sum_{x_k \epsilon S} I(x_j, x_k)$$

This step is repeated until the desired solution length has been reached.

Although MRMR mainly addresses challenges inherent in analyzing high-throughput genomic data such as microarray and next-generation sequencing, it can be used for survival data. In fact, it was adopted in several longitudinal studies (Chiu et al., 2022; Tylee et al., 2017). Therefore, based on its use in longitudinal studies, MRMR is employed in this thesis study to provide a more concise and interpretable set of predictors for modeling.

3.2 Marginal Models with Ordinal Response

The generalized estimating equation (GEE) method (Liang & Zeger, 1986) is a widely adopted statistical method for analyzing correlated longitudinal/clustered

data, often encountered in longitudinal or repeated measures studies. Due to the nature of the longitudinal data, not only measurements taken from the same subject over time are likely correlated but also traditional regression models' independence assumption is violated. The GEE approach's main advantage is allowing a population-averaged interpretation of the parameter of interest that requires correct specification of the mean structure of the response.

The primary focus of GEE has mostly been on continuous and binary responses, and therefore the literature is comparatively scarce for longitudinal ordinal responses (da Silva et al., 2019). Generalized estimating equations for ordinal outcomes require several aspects. The first aspect is to choose a model for the covariates and a non-linear link function to connect the model to the cumulative probabilities. Then the second aspect is to create a set of binary variables describing all possible outcomes for the ordinal observations (Clayton, 1992). The third aspect is to choose a working correlation matrix or working association structure to describe the possible association between all binary variables. The final aspect is the estimation method for the association parameters involved in the association structure.

3.2.1 Generalized Estimating Equation for Clustered Ordinal Responses

Clustered data frequently emerges in numerous applications, including longitudinal and repeated measure studies (Højsgaard et al., 2005). To utilize the GEE for clustered ordinal responses two approaches were highlighted in Heagerty and Zeger (1996). As one approach considers marginal cumulative probabilities for the ordinal responses, the other uses the pairwise association of pairs of observations from the same cluster. In that respect, the proportional odds model for the marginal means and model pairwise associations using the global odds ratio were applied for the ordinal response for clustered longitudinal data.

It was underlined that O_i is a vector of ordinal measurements for the i^{th} cluster and that O_{ij} represents the t^{th} observation in the i^{th} cluster. Let x_{it} denote the

covariates associated with O_{it}. The ordinal measure was stated as $O_{it} = c$, where $c \in [0,1,2,\dots,C-1]$ through a vector of cumulative indicator variables.

$$Y_{ijc} = I(O_{it} > c),$$

where $c \in [0,1,2,\dots,C-1]$. Thereby, the proportional odds model for the marginal means can be assumed as

$$logit[E(Y_{itc})] = \theta_c + x_{it}^T\beta$$

As an alternative approach, the proportional odds model can be derived by treating each possible binary observation of the ordinal response, $Y_{itc}, c = 0,1,\dots,C-1$ as a dichotomous variable at different levels. Simultaneously fitting logistic regressions for each pair of derived binary responses allows for the pooling of information, assuming parallel regressions as a function of the covariate for each binary response. This proportional odds assumption can be checked by fitting cutpoint-by-covariate interactions. Viewing the model as simultaneous logistic regressions introduces the odds ratio measure of association for observations within the same cluster. By doing that not only interfering neighbouring categories among them but also retaining the same interpretations for each binary pairs. In that respect, the global odds ratio for the pair of $O_{it}\ and\ O_{ik}$ is determined as

$$\varphi_{i(t,k)(c1,c2)} = \frac{P(O_{it} > c_1, O_{ik} > c_2)(O_{it} \le c_1, O_{ik} \le c_2)}{P(O_{it} > c_1, O_{ik} \le c_2)(O_{it} \le c_1, O_{ik} > c_2)}$$

The odds ratio, $\varphi_{i(t,k)(c1,c2)}$, for the ordinal response, the illustration of the marginal models is as follows:

$$\log\left(\varphi_{i(t,k)(c_1,c_2)}\right) = z_{i(t,k)(c_1,c_2)}^T\alpha$$

The covariates in z_{itk} be a subset of (x_{it}, x_{ik}) or any other relevant covariates for modeling the degree of association between the j and k observations.

3.2.2 Marginal Models for Correlated Ordinal Multinomial Responses

In various studies, the focus is on making inferences about the regression parameters of a marginal model for correlated, repeated, or clustered multinomial variables, whether they have ordinal or nominal response categories. While the association structure among the dependent responses is of lesser importance, the absence of a suitable multivariate distribution for multinomial responses and the sensitivity of traditional maximum likelihood methods to inaccuracies in the association structure has led various studies to extend the Generalized Estimating Equations (GEE) method of Liang and Zeger (1986) for multinomial responses. In that respect, significant contributions in this area have been made so far by Milleret al. (1993); Lipsitzet al. (1994); Lumley (1996); Heagerty and Zeger (1996); and Parsons et al. (2006). These adaptations solve the same estimating equations as Liang and Zeger (1986), but they vary in their parametrization and/or estimation of ρ, which typically describes a "working" assumption about the association structure. To be able to cover the GEE approach with the multinomial or ordinal response cases, the multgee R package has been implemented by Touloumis et al. (2013). This GEE approach addresses limitations in traditional GEE methods for handling correlated multinomial responses.

The theoretical background of the local odds ratios GEE approach with respect to longitudinal study with correlated ordinal multinomial responses is outlined below.

Let Y_{it} be the multinomial response for the subject i $(i = 1, \dots, N)$ at time t $(t = 1, \dots, t^*)$ that takes value in $\{1,2, \dots, C\}$, $C > 2$. Define the response vector for the subject i

$$Y_i = \left(Y_{it1}, \dots, Y_{i1(C-1)}, Y_{i21}, \dots, Y_{i2(C-1)}, \dots, Y_{it^*1}, \dots, Y_{it^*(C-1)}\right)^T$$

where $Y_{itc} = 1$ if the response for the subject i at time t falls at category c and $Y_{itc} = 0$ otherwise. Denote by x_{it} the covariate vector associated with Y_{it} and let $x_i = \left(x_{i1}^T, \dots, x_{it^*}^T\right)^T$ be the covariate matrix for the subject i . Define $\pi_{itc} =$

$E(Y_{itc}|x_i) = P(Y_{itc} = 1|x_i) = P(Y_{it} = c|x_i)$ as the probability of the response category c for subject i at time t, and let $\pi_i = \left(\pi_{i1}^T, \dots, \pi_{it^*}^T\right)^T$ be the mean vector of Y_i where $\pi_i = \left(\pi_{it1}, \dots, \pi_{it(C-1)}\right)^T$. It follows from the above that $Y_{itC} = 1 - \sum_{c=1}^{C-1} Y_{itc}$ and $\pi_{itC} = 1 - \sum_{c=1}^{C-1} Y_{itc}$.

The choice of the marginal models depends on the nature of the response scale. For ordinal multinomial responses, the family of cumulative link models

$$F^{-1}[P(Y_{it} \leq c|x_i)] = \beta_{0c} + \beta_*^T x_{it}$$

Where F is the cumulative distribution function of a continuous distribution and $\beta_{0c}: c = 1, \dots, C-1$ are the category-specific intercepts.

The category-specific intercepts need to satisfy a monotonicity condition $\beta_{01} \leq \beta_{02} \leq \cdots \leq \beta_{0(C-1)}$ only when the family of cumulative link models is employed.

3.3 Generalized Linear Mixed Effect Models

Generalized linear models (GLMs) serve as a comprehensive framework in statistical modeling, extending the concept of well-studied linear regression to accommodate non-normally distributed response variables by linking them to a linear combination of predictors through a link function, such as the logit, probit, or identity function. GLMs provide a flexible framework for analyzing various data types, from binary outcomes to count data, by specifying appropriate probability distributions and link functions.

However, the simplicity of GLMs comes with limitations, particularly when dealing with nested or hierarchical data structures where observations are not independent. This is where generalized linear mixed-effect models (GLMMs) step in, offering a powerful extension by incorporating random effects to model variability at multiple levels. In that respect, GLMMs combine the strengths of GLMs with the capacity to handle correlated data, making them invaluable in fields like biology, psychology, and ecology, where data often exhibit complex dependencies.

In GLMMs, fixed effects are evaluated similarly to those in GLMs, capturing the average relationship between predictors and the response variable. However, GLMMs introduce random effects, which account for variability across different levels of nesting, such as repeated measurements within individuals or sites. These random effects are typically assumed to follow a multivariate normal distribution, allowing for the estimation of both fixed and random effects simultaneously.

Furthermore, GLMMs offer flexibility in specifying the covariance structure among random effects, accommodating various correlation patterns within the data. This flexibility enables capturing the model complexity more accurately, which may be overlooked in traditional GLMs. Estimating parameters in GLMMs requires specialized approaches, such as restricted maximum likelihood estimation (REML), Maximum Likelihood Estimation (MLE) or Bayesian approaches, due to the increased complexity introduced by random effects. Additionally, model selection becomes more intricate, as researchers must balance model fit with parsimony while considering both fixed and random effects.

3.3.1 Generalized Linear Mixed Models by L1-Penalized Estimation

Throughout history, various methods have been proposed to tackle the high dimensional issues in longitudinal settings. As a result, variable selection has gained importance and led development of new penalization techniques and improving the existing ones in the field. In that respect, several approaches have been proposed so far such as SCAD penalty techniques (Ni et al., 2010), adaptive mixed Lasso methods (Wang et al., 2010), MLE-based L1 penalized estimation procedure (Schelldorfer et al., 2011).

The use of penalization techniques has been an ongoing process. Yet, in this thesis, a combination of L_1-penalty approaches and gradient ascent optimization with Fisher scoring suggested by Groll and Tutz (2014) are utilized. Even though Schelldorfer et al. (2011), which was covered the same problem, and was later extended by

Schelldorfer et al. (2014), these papers solely allow a few distribution families as well as simple structures for the variance-covariance matrix, which may not work for our cases. Therefore, Groll and Tutz's approach has been considered and triangulated here.

3.3.1.1 GLMM Formulation

To begin with, defining a proper model structure of GLMM is crucial. Let us assume that the dependent variable Y_{it} denote observation t in cluster i, $i = 1, \dots, N$, $t = 1, \dots, t^*$, collected in $y_i^T = (y_{i1}, \dots, y_{1t^*})$. Furthermore, let $x_{it}^T = (1, x_{it}, \dots, x_{it^*})$ be the covariate vector associated with fixed effects and $z_{it}^T = (z_{it}, \dots, z_{it^*})$ be the covariate vector associated with random effects. It is assumed that the observations y_{it} are conditionally independent with means $\mu_{it} = E(y_{it}|b_i, x_{it}, z_{it})$ and variances $var(y_{it}|b_i) = \phi v(\mu_{it})$, where $v(.)$ is a known variance function and ϕ is a scale parameter. The GLMM that we consider in the following has the form

$$g(\mu_{it}) = x_{it}^T\beta + z_{it}^T b_i = \eta_{it}^{par} + \eta_{it}^{rand},$$

where g is a monotonic and continuously differentiable link function, $\eta_{it}^{par} = x_{it}^T\beta$ is a linear parametric term with a parameter vector $\beta^T = (\beta_0, \beta_1, \dots, \beta_N)$ including intercept and $n_{it}^{rand} = z_{it}^T b_i$ contains the cluster-specific random effects. An alternative form that we also use is

$$\mu_{it} = h(\eta_{it}), \qquad \eta_{it} = \beta_0 + \eta_{it}^{par} + \eta_{it}^{rand},$$

where $h = g^{-1}$ is the inverse link function.

A closed presentation of the model (3) is obtained by using matrix notation. By collecting observations within one cluster, the model has the form

$$g(\mu_i) = X_i\beta + Z_i b_i$$

where $X_i^T = (x_{i1}, \dots, x_{it^*})$ denotes the design matrix of the i^{th} cluster and $Z_i^T = (z_{i1}, \dots, z_{it^*})$. For all observations one obtains

$$g(\mu) = X\beta + Zb$$

with $X^T = [X_1^T, \dots, X_N^T]$ and block-diagonal matrix $Z = Blockdiag(Z_1, \dots, Z_N)$. For the random effect vector $b^T = (b_1^T, \dots, b_N^T)$ one has a normal distribution with block-diagonal covariance matrix $Q_b = diag(Q, \dots, Q)$.

Focusing on GLMMs we assume that the conditional density of y_{it}, given explanatory variables and the random effect b_i, is of exponential family type

$$f(y_{it}|x_{it}, b_i) = \exp\left\{\frac{\left(y_{it}\theta_{it} - \mathrm{K}(\theta_{it})\right)}{\phi} + c(y_{it}, \phi)\right\},$$

where $\theta_{it} = \theta(\mu_{it})$ denotes the natural parameter, $\mathrm{K}(\theta_{it})$ is a specific function corresponding to the type of exponential family, $c(.)$ the log normalization constant and ϕ the dispersion parameter.

One popular method to maximize GLMMs is penalized quasi-likelihood (PQL), which has been suggested by Breslow and Clayton (1993) and Lin and Breslow (1996). Typically, the covariance matrix $Q(\varrho)$ of the random effects b_i depends on an unknown parameter vector ϱ. In penalization-based concepts, the joint likelihood function is specified by the parameter vector of the covariance structure ϱ together with the dispersion parameter ϕ, which are collected in $\psi^T = (\phi, \varrho^T)$, and parameter vector $\delta^T = (\beta^T, b^T)$. The corresponding log-likelihood is

$$l(\delta, \psi) = \sum_{i=1}^{N} \log\left(\int f(y_i|\delta, \psi)p(b_i, \psi)db_i\right),$$

where $p(b_i, \psi)$ denotes the density of the random effects. Breslow and Clayton (1993) derived the approximation

$$l^{app}(\delta, \gamma) = \sum_{i=1}^{N} \log\left(f(y_i|\delta, \psi) - \frac{1}{2}b^T Q(\varrho)^{-1}b\right),$$

where the penalty term $b^T Q(\varrho)^{-1}b$ is due to the approximation based on the Laplace method. PQL usually works within the profile likelihood concept. It is distinguished

between the estimation of δ, given the plugged-in estimate $\hat{\psi}$, resulting in the profile-likelihood $l^{app}(\delta, \psi)$, and the estimation of ψ.

3.3.1.2 Regularization in GLMM

In Groll and Tutz (2014), firstly the penalized log-likelihood term was addressed by including the penalty term in the log-likelihood function as follows:

$$l^{pen}(\beta, b, \psi) = \; l^{pen}(\delta, \psi) = \; l^{app}(\delta, \psi) - \lambda \sum_{i=1}^{N} |\beta_i|$$

where δ is the parameter vector, ψ is the joint likelihood function, which is a combination of the parameter vector of covariance structure and dispersion parameter. For given $\hat{\psi}$, the optimization problem of the above penalized log-likelihood function is reduced to

$$\hat{\delta} = \; arg \max_{\delta} l^{pen}\left(\delta, \hat{\psi}\right) = arg \max_{\delta} \left[l^{app}\left(\delta, \hat{\psi}\right) - \lambda \sum_{i=1}^{N} |\beta_i| \right].$$

A Full gradient algorithm is employed, based on the algorithm of Goeman (2010). As noted by Goeman (2010), the algorithm can be readily adapted for scenarios where some parameters should not be penalized. The penalty used in $\hat{\delta}$ equation can be seen as a partially penalized approach if the whole parameter vector $\delta^T = (\beta^T, b^T)$ is considered.

3.3.2 Cumulative Linear Mixed Models

Beforehand, cumulative link models (CLMs) have been called by different names based on the link function usage such as ordered logit models and ordered probit models (Greene & Hensher 2010); however, the name of the cumulative link models was mentioned in Agresti (2002). In the literature, the cumulative link model with the logit link function is usually referred to as the proportional odds model. Besides,

when a complementary log-log link function is used, it is sometimes known as a proportional hazards model for grouped survival times. Furthermore, Cumulative Link Mixed Models (CLMMs), which is an extended version of CLMs in multilevel data cases. The CLMMs have been used for estimating population-level intercepts and slopes or fixed effect parts as well as evaluating changes in these intercepts and slopes across some grouping factor or random effect part (Taylor et al., 2023).

To cover the basics of the cumulative linear mixed models, which are adapted from the cumulative link models while incorporating random effects in the predictor part, CLMMs with a single random effect part are illustrated in Christensen and Brockhoff's (2013) research as follows

$$Y_{itc} = F(\theta_t - x_{it}^T\beta - b_i)$$

where $i = 1, \dots, N$, $c = 1, \dots, C$ and $t = 1, \dots, t^*$. Besides, in order above model parameterization to be held random effects should be normally distributed and the conditional distribution of the observation given the random effects is multinomial.

Thereby, distributional properties can be illustrated as

$$(Y_{ic}|\beta_i = b_i) \sim Multinom(1, \pi_{ic}) \qquad B_i \sim N(0, \partial^2{}_b)$$

The β and θ parameters describe the structure in the conditional distribution of the observations, and ∂_b describes the heterogeneity in the population. This model has common properties as a normal linear mixed model where the response is treated as ordinal rather than normally distributed. If the inverse link function, F is taken to be the standard normal cumulative distribution function, this model corresponds to assuming the following linear mixed model for the latent scale:

$$S_{ic} = x_{ic}^T\beta + b_i + e_{ic} \qquad e_{ic} \sim N(0, \sigma^2)$$

This is possibly the simplest model for the latent scale that accounts for the grouping structure in the data.

The log-likelihood function for the models we consider may be written as

$$\ell(\theta, \beta, \sigma_b; y) = \sum_{i=1}^{N} log \int p(y_i|b_i)p(b_i)db_i$$

where $p(y_i|b_i)$ is the conditional multinomial probability mass function of the observations given the random effects, and $p(b_i)$ is the (marginal) normal density of the random effects. The log-likelihood is a sum of n independent contributions since observations from different individuals are assumed to be independent.

Estimation of CLMMs is complicated by the fact that the integral in eq. (11) does not have a closed-form solution. Several different approximations have been proposed and two of the most popular are the Laplace approximation and adaptive Gauss-Hermite quadrature (AGQ) (Skrondal & Rabe-Hesketh, 2004). The Laplace approximation is fast and reasonably accurate while AGQ is computationally more intensive, but it has the advantage that the accuracy can be increased by adding more quadrature nodes. Often the Laplace approximation is sufficiently accurate while essentially exact estimates can often be obtained from AGQ with a few, e.g. 5–10 nodes. Following Joe (2008) we recommend that the Laplace approximation is used initially; the final model may be estimated accurately with AGQ by increasing the number of nodes until the parameter estimates do not change by any relevant amount.

The Laplace approximation and AGQ are implemented in the R-package ordinal (Christensen, 2015) for CLMMs, and AGQ is also implemented in the NLMIXED procedure for SAS (Inc., 2008).

3.4 Bayesian Generalized Linear Mixed Effect Models

As a part of this thesis study, Bayesian Generalized Linear Mixed-Effect Models were also covered. Bayesian GLMMs offer a robust and flexible approach to analyzing hierarchical or nested data structures by integrating the principles of Bayesian statistics into the framework of generalized linear mixed-effect models (GLMMs). Unlike the frequentist approach, which relies on point estimates and asymptotic properties, the Bayesian paradigm provides a coherent framework for

incorporating prior information and directly estimating the posterior distributions of model parameters.

At the core of Bayesian GLMMs is the use of prior distributions, which represent the prior beliefs about the parameters before observing the data. These priors can be non-informative, reflecting minimal prior knowledge, or informative, incorporating existing domain knowledge. While working with limited data or the presence of missing observation, the choice of priors can significantly influence the results.

Bayesian GLMMs extend the flexibility of GLMMs by not only modeling fixed and random effects but also providing a full probability distribution for each parameter, leading to more nuanced uncertainty quantification. In Bayesian GLMMs, fixed effects describe the average relationship between predictors and the response variable, while random effects account for variations at different hierarchical levels, such as individual differences in longitudinal studies.

The process of fitting a Bayesian GLMM involves deriving the posterior distributions of the parameters using Bayes' theorem, which combines the likelihood of the observed data given the model parameters with the prior distributions. This results in a posterior distribution that incorporates both the data and prior information. Markov Chain Monte Carlo (MCMC) methods, such as the Gibbs sampler and the Metropolis-Hastings algorithm, are typically employed to approximate these posterior distributions, especially in complex models where analytical solutions are infeasible.

One of the key advantages of Bayesian GLMMs is their capacity to handle complex data structures and missing data more naturally than frequentist approaches. Bayesian inference provides a principled way to incorporate all sources of uncertainty, including uncertainty about model parameters and predictions. This may be particularly valuable in longitudinal datasets, where data may be missing at random intervals.

Furthermore, Bayesian model comparison tools, such as the Deviance Information Criterion (DIC) and the Widely Applicable Information Criterion (WAIC), allow for rigorous model evaluation and selection. These criteria assess model fit while penalizing for complexity, ensuring a balance between goodness-of-fit and parsimony. Bayesian model averaging can also be used to account for model uncertainty by averaging over multiple models, weighted by their posterior probabilities.

The interpretability of Bayesian GLMMs is enhanced through the direct probability statements about parameters and predictions. For instance, credible intervals provide a range within which the parameters are likely to lie with a certain probability, offering a more intuitive understanding of the uncertainty compared to frequentist confidence intervals.

3.4.1 Monte Carlo Markov Chain for Generalized Linear Mixed Effect Models

In GLMMs, the modeling framework has been generally designed to provide various ranges of data. Yet, for non-Gaussian distribution response cases, the likelihood function is not possible to be driven in a closed form. In that respect, Markov Chain Monte Carlo methods become useful to overcome this issue by sampling from a series of conditional distributions. Therefore, it follows the general GLMMs model formulation from the Bayesian perspective In Hadfield (2010), it was demonstrated that the model consists of three components, which are the probability density function of data y to latent variable l on the link scale, the linear mixed effect model with respect to l and variance structures that illustrate the variances between mixed effects, respectively.

The probability density function of data y to latent variable l of the i^{th} observation can be formulated as follows

$$f_i(y_i|l_i)$$

If y_i assumed to be an ordinal probit model with more than two categories, which is also called threshold or ordinal models, f has an additional parameter to the latent variable l. This additional parameter is the $\max(y) + 1$ cutpoints γ. Thus, the probability density function of f_i is translated into f_o, which refers to ordinal models, as follows:

$$f_O(y_i|l_i, \gamma) = F_N(\gamma_{y_i}|l_i, 1) - F_N(\gamma_{y_i+1}|l_i, 1).$$

In above formulation F_N is the cumulative density function for the normal.

The vector of a latent variable can be predicted by

$$l = X\beta + Zu + e$$

where X is the design matrix of the fixed effect part, whereas Z is the design matrix of the random effect part. Design matrices of X and Z have parameter vectors of β and u, and e stands for the residual vector.

Parameter vectors of predictors (both fixed and random effects) along with the residual vectors are assumed to follow a multivariate normal distribution. Hence, the covariance structures of the model parameter can be shown as

$$\begin{bmatrix} \beta \\ u \\ e \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} B & 0 & 0 \\ 0 & G & 0 \\ 0 & 0 & R \end{bmatrix}\right)$$

In the above model structures, β_0 stands for the prior means for the fixed effects with the prior covariance matrix B, while G and R are the covariances of the random effects and residuals, respectively. Besides, zero off-diagonal matrices imply a priori independence between matrix B, G and R. G and R are generally unknown and they are supposed to be estimated from the data.

The software application of the study, which MCMCglmm package in R, allows the following variance structure as well:

$$G = (V_1 \otimes A_1) \oplus (V_2 \otimes A_2) \oplus \ldots$$

where the parameter (co)variance matrices V are low-dimensional covariance matrices and A are high-dimensional structured matrices. As opposed to the matrices V, matrices of A are treated as known. While $\otimes$ denotes the Kronecker product, $\oplus$ refers to direct sums.

In the parameter estimation phase, latent variables and covariances should be updated. Since the latent variable l is not in a suitable form, Metropolis-Hastings or slice sampling method should be utilized to be updated. By using the Metropolis-Hasting algorithm latent variable l is sampled and during the burn-in phase proposed distribution has been determined via adaptive methods, which are triangulated in Haario et. al. (2001). Mixed effect model parameters, which are β and u, are estimated through Gibbs sampling. Lastly, having conjugate priors in the framework, covariance structures, G and R are also updated through Gibbs sampling since they are distributed as inverse-Wishard distribution. If that is not the case, instead of using Gibbs sampling, Metropolis-Hasting updates can be utilized to update the covariance structures. In ordinal models, introduced parameters and cutpoints, should be updated as well. For that reason, the method that has been proposed by Cowles and Carlin (1996) has iteratively updated the latent variable l and cutpoints γ by utilizing Hasting-with-Gibbs approach.

3.4.2 Bayesian Generalized Linear Multivariate Multilevel Models

In certain cases, Bayesian approaches has been superior to frequentist approaches as integrating the prior belief regarding parameters into the model, yet due to complexity of the posterior distribution analytical solution may not be driven. However, this issue has been overcome by utilizing MCMC to draw samples from the posterior distribution. MCMCglmm is one of the proposed algorithms in this context via applying Metropolis-Hasting updates and Gibbs sampling. Nevertheless, according to Gelman et. al. (2013), these algorithms have slow convergence while working with high-dimensional settings with correlated parameters. As opposed to Metropolis-Hasting and Gibbs sampling, brms package utilizes Hamiltonian Monte

Carlo (HMC) and its extension version No-U-Turn Sampler (NUTS) to address these drawbacks.

The model formulation is similar to the generalized linear mixed effect model, including fixed and random effects design matrices. Main distinguish between MCMCglmm and brms is that the way of modeling the covariance matrices. Inverse-Wishard prior is usually utilized in MCMCglmm and it is stated that conjugate of this prior distribution would lead to better Gibbs samplers' results (Gelman et al., 2004). Conversely, prior selections do not have to be conjugate in the NUTS sampler. Owing to HMC and NUTS, covariance matrices, V_k, can be determined by correlation matrix Ω_k and a vector of standard deviation σ_k as follows

$$V_k = D(\sigma_k)\Omega_k D(\sigma_k)$$

where $D(.)$ is equivalent to a diagonal matrix with diagonal elements σ_k. Later, prior selection is determined based on the LKJ correlation prior with $\varsigma > 0$ for correlation matrix (Lewandowski et al., 2009).

$$\Omega_k \sim LKJ(\varsigma)$$

Based on the value of a parameter of ς the prior distribution is shaped. By default, ς is equal to 1, meaning that the density is uniform.

When different levels of the grouping factor cannot be adopted to be independent, then the covariance matrix structure becomes equivalent to the MCMCglmm variance structure.

$$\sum\nolimits_k = V_k \otimes A_k$$

as previously stated in the MCMCglmm section, A_k the covariance matrix, which is known, and $\otimes$ is the product of Kronecker.

3.5 Machine Learning Models

This thesis explores various machine learning algorithms explicitly tailored for the analysis of longitudinal data with ordinal responses. Unlike traditional statistical models, machine learning algorithms offer robust frameworks for handling large, complex datasets with intricate dependencies. However, the application of these methods to ordinal response data, particularly within a longitudinal context, presents unique challenges and limitations that must be addressed to ensure accurate and reliable analysis.

Hyperparameter tuning is a critical step in the development of machine learning models. It involves selecting the optimal values for hyperparameters, which are pre-set parameters controlling the model's behavior and performance. These hyperparameters are not learned from the data but are chosen through an iterative process that involves training the model with different hyperparameter settings and evaluating its performance using a validation set. Therefore, models require careful tuning of their specific hyperparameters to enhance their predictive accuracy. The hyperparameters that yield the highest performance metrics, such as classification accuracy or other relevant metrics for ordinal data, are selected as the optimal values. After determining the optimal hyperparameters, the final models are retrained on the entire training dataset to fully capture the underlying patterns and relationships.

Despite their strengths, machine learning models face significant limitations when dealing with ordinal response data. Ordinal data, characterized by categories with a natural order but no fixed interval between them, require specialized techniques to preserve this order information. Standard machine learning models may treat ordinal outcomes as either nominal or continuous, leading to the loss of valuable information about the ordinal nature of the data. Techniques such as ordinal encoding or specialized loss functions are necessary to properly account for the order of categories.

Moreover, the longitudinal aspect adds another layer of complexity, as it involves repeated measurements over time for the same subjects. This nested data structure can lead to correlated observations, violating the assumption of independence in traditional machine learning algorithms. While mixed-effect models attempt to address this by incorporating random effects to account for within-subject correlation, they still face challenges in accurately modeling the temporal dependencies and the ordinal nature of the response.

Machine learning models also tend to struggle with interpretability, particularly in the context of longitudinal ordinal data. While these models can effectively capture complex, non-linear relationships, their "black-box" nature makes it difficult to understand the underlying mechanisms driving predictions. This lack of transparency can be a significant drawback, especially in fields where interpretability is crucial for making informed decisions based on the model's outputs.

Additionally, hyperparameter tuning in the context of ordinal longitudinal data can be more intricate. Performance metrics must be chosen carefully to reflect the ordinal nature of the response variable. Metrics that account for the order of categories are more appropriate than standard metrics.

In summary, while machine learning algorithms offer powerful tools for analyzing complex longitudinal data, their application to ordinal response data requires careful consideration of the unique characteristics of ordinal outcomes and the longitudinal structure. The integration of specialized techniques to handle the ordinal nature and the development of more interpretable models are essential for leveraging the full potential of machine learning in this context. Despite these challenges, innovative approaches like the Ordinal Mixed-Effects Random Forest are paving the way for more effective and insightful analysis of longitudinal ordinal data.

3.5.1 Ordinal Mixed Effect Random Forest

Various tree-based mixed effect models exist that integrate tree-based approaches into mixed effect models framework (Fokkema et al., 2018), yet few implementations have been developed for ordinal response models. Initially, to handle the hierarchical structure of data of the exponential family, a Generalized Mixed Effect Random Forest (GMERF) was proposed by Pellagatti et al. (2021). As an addition, the same authors proposed a novel algorithm, which is named Ordinal Mixed Effect Random Forest (OMERF), specifically designed for ordinal response data. OMERF takes CLMMs, which was proposed by Tutz and Hennevogl (1996) into account and combines it with the ordinal random forest approach (Hornung, 2020) to set ground for the tree-based method for nested data and ordinal responses.

A generic formulation of the CLMMs with a random intercept and θ random slopes can be expressed as

$$\eta_{itc} = g(\mu_{itc}) = \theta_c - (x_{it}^T\beta + z_{it}^T b_i)$$

$$b_i \sim N_\theta(0, \Sigma_b)$$

where $c = 1, \dots, C$, which is the ordinal response category, $i = 1, \dots N$ is the level of group index and $t = 1, \dots, t^*$ is the nested level index. Additionally, g is a monotonic link function, which is also differentiable. θ_c can be referred to as ordered thresholds or cutpoints.

$$\mu_{itc} = P(y_{it} \leq c) = \pi_{it1} + \cdots + \pi_{itc}$$

is the cumulative probability up to c^{th} category for the unit t in the group i.

Moreover, the ordinal response variable y_{it}, which has a number of C categories, is written in terms of the latent variable y_{it}^* with a set of $C - 1$ cutpoints. Therefore, CLMMs via the latent variable can be written as

$$y_{it}^* = x_{it}^T\beta^* + z_{it}^T b_i^* + \epsilon_{it}^*$$

$$e_{it}^* \sim N(0, \sigma_{\epsilon^*})$$

As a consequence, the cumulative probabilities can be found as $\gamma_{itc} = P(y_{it} \leq c) = P(y_{it}^* \leq \theta_c^*) = P(\epsilon_{it}^* \leq \theta_c^* - \pi_{it}^T \beta^* - z_{it}^T b_i^*) = g^{-1}(\theta_c - \pi_{it}^T \beta - z_{it}^T b_i)$.

In the context of CLMMs, parameter estimation is usually done via utilizing maximum likelihood methods or applying several approximation approaches such as Laplace approximation, Adaptive Gaussian Quadrature, Gauss-Hermite Quadrature or Non-Adaptive Gauss-Hermite Quadrature. The main difference among these approximation approaches comes from the computational time versus accuracy trade-off. Besides, based on the number of random effect terms available some approximations may not be valid. Besides, the Newton-Raphson algorithm is utilized to update the random effect part for the following approximations. Lastly, non-linear optimization techniques are constructed on the fixed effect part.

The proposed statistical model, which is OMERF, combines both CLMMs and random forest algorithms to model the random and fixed effect parts, respectively. The novel approach to be able to handle the ordinal response can be illustrated as

$$\eta_{itc} = g(\mu_{itc}) = \theta_c - (f(x_{it}) + z_{it}^T b_i)$$

$$g(\mu_{itc}) = logit(\mu_{itc}) = \ln\left(\frac{\mu_{itc}}{\mu_{itc} - 1}\right)$$

$$\mu_{itc} = P(y_{ic} \leq c)$$

$$b_i \sim N_\theta(0, \Sigma_b)$$

As seen from the above formulation, $f(x_{it})$ is the fixed part where it was estimated via random forest algorithms and μ_{itc} is the cumulative probabilities. The OMERF algorithm also allows offset, which does not require any parameters to be estimated from the data and is just a component of random forest estimates. When the offset variable is used, the formula becomes as

$$\eta_{itc} = g(\mu_{itc}) = \theta_c - (z_{it}^T b_i + offset_{it})$$

As η_{itc} cannot be directly calculated from the data, the OMERF algorithm considers ordinalforest R package (Hornung, 2020) approach, which is initially fitting fixed

covariates with respect to ordinal responses. Then, the cumulative probabilities are calculated based on this modeling. It follows by applying the inverse link function to initialize the unknown systematic component η_{itc}.

3.6 Model Evaluation Metrics

Evaluating the performance of trained machine learning and statistical models is a critical component of predictive analysis. This evaluation is essential for understanding how well a model can perform on new, unseen data, typically represented by a test set. To gain a comprehensive understanding of the model's performance, various metrics must be employed. By analyzing multiple performance metrics, we can identify the model's strengths and weaknesses and make necessary adjustments. Relying on a single metric, such as accuracy, can be misleading and may not accurately reflect the model's true predictive capabilities. For example, a model with high accuracy may still exhibit poor recall, failing to correctly identify positive instances. In some applications, particularly those involving imbalanced datasets, metrics such as recall and precision are more critical than overall accuracy. Therefore, evaluating multiple metrics helps to achieve a balance between precision, recall, and other performance measures, ensuring that the model generalizes well and provides reliable predictions on new data.

3.6.1 Confusion Matrix

The confusion matrix is a fundamental tool for evaluating the performance of classification models. It provides a detailed breakdown of the model's predictions and their alignment with actual classes in the dataset.

In a binary classification problem, the confusion matrix has a dimension of 2×2 (Figure 4.5.1.a) where one of the labels is considered to be “Positive” and the other one “Negative”. The matrix elements, which are titled true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), are categorized based

on the predicted and actual class labels. On the contrary, for multi-class classification problems, the confusion matrix has a dimension of CxC where C is the number of different class labels $C_1, \dots, C_C$ (In this thesis, there are three categories, C is equal to 3). The characterization of TP, TN, FP, and FN in a multiclass classification problem can be done via focusing on a specific class (Figure 4.5.1.b). By doing so, all performance metrics can be computed and then the whole confusion matrix can be referred to by utilizing the proper combination of these metrics.

- True Positive (TP): The number of correct positive predictions by the model.
- True Negative (TN): The number of correct negative predictions.
- False Positive (FP): The number of incorrect positive predictions (actual negative but predicted positive).
- False Negative (FN): The number of incorrect negative predictions (actual positive but predicted negative).

The confusion matrix examples for both binary and multiclass classification problems are displayed in Figure 3.5.

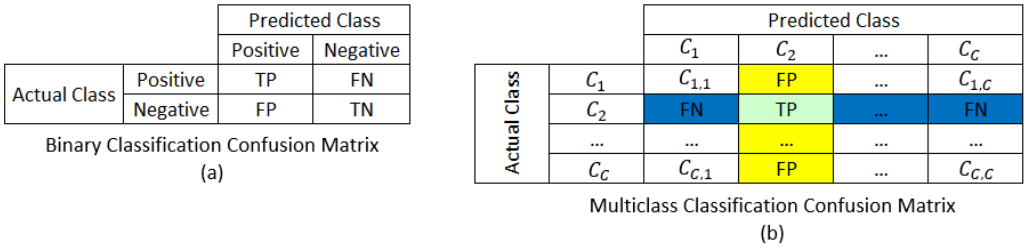


Figure 3.5 Confusion Matrix examples. (a) Binary Classification problem confusion matrix. (b) Multiclass Classification problem confusion matrix (Markoulidakis et al., 2021)

When assessing classification model performance, accuracy, sensitivity (also known as recall or true positive rate), and specificity are crucial metrics. Accuracy measures the proportion of correct classifications from the total number of instances. Nevertheless, accuracy may require careful examination for imbalanced datasets where one class is more frequent.

For clarification, P stands for summation of True Positive and False Negative, while N stands for summation of True Negative and False Positive.

$$Accuracy_c = \frac{TP_c + TN_c}{P_c + N_c}$$

Sensitivity, or recall, measures the proportion of correctly identified positive instances out of all positive instances in the dataset by focusing on a specific category. This metric indicates how well the model captures positive instances.

$$Sensitivity\ _c(Recall_c) = \frac{TP_c}{P_c}$$

Specificity measures the proportion of correctly identified negative instances out of all actual negative instances in the dataset by focusing on a certain category, indicating how well the model captures negative instances.

$$Specificity_c = \frac{TN_c}{TN_c + FP_c}$$

The F1 score is the harmonic mean of precision and recall. Precision is the proportion of true positive predictions out of all positive predictions. The F1 score provides a balanced measure since it considers both the false positives and false negatives.

$$F1\ Score_c = 2\ \times \frac{Precision\ _c \times\ Recall_c}{Precision_c + Recall_c}$$

where,

$$Precision_c = \frac{TP_c}{TP_c + FP_c}$$

While accuracy provides an overall view of the model's performance, sensitivity and specificity offer insights into its ability to distinguish between positive and negative instances. However, there is often a trade-off between sensitivity and specificity. Increasing sensitivity may decrease specificity and vice versa. The appropriate balance depends on the specific problem and the costs associated with false positives and false negatives. All these metrics range from 0 to 1, with values close to 1

indicating good model performance and low prediction error, while values close to 0 suggest poor performance. High accuracy implies the model correctly predicts a large proportion of instances, high precision indicates that the model's positive predictions are mostly correct, high recall shows the model effectively captures most positive instances and a high F1 score suggests a good balance between precision and recall.

3.6.2 Akaike Information Criterion

The Akaike Information Criterion (AIC), developed by Hirotugu Akaike in 1994 based on information theory, is a tool used in statistics and machine learning for model selection and comparison. AIC helps compare different models fitted to the same data, aiding in selecting the best model that balances goodness of fit and model complexity. AIC is employed in maximum likelihood estimation (MLE) and is based on minimizing information loss when a model approximates the true data-generating process.

The AIC formula is as follows:

$$AIC = -2 \times \ln(L) + 2 \times k$$

where k represents the number of model parameters.

AIC penalizes more complex models with more parameters, preventing overfitting where a model fits the training data well but may not generalize to new data. A lower AIC value indicates a more favorable model, providing a better trade-off between fit and complexity. AIC is used in various statistical modeling techniques, including linear regression, generalized linear models, and more complex machine learning models, helping avoid overfitting and obtaining more reliable and generalizable models.

3.6.3 Bayesian Information Criterion

The Bayesian Information Criterion (BIC), also known as the Schwarz criterion, is a model selection criterion used in statistics and machine learning, similar to AIC. BIC is based on maximizing the posterior probability of a model given the data and penalizes complex models more heavily, making it useful for small sample sizes or when overfitting is a concern.

The BIC formula is:

$$BIC = -2 \times \ln(L) + k \times \ln n$$

where k represents the number of model parameters and n stands for the sample size.

The penalty term in BIC increases with the sample size, favoring simpler models as the sample size grows. When comparing models, a lower BIC value is preferred, indicating a better balance between goodness of fit and complexity. Like AIC, BIC is commonly used in various statistical modeling techniques, including linear regression, generalized linear models, and machine learning models, providing reliable predictions on new data by effectively avoiding overfitting.

CHAPTER 4

DATA ANALYSIS

4.1 Design of Study

Analysis of longitudinal or repeated measure studies usually centers around classical methods like generalized estimating equations and random effects models due to the subject-specific and repeated structure of the observations. In addition, various machine learning methods are being continuously developed to handle longitudinal data structures and identify complex patterns within them. To address these challenges, this paper proposes a novel explainable AD prediction model using a multimodal dataset. This approach integrates clinical data, MRI segmentation data, and psychological data at the data level. Yet, despite ongoing research, there is still limited understanding of multi-class classification for AD. Therefore, this study aims to explore and compare appropriate methods for analyzing ordinal longitudinal data with missing multimodal data.

4.1.1 Introduction to Open Access Series of Imaging Studies (OASIS) Project

The Open Access Series of Imaging Studies (OASIS) initiative is a project that aims to provide the scientific community with unrestricted access to comprehensive neuroimaging datasets of the brain. Within the OASIS Brains project, multiple datasets serve various research objectives. The initial study, which was released in 2007, is OASIS-1 and it is a cross-sectional dataset. The second one is a longitudinal study designed for hypothesis-driven longitudinal data analysis. OASIS-3 is a longitudinal multimodal neuroimaging, clinical, cognitive, and biomarker dataset for

normal aging and AD. Lastly, OASIS-4, a clinical cohort, was evaluated for individuals who have experienced memory complaints.

The brain dataset used in this study was obtained from the OASIS-3 (Longitudinal Multimodal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer's Disease) dataset. OASIS-3 consists of 1378 participants that were collected across several ongoing projects through the Washington University of School of Medicine in St. Louis (WUSTL) Knight Alzheimer Disease Research Center (ADRC) over the course of 30 years. Participants include 755 cognitively normal adults and 622 individuals at various stages of cognitive decline ranging in age from 42-95 years. All participants were assigned a new random identifier and all dates were removed and normalized to reflect days from entry into the study. The dataset contains MR sessions, which are produced through Freesurfer processing. PET imaging scans and post-processed files can be accessed in OASIS-3. A cross-sectional TAU PET dataset for 451 subjects was also introduced and available to access in OASIS-3.

To identify the main factors leading to AD and its progression multiple sources have been considered non-imaging assessments such as cognitive, functional, physical, and behavioral assessments, along with Freesurfer outputs and Centiloid amyloid values have been evaluated to shed light on these research questions.

4.2 Data Pre-Processing

Various non-imaging datasets are used in this thesis which were obtained through Central XNAT. As a result of matching up sessions and merging eighteen different datasets, which were composed of various assessment scores, MRI and PET Scan sessions, the merged data was obtained. In the merged dataset, subjects have been monitored multiple times during their visits, yet their occurrences are unbalanced, which creates an irregularity in the dataset. Owing to these circumstances several anticipations have been taken before analyzing the data. First, due to the session matching-up process, some features resulted in having no observation. Hence, to

decrease the redundancy across the dataset, these variables and the unnecessary features are removed from the dataset. As covered in the Session Matching-Up and Merging section, the visit period was shaped by the Days From Entries variable, and within 365 days time period was considered. However, within this period multiple visits are observed. To keep the longitudinal structure in the dataset, a record that is the most recent one is kept among records that have multiple visits within the same period. By doing so, there is only one record within the visit period preserved within the cluster of subjects.

As stated in OASIS-3: Imaging Methods and Data Dictionary, to have a robust comparison across different region volumes, the volumes should be corrected for head size (LaMontagne et al., 2019; Buckner et al., 2004). Moreover, these correction steps are solely applicable to the volumes of the brain, rather than cortical thickness as they are found to be ineffective based on the head size.

Table 4.1 Normalization of MRI FreeSurfer-derived cortical volumes

	Steps Description
1	Compute the mean IntraCranial Volume (ICV) for the sample
2	Construct regression with ICV as an independent variable and Region of Interest (ROI) as a dependent variable to calculate B-weights, which are unstandardized coefficient values
3	Calculate the following formula for normalized volumes of ROIs: $Normalized\ Value\ of\ ROI\ Volume = Raw\ Value\ of\ ROI\ Volume - \left(B\text{-}weights * (Single\ Subject's\ ICV - Mean\ Value\ of\ ICV\ )\right)$

The above procedures are iteratively calculated for each subcortical and cortical ROI volume. If there is a change in participants within the sample, the steps should be re-run.

Additional regional calculations have been performed along with these calculations. In that respect, multiple global brain measures such as Whole Brain Volume, Cortex, Cortical White Matter, Sub Cort Gray and Total Ventricular Volume have been

added to the merged dataset by considering suggested calculations. The merged dataset has numerous categorical variables, hence encoding and decoding processes took place to maximize the information gain from the data. However, nominal variables in the merged dataset are kept as they are, since the one-hot encoding technique creates an issue called the dummy variable trap, which leads to the multicollinearity problem.

Since nearly one thousand features are available in the merged dataset, several dimension reduction techniques have been considered in the pre-processing phase. Due to the merging process of the different datasets in which the process was covered in Session Matching-Up and Merging, to avoid creating artificial observation an additional precaution had been taken, and as a result, features ended up having missing observations. Despite several approaches working on longitudinal or repeated measure studies with missing observations, most of the applications or packages that have been developed may not allow missing observations in their design. For that reason, an intuitive imputation approach has been utilized to minimize the features to work with while maximizing the information gain predicting the AD underlying factors.

Firstly, the missing rate has been considered. In that respect, variables whose missing rate is greater than 0.50 were removed from the dataset regardless of whether it is assumed to be a major underlying factor in predicting AD.

Secondly, the variance threshold technique, one of the feature elimination techniques called the filter method (Chandrashekar & Sahin, 2014), was introduced as an initial elimination approach. It removes variables that have variances below the predetermined threshold value. As a drawback of this approach, it does not account for the relationship between the variables and the response variable. Furthermore, having zero or less variation, meaning that all observations are more or less the same within the variable itself, may not provide valuable information to the response. Therefore, variables with zero or near zero variance were removed from the dataset by applying the nearZeroVar function in the caret package in R.

Thirdly, after eliminating the variables that have more missing observations than the non-missing observations and do not show any variability, the rest of the data has been divided into time-invariant and time-variant variables among missing ones. The main reason for this separation is to identify the type of imputation techniques used in the missing process. For the time-invariant features, a custom function, which utilizes The Last Observation Carried Forward (LOCF) approach, was written to be able to make a complete case for features like Gender, Hand, Race and so on. For this purpose, the na.locf function in the zoo package was used in R. For the time-variant features, initially, a predictor matrix was created using the variable selection procedure that was outlined in Van Buuren et al. (1999). The type of correlation along with the minimum threshold against which the absolute correlation and minimum threshold for the proportion of usable cases were specified as Spearman correlation, 0.4 and 0.3, respectively. The longitudinal structure was taken into account, thereby the predictor matrix was initialized by defining the index cluster and making a longitudinal model specification for the missing imputation. Later, a function called find.defaultMethod was run which systematically identifies and suggests multiple imputation techniques suitable for the variable considering the type of variable as well as the longitudinal structure of the data. Even though, "2l.glm.bin" was proposed for binary categorical covariates, it was replaced by "2l.jomo" for better imputation output. Likewise, "2l.glm.norm" was suggested for continuous variables, yet "2l.latentNormal" was used. Finally, multiple imputation by chained equations (mice) function was called setting maximum iteration to 5. Various imputation methods have utilized the mice package in R, along with its multilevel data extension, micemd.

To make a reliable interpretation of the features used in the model, all of the numeric covariates should be on the same scale. Thus, mean scaling and centering procedures were applied to the imputed complete dataset to improve both statistical models and machine learning algorithms' performances.

After imputing missing observations and scaling features, feature selection with the MRMR algorithm was employed to identify important features in a dataset. The

algorithm is similar to the Correlation-based Feature Selection (CFS) approach, which is based on maximizing the subject features' relevance with respect to the response variable while minimizing redundancy among other subject features. The relevance degree of features is computed using the F-statistics. The technique was initially applied on high-throughput genomic data, hence application of it does not specifically account for longitudinal structure. However, several cohort and longitudinal studies (Chiu et al., 2022; Tylee et al., 2017) in the literature used the MRMR approach to identify important features from the data.

In this thesis, a two-step feature selection process utilizing the MRMR algorithm was employed. The first phase involved categorizing the imputed dataset into distinct groups based on the characteristics of the original variables. For instance, the variables such as APOE-e4 Status, Gender, Hand, Father's Dementia Status and Mother's Dementia Status features were grouped as demographic variables as previously determined in the data sharing portal XNAT Central. The primary motivation for this step was to prevent the elimination of variables that might be crucial predictors of the response variable, given the large number of variables in the imputed dataset. By grouping variables with similar characteristics, the main goal was to select variables that maximize relevance with respect to the response variable. Doing so, not only ensures the inclusion of significant predictors but also allows for a comprehensive evaluation of various factors potentially linked to AD. In the second phase, the features selected in the first step of the MRMR algorithm from different datasets were merged, and a second step of feature selection was applied to this combined dataset.

To have a better understanding, the feature selection approaches are illustrated in Figure 4.1.

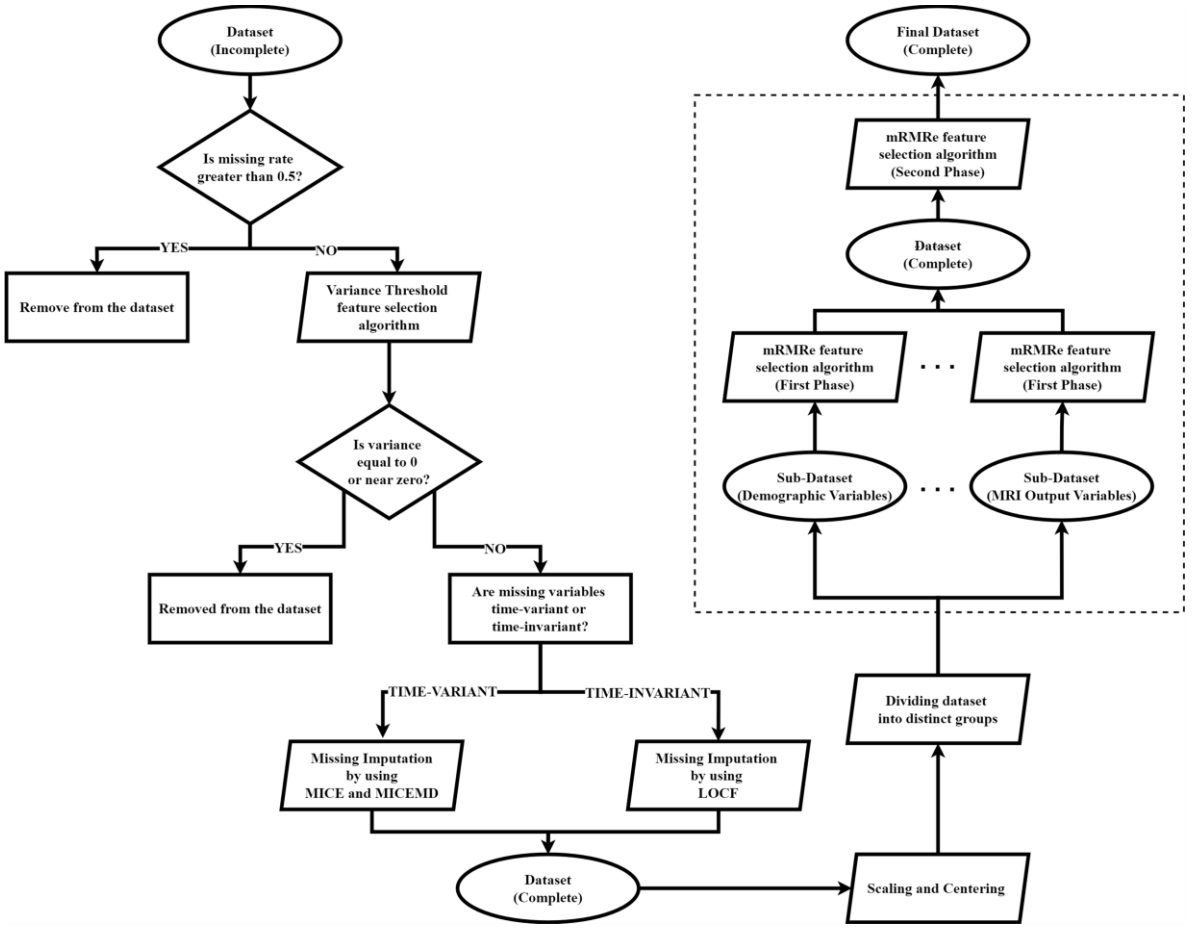


Figure 4.1 Flowchart of utilized feature selection approaches

4.3 Data Overview

In the final dataset, there are a total of 1764 observations and 73 features. Since there are vast amounts of features available in the dataset from demographic information to findings of different assessments, this thesis allows us to explore the relationships between key factors simultaneously and to appreciate other perspectives regarding disease prediction. Moreover, evaluating various underlying factors simultaneously might reveal potential risks associated with AD.

Table 4.2 provides detailed descriptions, classes, and definitions of the demographic variables in the dataset. Although most were excluded from the final dataset due to data pre-processing, they are included here to offer a brief overview of the subject's demographics.

Table 4.2 Demographic variables of the OASIS-3 dataset

Variable Name	Variable Type	Description
OASIS ID	Numeric	Subject's ID number
Days From Entry	Numeric	Visit day
CDR Group	Ordinal	Severity level of dementia status
Age at Entry	Numeric	Age when first entry occurred
Age at Death	Numeric	Age when death of subject occurred
Gender	Factor	Gender of subjects
Education	Factor	Education level of subjects
Social Economic Status	Factor	Socioeconomic status of subjects
Race	Factor	Race of subjects
Ethnic	Factor	Ethnicity of subjects
American Indians and Alaska Natives (AIAN)	Factor	Whether subject is AIAN
Native Hawaiian or Pacific Islander (NHPI)	Factor	Whether subject is NHPI
ASIAN	Factor	Whether subject is Asian
African American	Factor	Whether subject is African American
White	Factor	Whether subject is White
Mom Demented	Factor	Whether subject's mom has been diagnosed with dementia
Father Demented	Factor	Whether subject's father has been diagnosed with dementia
Hand	Factor	Handedness
APOE-e4 Status	Factor	Whether subject has APOE-e4 Gene

The dataset contains 590 subjects and in total for subjects where the maximum visit number is equal to 8. The response variable was categorized considering the Clinical Dementia Rating (CDR) score of the participants to build an ordinal scale structure.

Moreover, the response variable is determined based on the Global Clinical Dementia Rating (CDRTOT) and Clinical Dementia Rating Scale sum of boxes (CDRSUM) scores of participants which are one of the core assessment results in AD studies. For the categorization of dementia levels, different scales can be used for various research objectives, yet the most well-used categorization can be seen in Table 4.3 (O'Bryant et al., 2008; Morris, 1993).

Table 4.3 Dementia categorization of CDR scores

Staging Category	**CDRTOT**	**CDRSUM**
No Dementia	0	0
Very Mild Dementia	0.5	0.5-4.0
Mild Dementia	1	4.5-9.0
Moderate Dementia	2	9.0-15.5
Severe Dementia	3	16.0-18.0

Based on the intervals provided in Table 4.3, the distribution of the response variable was calculated, resulting in three dementia categories. Due to insufficient or absent observations in the latter categories, the moderate and severe dementia categories, which had very few records, were combined with the mild dementia category. Consequently, the distribution of the response variable is presented in Table 4.4.

Table 4.4 Distribution of Dementia Status

	No Dementia	**Very Mild Dementia**	**Mild Dementia**
Number of cases	1562	171	31
Percentages	88.5%	9.7%	1.8%

As mentioned previously, the response variable was determined by taking CDR scores into account. When subsequent visits of the subjects were examined, the severity level of the dementia status of some subjects was downgraded. Dementia is

generally known as a progressive neurodegenerative disease, which is not possible to be cured (Profyri et al., 2022). Due to this occurrence, either observations that lead to a downgrade should be replaced with the Last Observation Carried Forward (LOCF) replacement approach or listwise deletion should be performed. However, upon further examination of the dataset, it was found that subjects were also diagnosed with different types of disorders or symptoms such as “remote drug-induced cognitive dysfunction”, “active other neurological/medicated disorders”, “active hypothyroidism”, “remote cardiovascular diseases”, “remote global cerebellar hypoplasia, “diabetes” and so on. Moreover, all of these medical conditions in the dataset have a connection with AD (Salehipouret et al., 2023; Leszek et al., 2021; Tzeng et al., 2022) and the study of Tzeng et al. (2022) revealed that 0.9% of the subjects within the study was reverted to No Dementia category from higher severity level of Dementia. Consequently, rather than manipulating the data or creating a bias in the response variable, the changes among severity levels were retained as they were.

Figures 4.2 - 4.5 display boxplots and bar graphs of the features available in the final set. In figures; MMSE, Total Ventricular Volume (Normalized), LH Bankssts Thickness, RH Lingual Thickness, Simon Number, LOGIMEM, APOE-e4 Status and Multivitamin Use variables’ outputs are given. (For the determination of these variables, kindly refer to the “Implementation of The Methods” section.) Blue boxes represent the most severe category of dementia status, green boxes stand for Very Mild Dementia status, while red boxes have No Dementia status.

On average, non-demented subjects tend to have smaller ventricular volumes and higher values for Right Hemisphere Lingual Thickness, Left Hemisphere Bankssts Thickness, MMSE, Logical Memory and Simon Number compared to subjects with dementia. This implies that regions such as the lingual gyrus and superior temporal sulcus generally exhibit greater cortical thickness and less ventricular enlargement in non-demented individuals. Ventricular enlargement may be linked with brain atrophy (Nestor et al., 2008). Additionally, lower MMSE values correspond to more severe dementia (Tahami Monfared et al., 2022).

The majority of participants without the APOE-e4 allele do not have dementia, as indicated by the large red bar. Among APOE-e4 carriers, there is a noticeable decrease in the number of individuals without dementia compared to those who do not carry the APOE-e4 allele. When multivitamin usage is taken into account, the number of individuals with very mild dementia is considerably low in both groups, with slightly more represented among those who do not use multivitamins.

In general, the presence of the APOE-e4 allele appears to be associated with a higher likelihood of developing dementia, as indicated by the larger proportion of non-demented individuals among those who do not have the APOE-e4 allele. Moreover, multivitamin use seems to be linked with a lower likelihood of developing dementia, as the proportion of demented individuals is higher among individuals who do not take it.

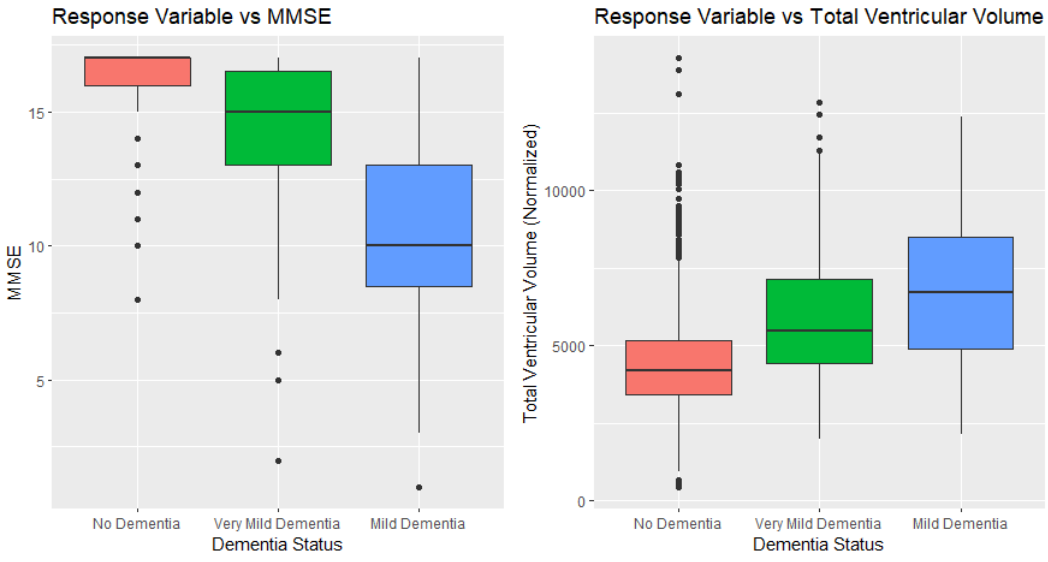


Figure 4.2 Boxplots of MMSE and TVV by response variable

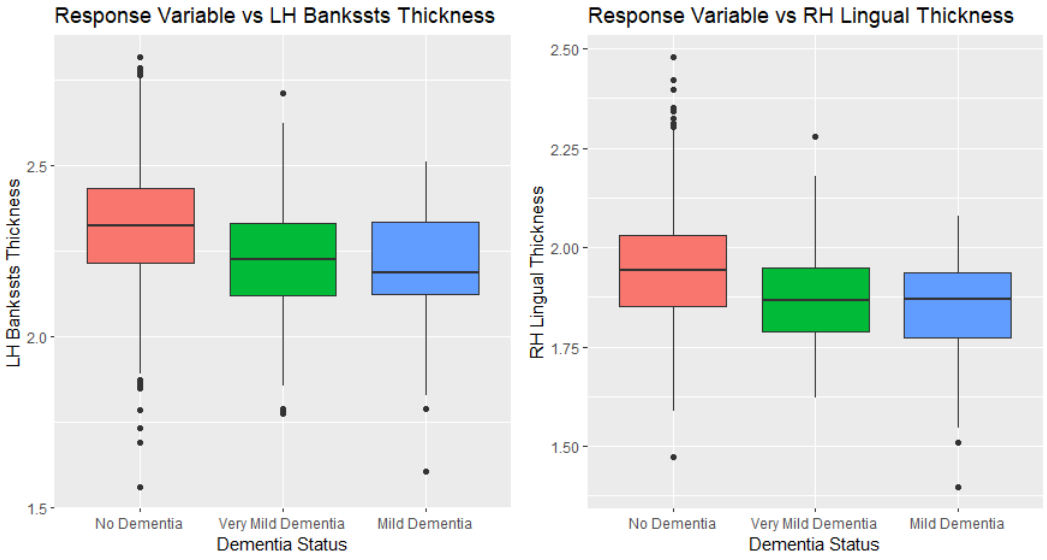


Figure 4.3 Boxplots of LH Bankssts Thickness and LH Lingual Thickness by response variable

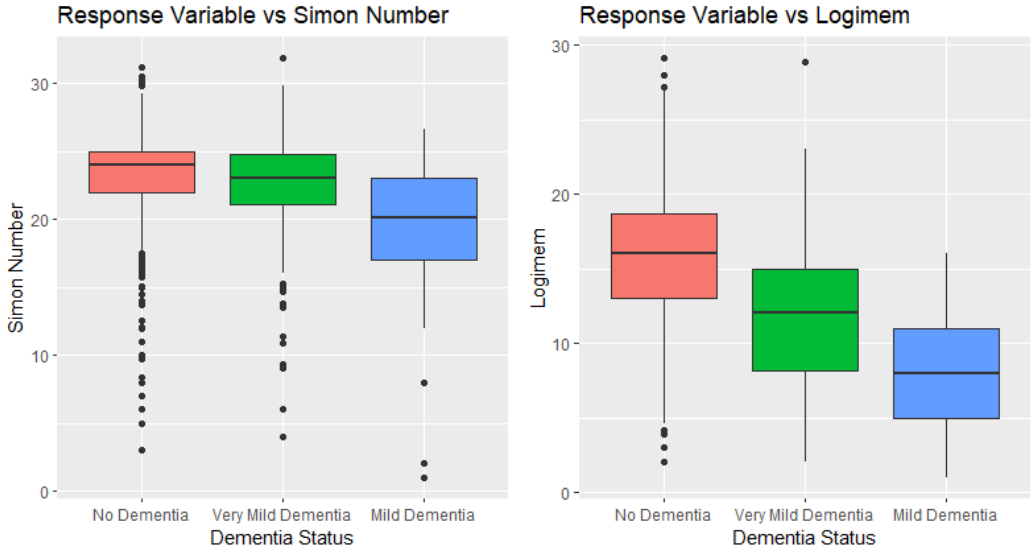


Figure 4.4 Boxplots of Simon Number and LOGIMEM by response variable

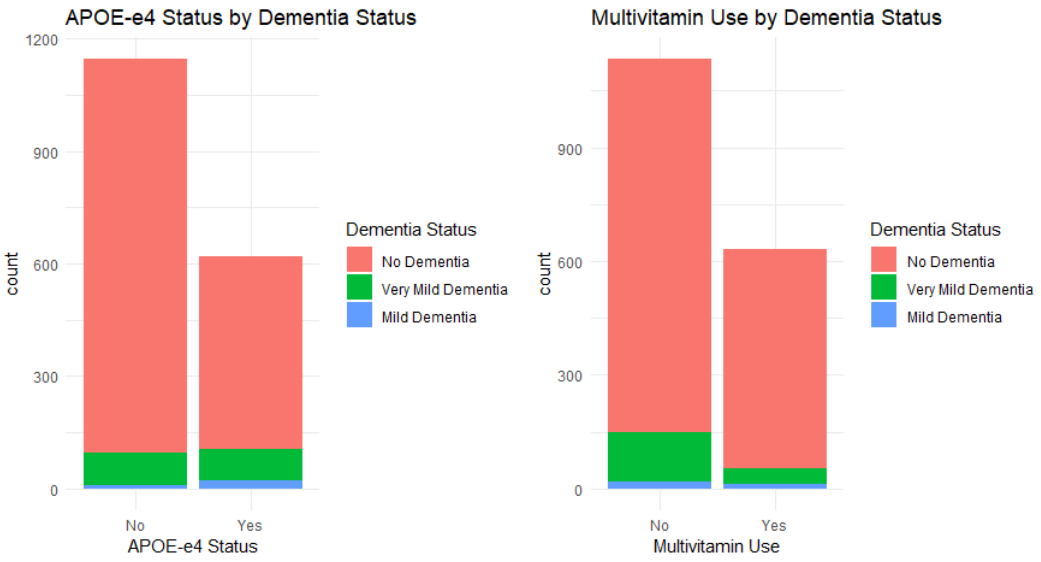


Figure 4.5 Bar graphs of APOE-e4 Status and Multivitamin Use by response variable

4.4 Implementation of Methods

One of the most important steps in ensuring the robustness and generalizability of both statistical models and machine learning algorithms is to generate a more reliable and robust estimate of the model's performance. To reflect this, data was separated into three parts as training set, test set, and validation set.

- Training Set (60% of the subjects): This subset was used for model training, allowing the algorithms to learn the underlying patterns and relationships within the data.
- Validation Set (20% of the subjects): This subset served as a basis for tuning the model's hyperparameters and for implementing techniques to prevent overfitting.
- Testing Set (20% of the subjects): This subset was reserved for the final evaluation of the model's performance, providing an unbiased assessment of how well the model is expected to generalize to unseen data.

To address the imbalanced class structure present in the dataset, stratified sampling was employed during the data splitting process. Stratified sampling ensured that the distribution of the response variable was preserved across all subsets, maintaining the representativeness of class proportions in the training, validation, and testing sets. This approach not only enhanced the model's ability to learn from the data but also ensured that the evaluation metrics reflected a fair assessment of the model's performance in imbalanced classes.

Table 4.5 Summary of Cross-validation

	Subject Size	**Sample Size**
Training set (60%)	354	1070
Test set (20%)	118	343
Validation set (%20)	118	351

After performing the cross-validation, seventy-three features were available in each dataset. Our objective was to explain the response variable, which is the severity level of AD, by constructing the largest possible regression model that converges across various regression approaches, including mixed effect models, machine learning model, and Generalized Estimating Equations (GEE) models.

In that respect, the buildclmm function, which is a part of the buildmer package in R, was constructed. The algorithm evaluates all the covariates step by step with respect to the response and performs stepwise elimination based on the change in different criteria such as likelihood ratio test, log-likelihood, AIC, BIC or deviance. In this thesis, AIC was used as a criterion in the buildclmm function, as it effectively handles model complexity while providing better goodness of fit overall (Akaike, 1974). The algorithm was run solely on the training dataset, and the selected features from the training dataset were subsequently filtered on the other datasets.

As a result, the features "Simon Number", "Total Ventricular Volume", "LOGIMEM", "RH Lingual Thickness", "MMSE" and "APOE-e4 Status" were

selected by the algorithm. Besides, “Multivitamin Use” and “LH Bankssts Thickness” covariates were included in the model due to their link with cognitive and memory changes in subjects (Sacchi et al., 2023; Yeunget al., 2023).

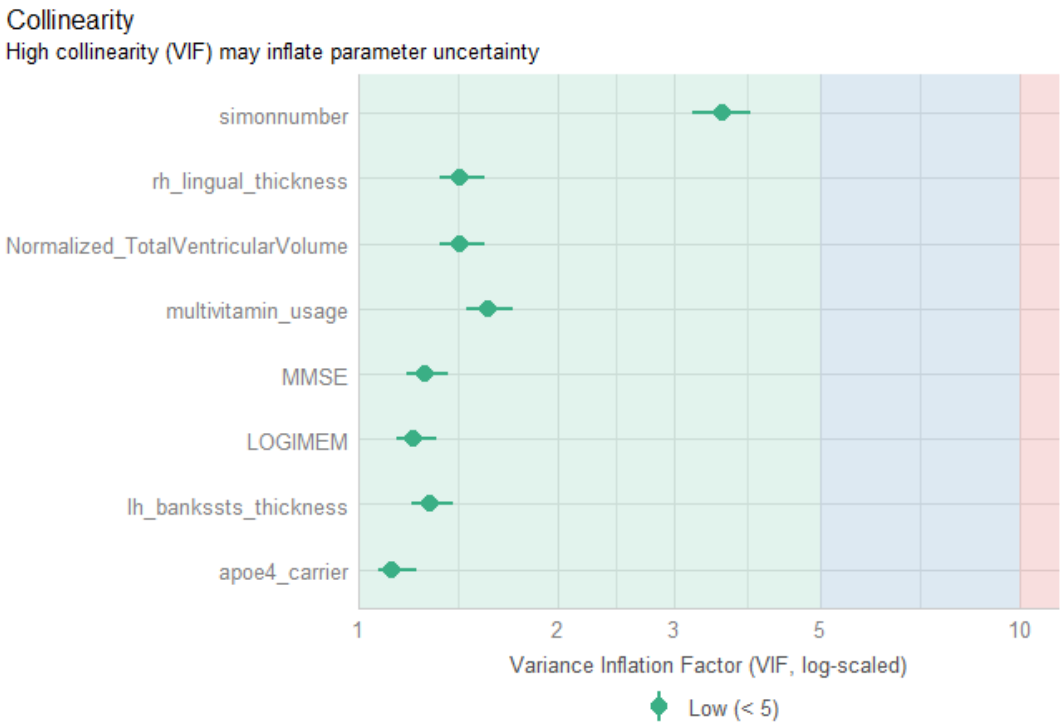


Figure 4.6 Multicollinearity plot

Figure 4.6 shows a multicollinearity of model terms, calculated by the variance inflation factor (VIF). It occurs mainly when two or more explanatory variables are highly correlated. Including these variables in the model leads to estimates of the models becoming unstable and making it difficult to interpret the relationship between the covariates and the response accurately. Features were evaluated in terms of multicollinearity issues by utilizing the performance package in R, and it was found that all of the features have considerably low VIF scores.

The description of the variables in the data is given in Table 4.6.

Table 4.6 Description of the final dataset

Variable Name	**Variable Type**	**Description**
OASIS ID	Factor	ID number of subjects
Simon Number	Numeric	Number of correct responses on all 120 trials (Simon, 1969)
LOGIMEM	Numeric	Logical memory score for immediate recall (Wechsler, 1987)
Multivitamin Use	Factor	Whether the subject is using multivitamins. (1=False, 2=True)
APOE-e4 Status	Factor	Whether the subject carries the APOE-e4 gene. (1=True, 2=False)
MMSE	Numeric	Mini-mental state examination scores
Right Hemisphere Lingual Thickness	Numeric	The average thickness of the lingual gyrus in the right hemisphere of the brain
Normalized Total Ventricular Volume	Numeric	Total volume of the brain's ventricles normalized for head size
Left Hemisphere Bankssts Thickness	Numeric	The average thickness of the banks of the superior temporal sulcus in the left hemisphere
Days From Entry	Factor	Number of days since the subject's initial entry into the study
CDR Groups	Ordinal	Clinical Dementia Rating (CDR) groups indicating the severity of dementia (0=No Dementia, 1= Very Mild Dementia, 2= Mild Dementia)

4.4.1 Marginal Models with ORDGEE

The function ordgee in geepack (Yan et al., 2012a) estimates the marginal effects of covariates on the severity status of AD while accounting for the correlated structure of longitudinal data and the predictor variables. The estimations were calculated via the method suggested by Heagerty and Zeger (1996) and Højsgaard et al. (2006). Likewise, the function allows several links as well as different association structures (independence, exchangeable, unstructured, and user-defined). It was mentioned that the package of geepack, in which the function of ordgee is built-in, produces unreliable estimation results for multinomial responses (Touloumis, 2015). However, just for the sake of the comparison, the function was utilized.

The analysis was performed considering the “exchangeable” working structure and “probit” link function. Probit link function was utilized here due to logit link was resulted in not producing any estimates. Interpretation of probit estimates is challenging as compared to the logit link function. Yet, approximate conversion is possible via utilizing a constant, which is equal to $\left(\frac{15}{16}\right)\left(\frac{\pi}{\sqrt{3}}\right)$ (Johnson et al., 1995, as cited in Asar, 2013).

In the application of ordgee from the geepack package in the R model, when we add variables "Multivitamin Use," "APOE-e4 Status," "RH Lingual Thickness," and "LH Bankssts Thickness", the model does not converge. Therefore, these covariates were extracted from the model and refined accordingly.

Table 4.7 Summary of the ordgee Random Effect model with "OASIS ID"

	Estimate	**Std Error**	**Z Value**	**P Value**
Inter:No Dementia	1.65	0.15	127.87	<0.001
Inter:Very Mild Dementia	3.24	0.20	261.13	<0.001
Simon Number	0.23	0.05	18.37	<0.001
LOGIMEM	0.33	0.08	16.85	<0.001
MMSE	0.46	0.09	27.59	<0.001
Normalized TVV	-0.23	0.06	12.64	<0.001

TVV: Total Ventricular Volume

The estimated model incorporating the probit-logit conversion constant can be displayed as

$$\ln\left(\frac{P(Y_{it}\leq c)}{1-P(Y_{it}\leq c)}\right) = \widehat{\beta_c} + 0.40 * (Simon\ Number) + 0.57 * (LOGIMEM) + 0.78 * (MMSE) - 0.39 * (Normalized\ TVV)$$

$$logit\big(P(Y \leq 0)\big) = 2.81 + 0.40 * (Simon\ Number) + 0.57 * (LOGIMEM) + 0.78 * (MMSE) - 0.39 * (Normalized\ TVV)$$

$$logit\big(P(Y \leq 1)\big) = 5.51 + 0.40 * (Simon\ Number) + 0.57 * (LOGIMEM) + 0.78 * (MMSE) - 0.39 * (Normalized\ TVV)$$

where $C = 0$ refers to No Dementia, $C = 1$ stands for Very Mild Dementia in ordgee function.

When no covariates are considered, the odds of No Dementia are $exp(2.81)$= 16.61 compared to Very Mild or Mild Dementia. Similarly, the odds of No Dementia or Very Mild Dementia are $exp(5.51) = 247.15$ compared to Mild Dementia. These results align with the observed distribution of the response.

The estimate for Simon Number is 0.40, indicating that higher Simon Number scores are associated with lower log odds of being in a higher dementia category. The

coefficient estimate of MMSE is 0.78, indicating that higher MMSE scores are associated with lower log odds of being in a higher dementia category. On the other hand, Normalized TVV is estimated as -0.39, meaning that larger ventricular volumes are associated with higher log odds of being in a higher dementia category. Increased ventricular volume is often linked with brain atrophy and higher dementia severity. Higher scores in cognitive assessments (Simon Number, LOGIMEM, MMSE) are associated with lower dementia severity, while larger normalized ventricular volumes are associated with higher dementia severity

4.4.2 Marginal Models with ORDLORGEE

The function ordLORgee in multgee (Touloumis, 2015) was implemented to acquire the population-averaged estimates of the parameter of interest and to manage the correlation structure among covariates and response. The function allows several links (logit, probit, cauchit, cloglog) as well as different association structures (uniform, time.exch, category.exch, independence, RC and fixed). In addition, intrinsic parameters were also introduced in the ordLORgee function to be helpful regarding the selection of the working structure. Based on the range of the intrinsic parameter estimates working structure can be useful in identifying between structures. For the ordinal response case, all working structures can be used depending on the situation where intrinsic parameters are dispersed. If the range of the intrinsic parameters is close to each other within a margin of two units, a uniform working structure should be utilized (Touloumis, 2013). Besides, the time-exch working structure was mentioned to address the convergence issue on the performance of the model. To display different working structure and their comparison several information criterion were used. Quasi Information Criterion (QIC), Correlation Inflation Criterion (CIC), Rotnitzky–Jewell Criterion (RJC) and Quasi-likelihood under the independence model criterion (QICu).

Table 4.8 Correlation Structures of Marginal Models

Correlation Structure	**QIC**	**CIC**	**RJC**	**QICu**
Time Exchangeable	164.65	167.21	7.72	1.41
Category Exchangeable	171.75	168.26	10.74	1.41

Table 4.9 Tukey's Five-Number Summary of Estimated Intrinsic Parameters

	Min	**L. Hinge**	**Median**	**U. Hinge**	**Max**
Estimated values	-14.38	<0.001	2.59	5.28	18.80

Before implementing the algorithm, the intrinsic parameters were examined and the difference was varied. In that respect, time.exch working structure was utilized by constructing the algorithm. Moreover, "LH Bankssts Thickness" was removed from the estimated model due to its insignificance, as the p-value exceeded the threshold value of 0.05.

Table 4.10 Summary of the ordLORgee Random Effect model with "OASIS ID"

	Estimate	**Std Error**	**Z Value**	**P Value**
β_1	2.11	0.22	9.69	<0.001
β_2	5.17	0.34	15.26	<0.001
Simon Number	0.32	0.09	3.57	<0.001
LOGIMEM	0.59	0.09	6.44	<0.001
Multivitamin Use (Y)	0.39	0.22	1.76	0.079
APOE-e4 Status (NC)	0.54	0.26	2.08	0.038
MMSE	0.50	0.09	5.68	<0.001
RH Lingual Thickness	0.38	0.13	3.02	0.003
Normalized TVV	-0.42	0.11	-4.00	<0.001

TVV: Total Ventricular Volume, Y: Yes, NC: Non-carrier

The estimated model can be written as follows:

$$\ln\left(\frac{P(Y_{it}\leq c)}{1-P(Y_{it}\leq c)}\right) = \widehat{\beta_c} + 0.32 * (Simon\ Number) + 0.59 * (LOGIMEM) + 0.39 * (Multivitamin\ Use) + 0.54 * (APOE\text{-}e4\ Status) + 0.50 * (MMSE) + 0.38 * (RH\ Lingual\ Thickness) - 0.42 * (Normalized\ TVV)$$

$$logit\big(P(Y \leq 1)\big) = 2.11 + 0.32 * (Simon\ Number) + 0.59 * (LOGIMEM) + 0.39 * (Multivitamin\ Use) + 0.54 * (APOE\text{-}e4\ Status) + 0.50 * (MMSE) + 0.38 * (RH\ Lingual\ Thickness) - 0.42 * (Normalized\ TVV)$$

$$logit\big(P(Y \leq 2)\big) = 5.17 + 0.32 * (Simon\ Number) + 0.59 * (LOGIMEM) + 0.39 * (Multivitamin\ Use) + 0.54 * (APOE\text{-}e4\ Status) + 0.50 * (MMSE) + 0.38 * (RH\ Lingual\ Thickness) - 0.42 * (Normalized\ TVV)$$

where $C = 1$ refers to No Dementia, $C = 2$ stands for Very Mild Dementia in ordLORgee function.

When marginal models with ordgee function estimates are compared with estimates taken from ordLORgee function, it is seen that estimate coefficients are close to each other, indicating consistency among the marginal model's applications.

The ordLORgee model identifies several significant predictors of dementia severity. Higher scores in cognitive assessments (Simon Number, LOGIMEM, MMSE) and greater RH Lingual Thickness are associated with less severe dementia, while larger ventricular volumes are associated with more severe dementia. Besides, the APOE-e4 allele is a significant genetic factor linked to increased dementia severity and multivitamin use shows a positive association with lower dementia severity.

4.4.3 Generalized Linear Mixed Effect Models with GLMMLASSO

The Generalized Linear Mixed Model with L1-penalized estimation (glmmLasso) was employed to assess the relationship between various predictors and the severity levels of dementia, measured through the ordinal response variable. The analysis

indicated that all the variables, except for Left Hemisphere Bankssts Thickness were statistically significant since their p-values were below the significance level of 0.05. The p-value for the non-significant variable is measured as 0.07, which is marginally above the threshold of 0.05. Despite the slight margin, it can be inferred that LH Bankssts Thickness still shows a meaningful association with the response variable, dementia severity levels, when accounting for the random intercept. Therefore, the random effect model was retained in its original form, and the model parametrization was constructed by considering all covariates.

Table 4.11 Summary of the glmmLasso Random Effect model with "OASIS ID"

	Estimate	**Std Error**	**Z Value**	**P Value**
θ_1	2.21	0.16	13.74	<0.001
θ_2	5.34	0.24	22.1	<0.001
Simon Number	0.32	0.08	4.04	<0.001
LOGIMEM	0.62	0.11	5.94	<0.001
Multivitamin Use (Y)	0.57	0.23	2.54	0.011
APOE-e4 Status (NC)	0.46	0.21	2.20	0.028
MMSE	0.65	0.09	6.87	<0.001
RH Lingual Thickness	0.35	0.12	2.93	0.003
Normalized TVV	-0.41	0.09	-4.44	<0.001
LH Bankssts Thickness	0.20	0.11	1.81	0.070

TVV: Total Ventricular Volume, Y: Yes, NC: Non-carrier

The estimated model can be written as follows:

$$\ln\left(\frac{P(Y_{it}\leq c)}{1-P(Y_{it}\leq c)}\right) = \widehat{\theta_c} + 0.32 * (Simon\ Number) + 0.62 * (LOGIMEM) + 0.57 * (Multivitamin\ Use) + 0.46 * (APOE\text{-}e4\ Status) + 0.65 * (MMSE) + 0.35 * (RH\ Lingual\ Thickness) - 0.41 * (Normalized\ TVV) + 0.20 * (LH\ Bankssts\ Thickness) + \widehat{b_\iota}$$

$\widehat{b_\iota}$ is the random intercept term and the term is distributed as $\widehat{b_\iota} \sim N(0,0.12)$.

The coefficient estimates from the glmmLasso method are similar to those obtained from the marginal models' methods, suggesting that their interpretations align.

$$logit\big(P(Y \leq 1)\big) = 2.11 + 0.32 * (Simon\ Number) + 0.62 * (LOGIMEM) + 0.57 * (Multivitamin\ Use) + 0.46 * (APOE\text{-}e4\ Status) + 0.65 * (MMSE) + 0.35 * (RH\ Lingual\ Thickness) - 0.41 * (Normalized\ TVV) + 0.20 * (LH\ Bankssts\ Thickness) + \widehat{b_i}$$

$$logit\big(P(Y \leq 2)\big) = 5.34 + 0.32 * (Simon\ Number) + 0.62 * (LOGIMEM) + 0.57 * (Multivitamin\ Use) + 0.46 * (APOE\text{-}e4\ Status) + 0.65 * (MMSE) + 0.35 * (RH\ Lingual\ Thickness) - 0.41 * (Normalized\ TVV) + 0.20 * (LH\ Bankssts\ Thickness) + \widehat{b_i}$$

where $C = 1$ refers to No Dementia, $C = 2$ stands for Very Mild Dementia in glmmLasso function.

Higher cognitive scores and certain brain thickness measurements are associated with less severe dementia, while larger ventricular volumes and the presence of the APOE-e4 allele are linked to more severe dementia.

4.4.4 Generalized Linear Mixed Effect Models with CLMM

The OASIS-3 longitudinal data analysis was performed with CLMM using maximum likelihood estimation with adaptive Gauss-Hermite quadrature approximation. The statistical analysis revealed that only LH Bankssts Thickness was not statistically significant since their p-values were above the significance level of 0.05. This result suggests that this variable does not have a significant association with the severity levels of dementia. Thus, the cumulative link mixed effect model is reconducted by removing insignificant variable from the model.

The refitted model's summary and outputs can be displayed below.

Table 4.12 Summary of the clmm Random Effect model with "OASIS ID"

	Estimate	**Std Error**	**Z Value**	**P Value**
Simon Number	-0.52	0.16	-3.29	<0.001
LOGIMEM	-1.04	0.19	-5.58	<0.001
Multivitamin Use (Y)	-0.71	0.35	-2.06	0.040
APOE-e4 Status (NC)	-0.91	0.44	-2.05	0.040
MMSE	-0.97	0.16	-5.95	<0.001
RH Lingual Thickness	-0.71	0.22	-3.29	<0.001
Normalized TVV	0.73	0.19	3.84	<0.001

TVV: Total Ventricular Volume, Y: Yes, NC: Non-carrier

Threshold values or cutpoints regarding the clmm model can be found below.

Table 4.13 Summary of the threshold or cutpoints

	Estimate	**Std Error**	**Z Value**
No Dementia|Very Mild Dementia	4.05	0.50	8.08
Very Mild Dementia|Mild Dementia	9.38	0.90	10.44

The estimated model can be written as follows:

$$\ln\left(\frac{P(Y_{it}\leq c)}{1-P(Y_{it}\leq c)}\right) = \widehat{\theta_c} + 0.52 * (Simon\ Number) + 1.04 * (LOGIMEM) + 0.71 * (Multivitamin\ Use) + 0.91 * (APOE\text{-}e4\ Status) + 0.97 * (MMSE) + 0.71 * (RH\ Lingual\ Thickness) - 0.73 * (Normalized\ TVV) + \widehat{b_i}$$

$\widehat{b_i}$ is the random intercept term and the term is distributed as $\widehat{b_i} \sim N(0, 7.513)$.

$$logit\big(P(Y \leq 1)\big) = 4.05 + 0.52 * (Simon\ Number) + 1.04 * (LOGIMEM) + 0.71 * (Multivitamin\ Use) + 0.91 * (APOE - e4\ Status) + 0.97 * (MMSE) + 0.71 * (RH\ Lingual\ Thickness) - 0.73 * (Normalized\ TVV) + \widehat{b_i}$$

$$logit\big(P(Y \leq 2)\big) = 9.38 + 0.52 * (Simon\ Number) + 1.04 * (LOGIMEM) + 0.71 * (Multivitamin\ Use) + 0.91 * (APOE - e4\ Status) + 0.97 * (MMSE) + 0.71 * (RH\ Lingual\ Thickness) - 0.73 * (Normalized\ TVV) + \widehat{b_i}$$

where $C = 1$ refers to No Dementia, $C = 2$ stands for Very Mild Dementia in clmm function.

Although the absolute values of the coefficients from the clmm outputs are similar to those from other approaches such as ordgee, ordLORgee, and glmmLasso, the signs are opposite. This discrepancy arises because clmm uses different parametrization compared to the other models. When incorporating a negative sign in the model parametrization, the estimated model aligns with the others and the interpretation will be the same.

The output shows that for individuals who do not have the APOE-e4 gene, the log odds of being non-demented (versus being in a Very Mild and Mild Dementia category) are actually $\eta_1 = (-0.91) = 0.91$ points higher than individuals who have the APOE-e4 gene. Likewise, the odds ratio interpretation will be further expanded as individuals who do not have the APOE-e4 gene have $exp(0.91) = 2.48$ times the odds of being non-demented (versus being in a Very Mild or Mild Dementia category) compared to individuals who have the APOE-e4 gene. Therefore, it can be concluded that the presence of the APOE-e4 allele gene is correlated with higher dementia severity. Similarly, right hemisphere lingual thickness is linked to severe dementia severity as the thickness decreases. Additionally, multivitamin use is associated with lower dementia severity. These results are consistent across all models (ordgee, ordLORgee, and glmmLasso), suggesting robust relationships between these predictors and dementia severity, despite differences in model parametrization.

4.4.5 Bayesian Generalized Linear Mixed Effect Models with MCMCGLMM

MCMCglmm package was designed for generalized linear mixed effect models via utilizing MCMC techniques that are leveraging on several sampling approaches such as Gibbs sampling, slice sampling and Metropolis-Hasting algorithms (Hadfield J., 2010). MCMCglmm algorithm supports probit link for ordinal response cases, hence estimated values were based on probit link. Conversion of probit to logit was computed for all posterior means by using the same constant used in ordgee (Johnson et al., 1995, as cited in Asar, 2013) and provided in the estimated models. MCMCglmm uses inverse-Wishard prior for covariance and implementing normal priors for the fixed effect part.

The analysis resulted in all variables being significant, meaning that the p-values of MCMC were below the significance level of 0.05. Similar to glmmLasso algorithms, this method suggests that all have a significant association with the response variable.

Table 4.14 Hyperparameters of MCMCglmm algorithm

	Value
Iterations	3001:99901
Thinning interval	100
Sample size	970

Table 4.15 Summary of the MCMCglmm cutpoints and structures

	Posterior Mean	**l-95% CI**	**u-95% CI**	**Effective Sample**
Cutpoints	4.42	3.78	5.43	168
G-structure (~OASIS ID)	5.63	2.96	8.45	93
R-structure (~units)	1	1	1	0

The cutpoints, which divide the ordinal response variable as a function of the linear predictor, represent the bin boundaries for the latent variables associated with each category of the response variable. The G-structure is the variance-covariance matrix of the random effects, while the R-structure is the variance-covariance matrix of the residuals. These structures are essential for calculating the intraclass correlation and the Highest Posterior Density (HPD) intervals.

In Table 4.15, the cutpoints were determined alongside the variance-covariance matrices of the random effects (G-structure) and residuals (R-structure), with a posterior mean of 4.42, 5.63, and 1, respectively. The MCMCglmm algorithm always fits the R-structure term due to the mixing properties of the chain. However, observation-level heterogeneity cannot be estimated from the data in ordinal response models, so fixing the residual at a specific value is often recommended. In that regard, a redundant non-identified parameter was introduced in the Inverse-Wishart prior to setting the R-structure at 1, enhancing the mixing properties of the chain.

Table 4.16 Summary of the MCMCglmm algorithm

	Posterior Mean	**l-95% CI**	**u-95% CI**	**Effective Sample**
Intercept	-3.32	-4.13	-2.59	173
Simon Number	-0.43	-0.69	-0.17	323
LOGIMEM	-0.84	-1.12	-0.55	369
Multivitamin Use (Y)	-0.54	-1.09	-0.03	750
APOE-e4 Status (NC)	-0.81	-1.57	-0.08	472
MMSE	-0.77	-1.03	-0.50	575
RH Lingual Thickness	-0.54	-0.92	-0.18	519
Normalized TVV	0.58	0.24	0.89	472
LH Bankssts Thickness	-0.27	-0.63	-0.02	437

pMCMC <0.05

TVV: Total Ventricular Volume, Y: Yes, NC: Non-carrier

Even though pMCMC was reported, it is beneficial to evaluate the credible intervals for each covariate. Since both lower and upper credible intervals do not include zero, it is safe to state that the fixed effects are significant.

The main distinction of MCMCglmm from other models is that instead of two intercepts or coefficient estimates to be used in constructing cumulative probabilities, it provides only one intercept and one cutpoint value. The estimated model can be illustrated as follows considering cutpoint and intercept values:

$$\ln\left(\frac{P(Y_{it}\leq c)}{1-P(Y_{it}\leq c)}\right) = \widehat{\gamma_c} - \hat{l} + 0.73 * (Simon\ Number) + 1.42 * (LOGIMEM) + 0.92 * (Multivitamin\ Use) + 1.38 * (APOE\text{-}e4\ Status) + 1.32 * (MMSE) + 0.92 * (RH\ Lingual\ Thickness) - 0.99 * (Normalized\ TVV) + 0.47 * (LH\ Bankssts\ Thickness) + \widehat{b_i}$$

where γ stands for cutpoint (γ_1 set to 0 in ordinal models), l refers to intercept and $\widehat{b_i}$ is the (posterior) random intercept term with a variance of 2.06.

$$logit\big(P(Y \leq 1)\big) = (5.64) + 0.73 * (Simon\ Number) + 1.42 * (LOGIMEM) + 0.92 * (Multivitamin\ Use) + 1.38 * (APOE\text{-}e4\ Status) + 1.32 * (MMSE) + 0.92 * (RH\ Lingual\ Thickness) - 0.99 * (Normalized\ TVV) + 0.47 * (LH\ Bankssts\ Thickness) + \widehat{b_i}$$

$$logit\big(P(Y \leq 2)\big) = 13.15 + 0.73 * (Simon\ Number) + 1.42 * (LOGIMEM) + 0.92 * (Multivitamin\ Use) + 1.38 * (APOE\text{-}e4\ Status) + 1.32 * (MMSE) + 0.92 * (RH\ Lingual\ Thickness) - 0.99 * (Normalized\ TVV) + 0.47 * (LH\ Bankssts\ Thickness) + \widehat{b_i}$$

where $C = 1$ refers to No Dementia, $C = 2$ stands for Very Mild Dementia in MCMCglmm function. For consistency across all models, probit-logit conversion was applied to the estimated coefficients.

In general, the results from the MCMCglmm algorithm indicate that better cognitive scores (Simon Number, LOGIMEM, MMSE) and greater brain thickness (RH Lingual and LH Bankssts) are associated with lower dementia severity. In contrast, larger ventricular volumes are linked to higher dementia severity. The association of multivitamin use with dementia severity is less definitive but suggests potential protective effects. Furthermore, the estimated variance of the random intercept for the OASIS ID variable was calculated as 2.06. This result suggests that variability among subjects exists, which indicates that different subjects have different probabilities of having different dementia statuses.

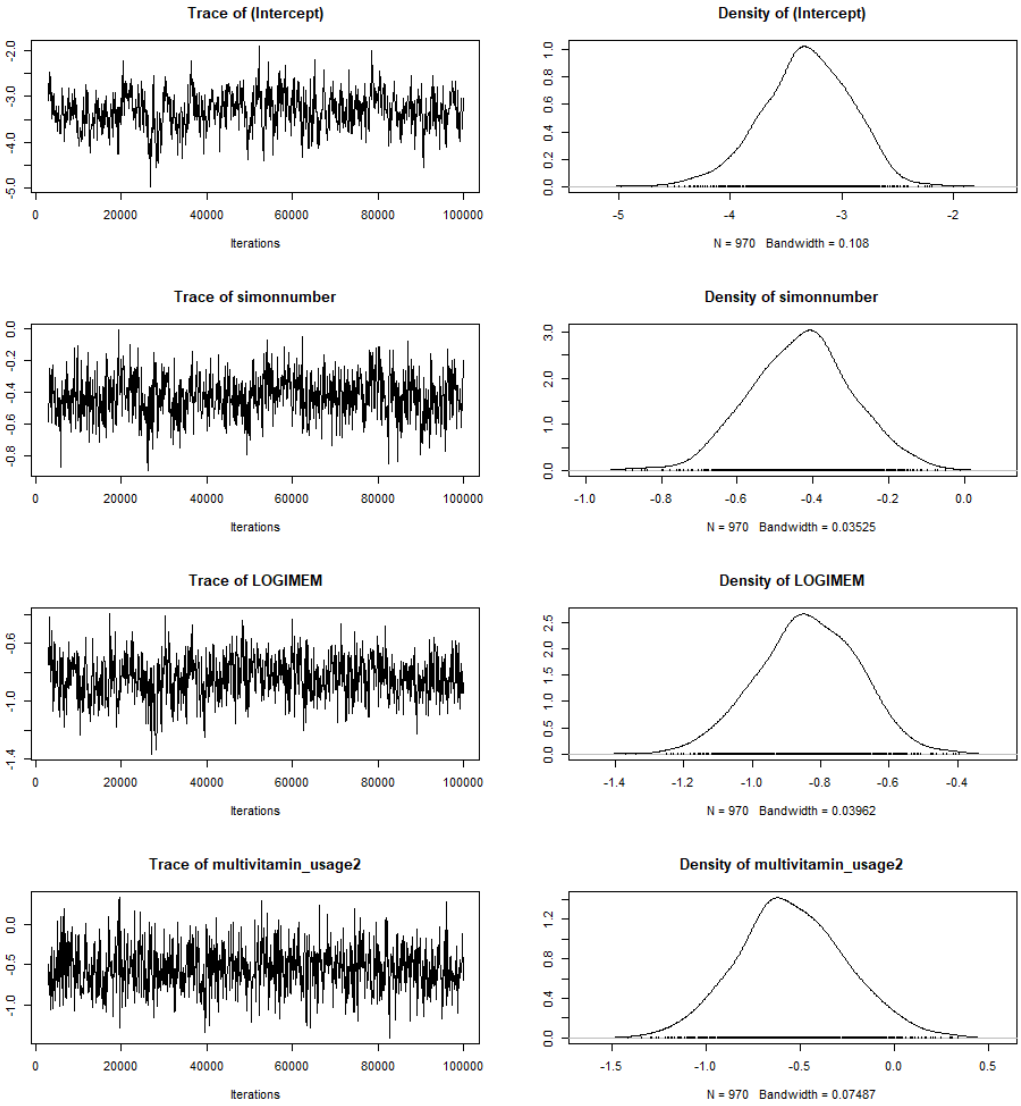


Figure 4.7 Trace and density plots of the estimate of the intercept and covariates

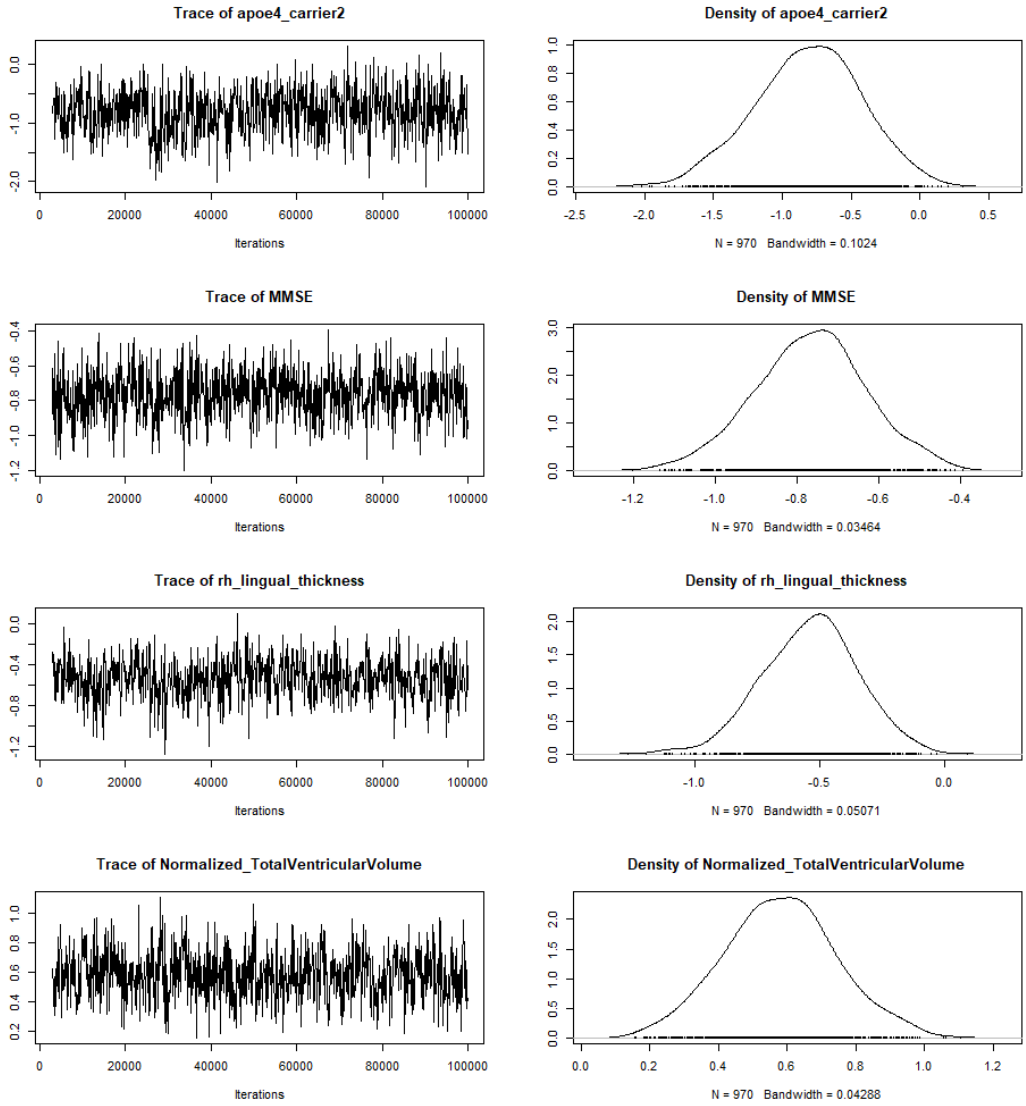


Figure 4.8 Trace and density plots of the estimate of the covariates

In Figures 4.7 and 4.8, the trace and density estimate plots of the intercept and covariates can be seen. When trace plots are evaluated, the model looks mixed well and chain convergence seems to be satisfied.

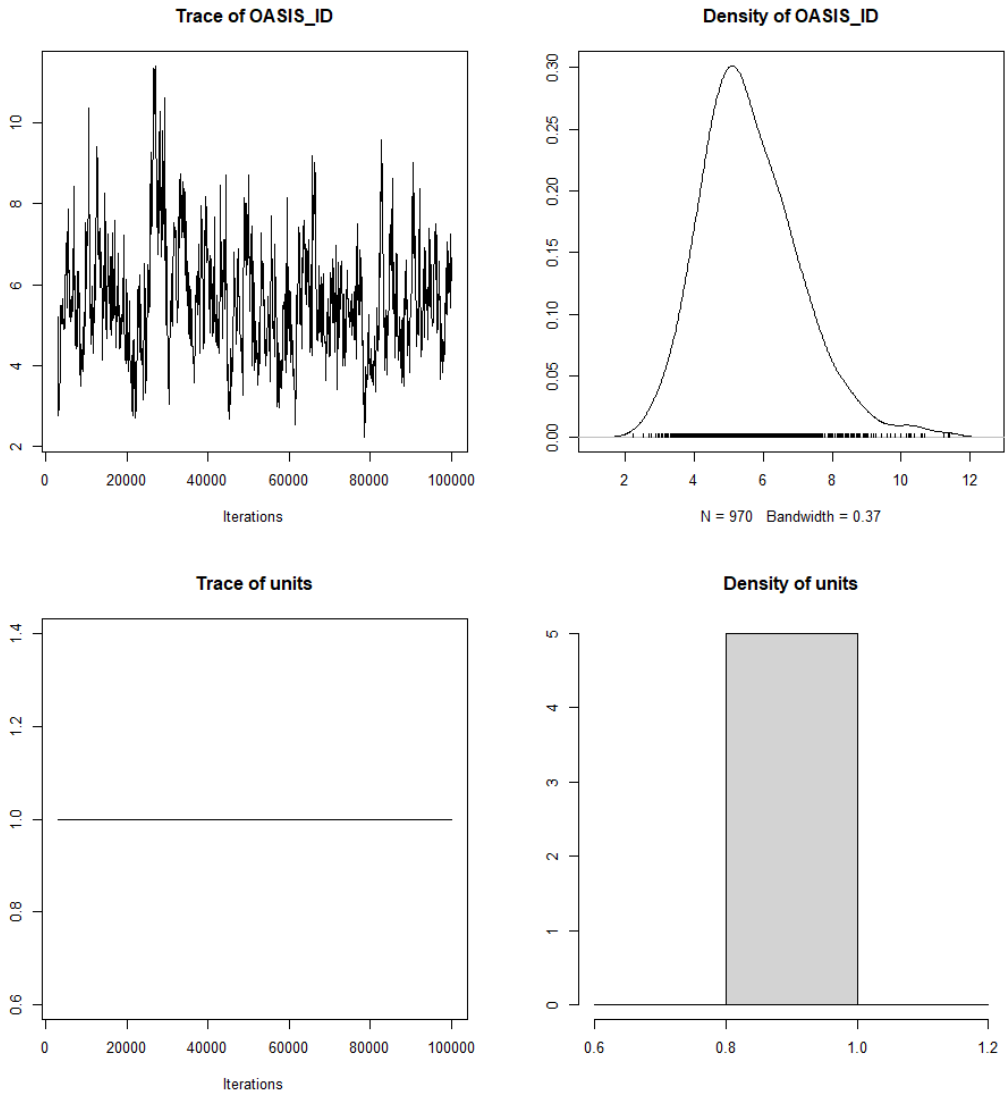


Figure 4.9 Trace and density plots of the estimate of the random effect and units

Upon evaluating the variances of both random effects and residuals, it appears that the residual variances are static and equal to 1 due to the prior used in the algorithm. For random effect variances, increasing the burn-in period and the number of iterations may improve the mixing.

4.4.6 Bayesian Mixed Effects Ordinal Regression Models with BRMS

In this thesis, ordinal mixed effect models in the context of the Bayesian perspective have been evaluated to compare and contrast the other applications specifically designed for the frequentist side. For this purpose, the brms package in R software was employed. The algorithm was designed in Stan to perform Bayesian generalized linear models via Stan. As previously mentioned in the Methodology section, there are similarities with the MCMCglmm function in terms of theoretical aspects, but brms has utilized Hamiltonian Monte Carlo (HMC) and its extension version of No-U-Turn Sampler (NUTS) for posterior distribution sampling.

The brms package does not provide a p-value in its summarization, instead, a significant variable is determined via lower and upper credible intervals decision. The analysis suggests that solely the “LH Bankssts Thickness” is negative with a zero overlapping 95%-CI. Moreover, it can be concluded that, on average, the bankssts thickness on the left hemisphere of the brain may reduce the severity of dementia by some amount, but is still not significant by standard decision rules.

Therefore, insignificant covariates were removed from the model and the model was reconstructed.

Table 4.17 Summary of the brms algorithm

	Estimate	Est. Error	l-95% CI	u-95% CI	Rhat	Bulk Ess.	Tail Ess.
Intercept[1]	4.02	0.51	3.11	5.12	1.00	2380	2342
Intercept[2]	9.51	0.94	7.85	11.53	1.00	2158	2337
Simon Number	-0.53	0.16	-0.85	-0.22	1.00	3778	3162
LOGIMEM	-1.07	0.19	-1.45	-0.71	1.00	4168	3277
Multivitamin Use (Y)	-0.75	0.36	-1.46	-0.04	1.00	5732	3478
APOE-e4 Status (NC)	-0.96	0.48	-1.90	-0.01	1.00	3519	3401
MMSE	-0.97	0.16	-1.30	-0.66	1.00	4994	3294
RH Lingual Thickness	-0.76	0.23	-1.21	-0.31	1.00	3346	2768
Normalized TVV	0.75	0.20	0.39	1.16	1.00	3594	3236

TVV: Total Ventricular Volume, Y: Yes, NC: Non-carrier

The cumulative density family of brms uses similar parametrization used in clmm. Therefore, the estimated model can be written as follows:

$$\ln\left(\frac{P(Y_{it}\leq c)}{1-P(Y_{it}\leq c)}\right) = \widehat{\theta_c} + 0.53 * (Simon\ Number) + 1.07 * (LOGIMEM) + 0.75 * (Multivitamin\ Use) + 0.96 * (APOE\text{-}e4\ Status) + 0.97 * (MMSE) + 0.76 * (RH\ Lingual\ Thickness) - 0.75 * (Normalized\ TVV) + \widehat{b_i}$$

$\widehat{b_i}$ is the random intercept term and the term is distributed as $\widehat{b_i} \sim N(0, 8.066)$.

$$logit\big(P(Y \leq 1)\big) = 4.02 + 0.53 * (Simon\ Number) + 1.07 * (LOGIMEM) + 0.75 * (Multivitamin\ Use) + 0.96 * (APOE - e4\ Status) + 0.97 * (MMSE) + 0.76 * (RH\ Lingual\ Thickness) - 0.75 * (Normalized\ TVV) + \widehat{b_i}$$

$$logit\big(P(Y \leq 2)\big) = 9.51 + 0.53 * (Simon\ Number) + 1.07 * (LOGIMEM) + 0.75 * (Multivitamin\ Use) + 0.96 * (APOE - e4\ Status) + 0.97 * (MMSE) + 0.76 * (RH\ Lingual\ Thickness) - 0.75 * (Normalized\ TVV) + \widehat{b_i}$$

where $C = 1$ refers to No Dementia, $C = 2$ stands for Very Mild Dementia in the brms function.

According to the results, all models (ordgee, ordLORgee, glmmLasso, clmm, MCMCglmm and brms), suggest the same relationships between these predictors and dementia severity, despite some differences in the model parametrization in clmm and MCMCglmm.

Additionally, the 'ESS_bulk,' 'ESS_tail,' and 'Rhat' outputs provide insight into the quality of the posterior distribution estimates for the parameter associated with dementia severity. Convergence of the coefficients is assessed using the "Rhat" statistic. If "Rhat" values exceed 1.00, this suggests that additional iterations or a more robust prior might be necessary. However, in our analysis, all parameters appear to have converged without any issues. The effective sample size (ESS) for both Bulk and Tail represents the number of draws that contain equivalent information to a dependent sample obtained via the MCMC algorithm. Ideally, ESS values should be maximized, with values greater than 1000 being considered sufficient for stable estimates (Bürkner, 2017). Typically, for each Markov Chain, which defaults to four chains, the bulk-ESS and tail-ESS should each be at least around 100 to ensure the reliability of the respective quantiles.

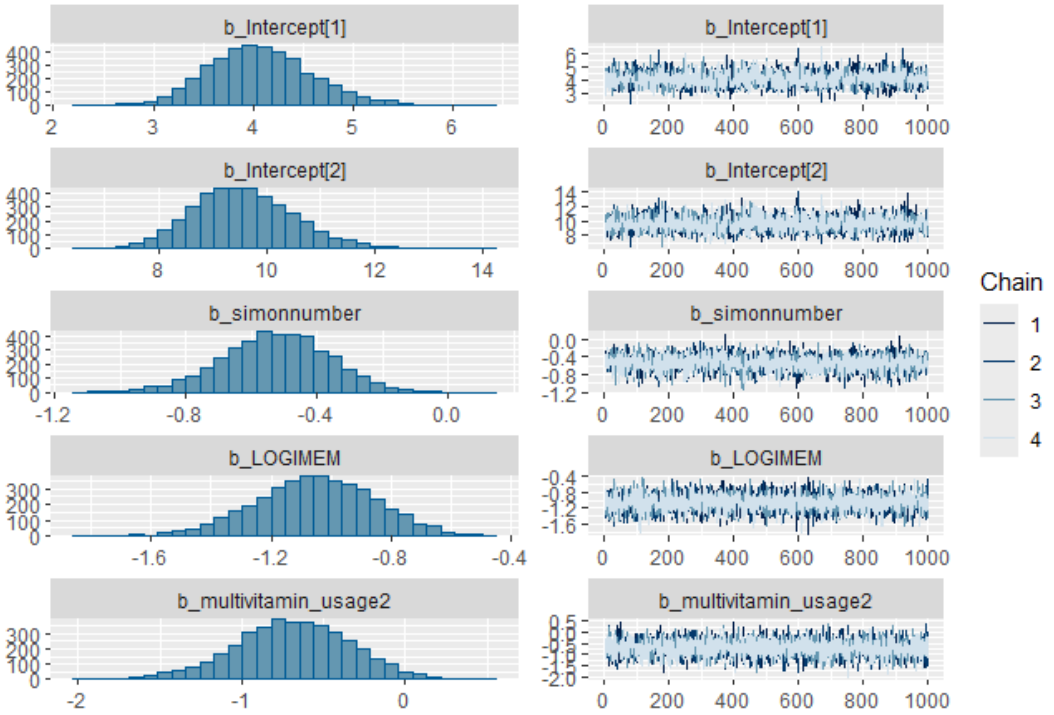


Figure 4.10 Convergence plots of covariates

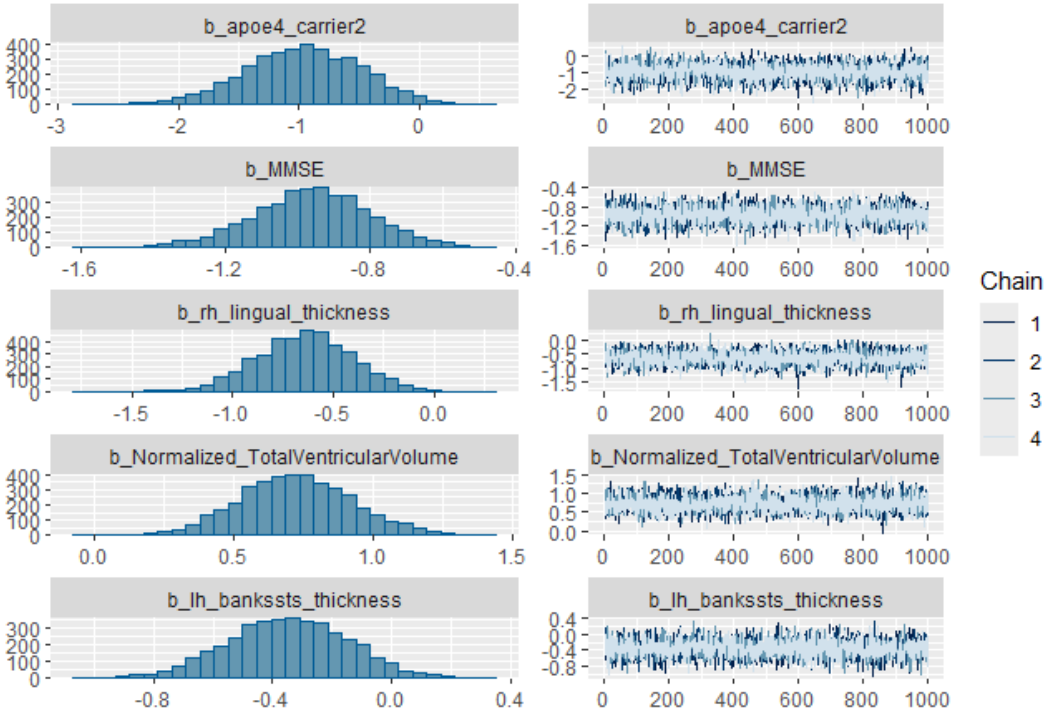


Figure 4.11 Convergence plots of covariates

When the convergence was evaluated, as the number of chains increased better convergence was satisfied for all variables.

4.4.7 Ordinal Mixed Effect Random Forest with OMERF

Several machine learning algorithms are available in the literature for mixed effect models. However, most of them were not designed for the mixed effect models with ordinal response, namely cumulative link models rarely exist. Bergonzoli (2023) and in the revised version Bergonzoli et al. (2024) proposed a statistical method, which is titled Ordinal Mixed-Effects Random Forest (OMERF), for the analysis of hierarchical data with ordinal response variable. This method merged two approaches, the standard cumulative link mixed models (CLMMs) and random forest (RF) algorithms, to explain the multi-level structure of the data by utilizing random and fixed effect parts separately through these algorithms. The algorithm also incorporates offset setup in the model specification, which comes from the ordinalForest (Hornung, 2020) function's estimation approach.

Firstly, the initialization step was built by utilizing the ordinalForest algorithm to capture potentially nonlinear trends and interactions. Estimates, computed from the ordinalForest, were used as an offset variable in the CLMM algorithm to be able to estimate the random effect in which the model is constructed as a random intercept model. Next, randomForest algorithm (Breiman, 2001) was applied to estimate the fixed effect. These processes were iteratively computed till the convergence was made.

Table 4.18 Summary of the CLMMs using the OMERF algorithm with offset variable

	Estimate	**Std Error**	**Z Value**	**P Value**
Offset(eta.est.rf)	0.61	0.11	5.34	<0.001

Threshold values or cutpoints regarding the clmm model using OMERF can be found below.

Table 4.19 Summary of the threshold or cutpoints via OMERF

	Estimate	**Std Error**	**Z Value**
No Dementia|Very Mild Dementia	9.00	0.93	9.70
Very Mild Dementia|Mild Dementia	14.37	1.38	10.40

As part of implementing a random forest, it's necessary to calculate the importance scores of the available variables. These scores reveal each variable's contribution to the model's performance and identify which variables are more effective in classifying the ordinal response variable. The Gini impurity and mean decrease in accuracy metrics were used for this purpose. These metrics indicate which variables are more relevant and informative, as well as identify the covariates that have little to no influence on the model.

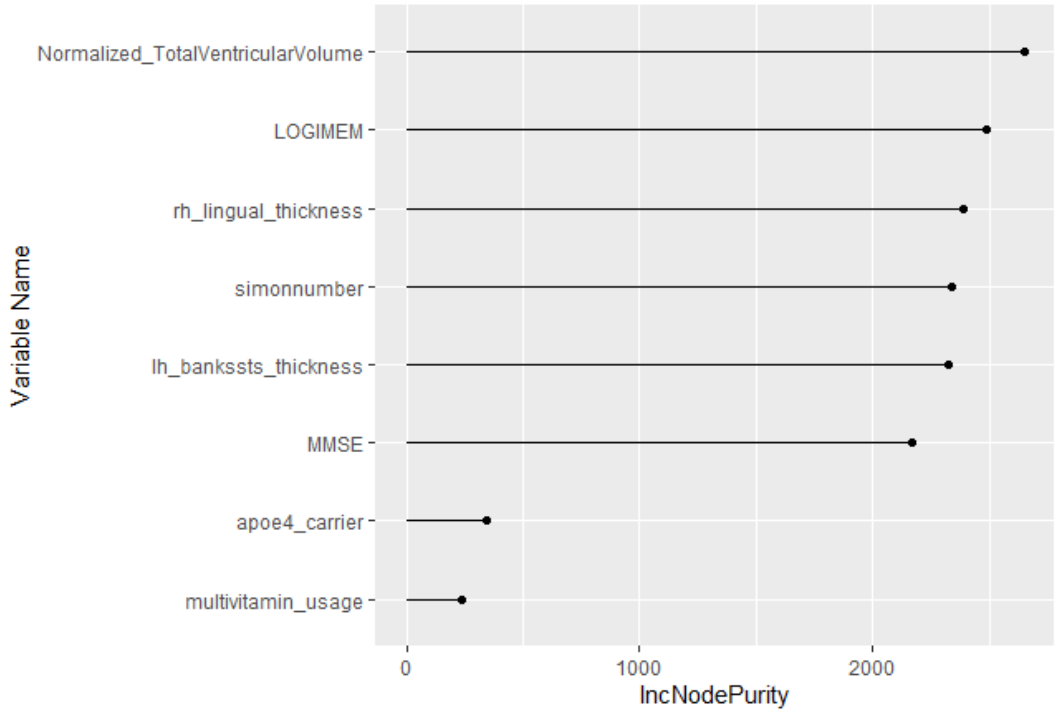


Figure 4.12 Variable importance plot of increase in node purity

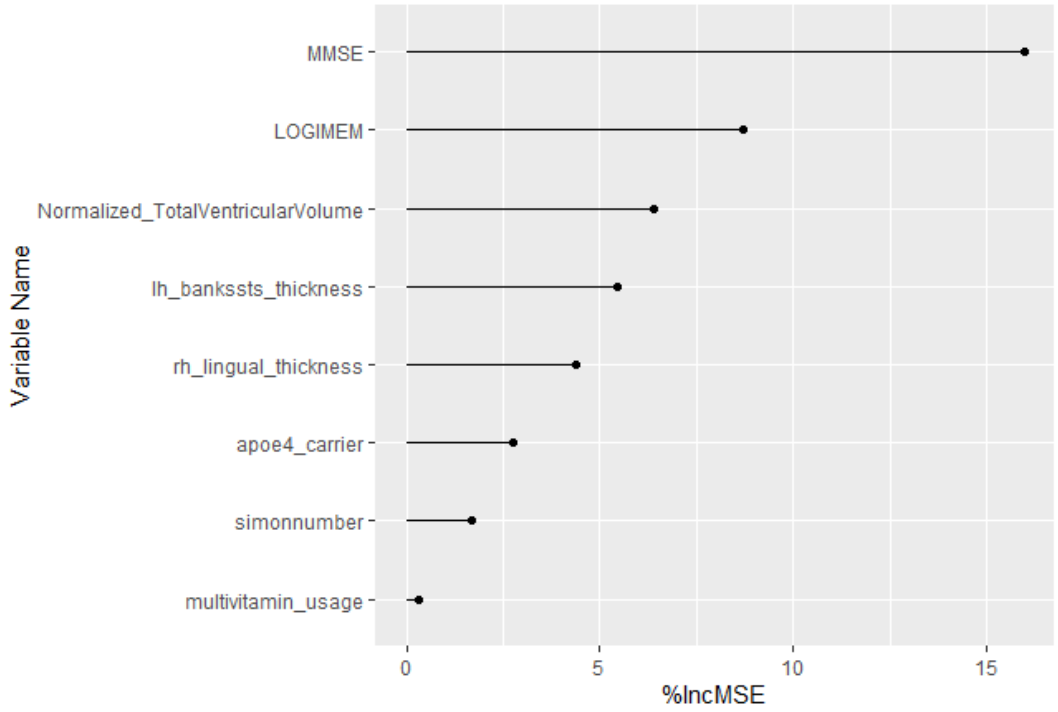


Figure 4.13 Variable importance plot of increase of Mean Square Error (MSE)

“%IncMSE” is robust and informative metric to identify the variable importance. Based on this metric, the “MMSE” and “LOGIMEM” variables were the most influential variables in classifying the dementia severity levels. On the other hand, “IncNodePurity” is more likely to relate to loss function and has importance in classification. Partial Dependence Plots (PDP) were also drawn from the part of the RF algorithm, which provides crucial information about the underlying trends of the ordinal response model.

The PDPs illustrate how the predicted outcome changes as a single predictor variable changes, while all other predictors are held constant. Examining each plot separately reveals all of the continuous variables were non-linear. For instance, the PDP of MMSE shows a steep decline in predicted dementia severity with increasing MMSE scores up to around -2. After this point, the severity levels off, indicating that higher MMSE scores are strongly associated with lower dementia severity. This finding aligns with the constructed statistical models. A similar pattern is observed for LOGIMEM as well. There is a sharp decrease in predicted dementia severity with increasing LOGIMEM scores up to around -2, after which the curve levels off. This suggests that better performance on the LOGIMEM, indicating better memory, is associated with lower dementia severity. While, for Normalized Total Ventricular Volume, the relationship appears more complex as compared to other PDPs. Initially, there is a slight increase in dementia severity with increasing ventricular volume, followed by fluctuations. Overall, larger ventricular volumes are associated with higher dementia severity, supporting the idea that brain atrophy, which leads to enlarged ventricles, is linked to more severe dementia.

When all plots were evaluated collectively, it was found that the PDPs provided valuable insights into the relationships between predictor variables and dementia severity. Higher cognitive scores (MMSE, LOGIMEM, Simon Number) and greater brain thickness (RH Lingual Thickness) are associated with lower dementia severity whereas, larger ventricular volumes and the presence of the APOE-e4 allele are linked to higher dementia severity. Multivitamin use shows a marginal association with lower severity.

These interpretations are consistent with the findings of the statistical models examined in this study. Hence, these results underline the importance of these variables in predicting dementia severity. The consistency of these findings across different modeling approaches suggests robust relationships between these predictors and dementia severity, despite differences in model parametrization. (In Figure 4.14, “1” refers to “No”, while “2” stands for “Yes”.)

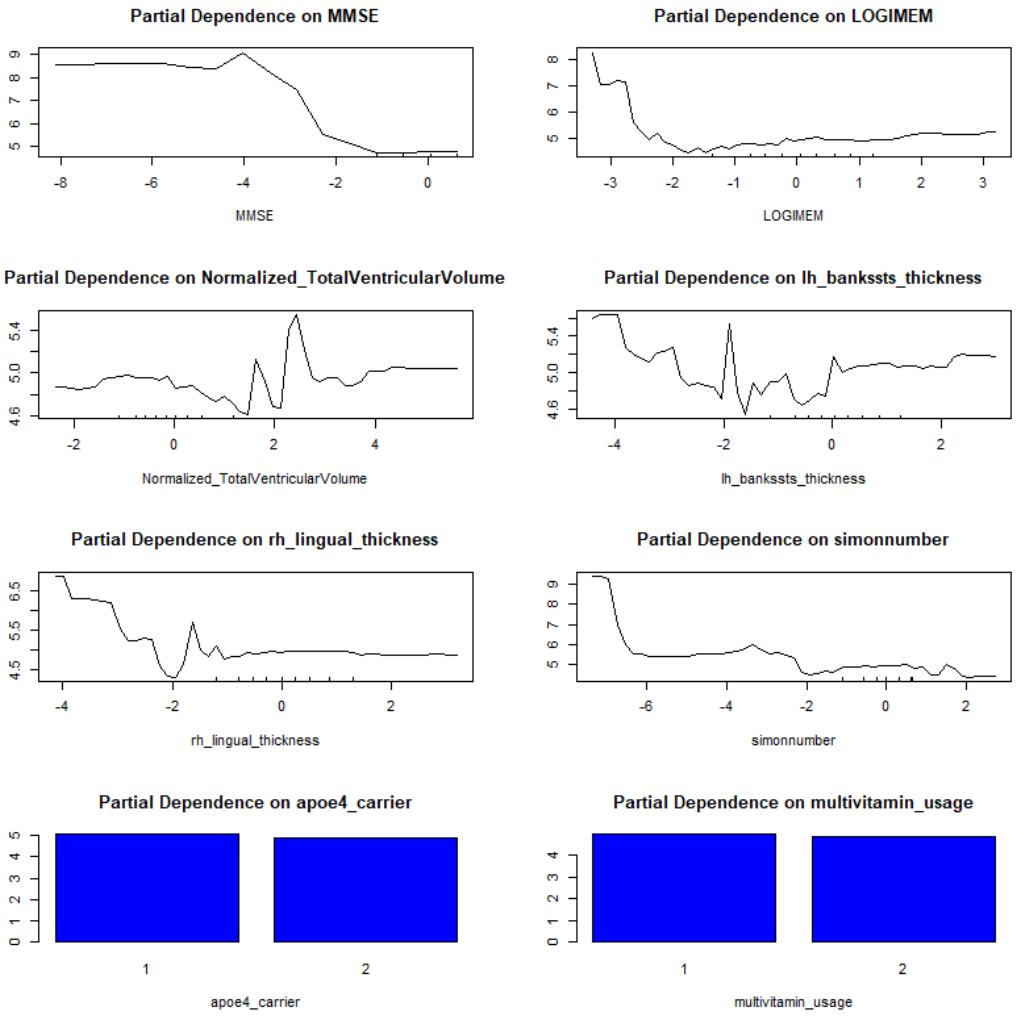


Figure 4.14 Partial Dependence Plots of covariates used in OMERF

4.5 Comparison of Algorithms

In this section, all the statistical and machine learning models mentioned in the previous sections were used to predict the severity level of dementia status using the test dataset. The performance of each model was evaluated using various metrics such as accuracy, sensitivity, specificity and F1 score. The F1 score measure offers a balanced evaluation of the model's performance by considering positive and negative instances. A higher F1 score, in particular, reflects a model's superior overall performance, especially given the imbalanced distribution of the response variable. As such, the F1 score is considered more informative compared to other performance metrics (Nguyen, 2019).

Table 4.20. provides the performance metrics of models in predicting the severity level of dementia using the test dataset. The average accuracy values range from 0.81 to 0.91, indicating the overall correctness of the models' predictions. The models with the highest average accuracy are the Marginal model with ordgee, the Cumulative Link Mixed Effect Model (clmm) and both the Bayesian Mixed Effect Model approach (MCMCglmm and brms), achieving an average accuracy of 0.91. Average sensitivity values, range from 0.80 to 0.92, indicating the models' ability to correctly identify positive instances. The model with the highest average sensitivity is the Bayesian Mixed Effect Model with the brms model, achieving an average sensitivity of 0.92. Average specificity values range from 0.41 to 0.84 reflecting the models' ability to correctly identify negative instances. The model with the highest average specificity is the Bayesian Mixed Effect Model with the MCMCglmm model, achieving an average specificity score of 0.84. The average F1 scores, which provide a balanced measure of precision and recall, range from 0.80 to 0.93. The models with the highest average F1 score are the Bayesian Mixed Effect Models, both achieving an average F1 score of 0.93. The Marginal model with ordgee, the mixed effect model with clmm, and the Bayesian mixed effect models (MCMCglmm and brms) consistently demonstrated high and similar performance across various metrics.

Table 4.20 Test data performance (Overall results)

	Accuracy	**Sensitivity**	**Specificity**	**F1 Score**
ordLORgee	0.87	0.87	0.64	0.86
ordgee	0.91	0.91	0.83	0.92
glmmLasso	0.86	0.86	0.65	0.86
clmm	0.91	0.91	0.80	0.92
MCMCglmm	0.91	0.91	0.84	0.93
brms	0.91	0.92	0.81	0.93
OMERF	0.81	0.80	0.41	0.80

Although overall performance metrics provide valuable information; to underline the reliability and robustness of the models and their prediction performances, category-specific metrics should be examined. In that regard, F1-Score, which provides better information when working with unbalanced distributed response variables, and Recall, which highlights whether the models performed well in predicting certain categories such as the severe category of Mild Dementia, were depicted.

Hence, the F1-Score by class values suggests that the Bayesian Mixed Effect Model using the MCMCglmm algorithm has the highest F1-Score in predicting the most severe dementia category, Mild Dementia, as shown in Table 4.21. Following this, the Marginal model with ordgee reported an F1-Score of 0.50 for this category. However, the machine learning algorithm modified for this thesis, OMERF, underperformed in all dementia categories compared to the other approaches.

One of the key factors contributing to the lower F1 scores in the more severe dementia categories, "Very Mild Dementia" and "Mild Dementia," is the unbalanced nature of the data. The distribution of dementia categories in the dataset is heavily skewed, with 88.5% of the sample representing "No Dementia," while only 9.7% and 1.8% fall into the "Very Mild Dementia" and "Mild Dementia" categories, respectively. This imbalance in the data, coupled with the limited sample sizes in the

more severe categories, weakens the models' ability to predict these cases accurately. As the algorithms were tested on the same data distribution, the smaller sample sizes for the more severe dementia categories limited their capacity for correct classification.

Table 4.21 F1 Scores by Class

	No Dementia	**Very Mild Dementia**	**Mild Dementia**
ordLORgee	0.94	0.34	0.40
ordgee	0.95	0.45	0.50
glmmLasso	0.95	0.33	0.35
clmm	0.95	0.48	0.46
MCMCglmm	0.96	0.49	0.56
brms	0.95	0.48	0.46
OMERF	0.90	0.22	0.27

This explanation holds for Recall as well. The Cumulative Link Mixed Model (CLMM) and the Bayesian Mixed Effect Model (brms) both achieved perfect Recall scores for "Mild Dementia," the most severe category, with a value of 1.00 (Table 4.22). Additionally, MCMCglmm outperformed other models in identifying the second category, "Very Mild Dementia," with a Recall score of 0.75. However, due to the imbalance in the data distribution, the models struggled to correctly classify the few available cases of "Very Mild Dementia" and "Mild Dementia."

Table 4.22 Recall by Class

	No Dementia	**Very Mild Dementia**	**Mild Dementia**
ordLORgee	0.95	0.36	0.33
ordgee	0.92	0.69	0.67
glmmLasso	0.96	0.32	0.26
clmm	0.93	0.62	1.00
MCMCglmm	0.93	0.75	0.63
brms	0.92	0.71	1.00
OMERF	0.91	0.19	0.40

Considering the performance across both F1-Score and Recall, the Bayesian approaches (MCMCglmm and brms) consistently stood out as the top-performing models. The Marginal Model with ordgee and the Cumulative Link Mixed Models (clmm) also performed well but slightly behind the Bayesian models. In general, the OMERF algorithm performed poorly considering other approaches. One of the main reasons for this occurrence is due to the model specification. In the CLMM part of the OMERF algorithm, where the random effect was estimated, no covariates were introduced as random effect to align with all models studied in this thesis. (Only the offset variable was treated as a fixed part and OASIS ID was considered a random effect). Therefore, different model specifications may yield different results, so the comparison between machine learning model and statistical models is subject to this study.

CHAPTER 5

CONCLUSION AND DISCUSSION

In this study, we presented a longitudinal data analysis on the multimodal non-neuroimaging OASIS-3 dataset, which includes demographic information, MRI scans, PET scans and numerous assessment scores ranging from cognitive to physical evaluation. AD is a progressive neurological condition characterized by brain shrinkage, which ultimately leads to severe impairments in an individual's cognitive abilities and memory. Although the main cause of AD is not fully understood, several factors such as genetics and environmental influences are thought to play a role (NIH National Institute on Aging, 2024). The OASIS-3 dataset offers a comprehensive resource for exploring these factors through its multimodal design.

Previous studies on OASIS-3 focused primarily on MRI image processing. Yet, to our knowledge, this is the first time a comprehensive longitudinal study was conducted on ordinal response utilizing non-imaging assessments of the OASIS-3 dataset. Although Jahan et al. (2023) conducted similar work, they relied solely on machine learning, disregarding the longitudinal structure and intraclass correlations.

We used intuitive feature selection approaches, including combining variance and missingness threshold methods, MRMR feature selection, backward stepwise elimination based on AIC, and addressing multicollinearity. The final set of covariates aligns well with those commonly identified in AD classification studies. Our study systematically explored potential factors influencing AD severity classification by combining statistical models and machine learning algorithms, all employing cumulative link models to accommodate the ordinal response with three severity levels.

The results showed that increases in MMSE, Simon Number, LOGIMEM, RH Lingual Thickness, and LH Bankssts Thickness are associated with a reduced risk of dementia. The presence of Apolipoprotein E of allele 4 is associated with an increased risk of having AD. Yet, despite ongoing debate, multivitamin usage appears to have a positive effect on cognitive function, possibly indicating a protective role against dementia. Additionally, an increase in total ventricular volume, due to brain tissue shrinkage, is observed in individuals diagnosed with AD compared to non-demented individuals.

A significant challenge in this study was managing the large and diverse datasets to achieve a cohesive multimodal perspective. Merging MRI and PET scan data with non-imaging assessments was complex, particularly given the irregular time points, dropout rates, and missing data. The imbalance in the dataset, where severe AD patients were less likely to have MRI scans, likely influenced prediction performance.

In terms of modeling, brms and MCMCglmm posed challenges due to the complexity of Bayesian algorithms. Additionally, both ordgee and MCMCglmm provide only probit links, so we applied a probit-logit conversion to ensure comparability with other models, which may have slightly affected consistency across results.

Despite these findings, the study faced several limitations. Firstly, the irregular visits and dropout rates, whether due to death or lack of follow-up, significantly affected the merging and matching-up sessions and results. Secondly, the limited availability of machine learning models for analyzing ordinal responses with random effects made this study predominantly focus on statistical models. Lastly, evaluating model performance on an unseen test dataset was challenging due to the lack of built-in functions for predicting severity classes. Despite we proposed custom prediction functions for each method, further evaluation is necessary.

For future works, several steps can be considered to enhance the robustness of the study. The MRMR feature selection method used here, which was originally

designed for genomic data, does not account for temporal dependence among the cluster. Although Radovic, M., et al. (2017) extended the algorithm to address the issue, a suitable algorithm for longitudinal data with ordinal responses that maintains the model performance is still needed. Additionally, while we emphasized AD classification through various methods such as marginal models, mixed-effects models and machine learning algorithms, incorporating transition models and deep learning algorithms could provide more comprehensive comparisons. Finally, introducing cure models (survival analysis) on the OASIS-3 dataset may be crucial as identification of the individuals with a high probability of being cured could eliminate excessive treatments.

REFERENCES

Agresti, A. (2002) Categorical Data Analysis. 2nd Edition, John Wiley & Sons, Inc., New York, 320-332. http://dx.doi.org/10.1002/0471249688

Akaike, H. (1974). A New Look at the Statistical Model Identification. In: Parzen, E., Tanabe, K., Kitagawa, G. (eds) Selected Papers of Hirotugu Akaike. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-1694-0_16

Archer, K.J., Hou, J., Zhou, Q., Ferber, K., Layne, J.G. and Gentry, A.E. (2014) Ordinalgmifs: An R Package for Ordinal Regression in High-Dimensional Data Settings. Cancer Informatics, 13, CIN.S20806. https://doi.org/10.4137/CIN.S20806

Asar, O., & Ilk, O. (2013). Marginally specified models for analyzing multivariate longitudinal binary data. arXiv preprint arXiv:1310.5440

Bergonzoli, G., Rossi, L., & Masci, C. (2024). Ordinal Mixed-Effects Random Forest. arXiv preprint arXiv:2406.03130

Bergonzoli, G. (2023). Ordinal Mixed-Effects Random Forest. https://www.politesi.polimi.it/retrieve/dc4a6762-26ea-4d86-8d6a-ae7c4bd214c3/

Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324

Breslow, N. E., & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association, 88(421), 9–25. https://doi.org/10.1080/01621459.1993.10594284

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Buckner, R. L., Head, D., Parker, J., Fotenos, A. F., Marcus, D., Morris, J. C., & Snyder, A. Z. (2004). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. NeuroImage, 23(2), 724–738. https://doi.org/10.1016/j.neuroimage.2004.06.018

Cascarano, A., Mur-Petit, J., Hernández-González, J., Camacho, M., de Toro Eadie, N., Gkontra, P., Chadeau-Hyam, M., Vitrià, J., & Lekadir, K. (2023). Machine and deep learning for longitudinal biomedical data: A review of methods and applications. Artificial Intelligence Review, 56(Suppl 2), S1711–S1771. https://doi.org/10.1007/s10462-023-10561-w

Cass S. P. (2017). Alzheimer's Disease and Exercise: A Literature Review. Current sports medicine reports, 16(1), 19–22. https://doi.org/10.1249/JSR.0000000000000332

Chandrashekar, G. and Sahin, F. (2014) A Survey on Feature Selection Methods. Computers and Electrical Engineering, 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chen, S. and DuBois Bowman, F. (2011), A novel support vector classifier for longitudinal high-dimensional data and its application to neuroimaging data. Statistical Analy Data Mining, 4: 604-611. https://doi.org/10.1002/sam.10141

Chiu, H. R., Hwang, C. K., Chen, S. Y., Shih, F. Y., Han, H. C., King, C. C., Gilbert, J. R., Fang, C. C., & Oyang, Y. J. (2022). Machine learning for emerging infectious disease field responses. Scientific reports, 12(1), 328. https://doi.org/10.1038/s41598-021-03687-w

Christensen, R. H. B., & Brockhoff, P. B. (2013). Analysis of sensory ratings data with cumulative link models. Journal de la Societe Francaise de Statistique & Revue de Statistique Appliquee, 154(3), 58-79.

Christensen, R.H.B. (2015) Ordinal-Regression Models for Ordinal Data. R Package Version.

Clayton, D. (1992). Repeated ordinal measurements: A generalised estimating equation approach. Medical Research Council Biostatistics Unit Technical Reports Cambridge, England.

Conaway, M. (1990). A random effects model for binary data. Biometrics 46, 317-328.

Cowles, M. K., & Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. Journal of the American Statistical Association, 91(434), 883–904. https://doi.org/10.1080/01621459.1996.10476956

Cui, R., Liu, M., & Alzheimer's Disease Neuroimaging Initiative (2019). RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society, 73, 1–10. https://doi.org/10.1016/j.compmedimag.2019.01.005

da Silva, J.L., Colosimo, E.A., & Demarqui, F. (2019). A general GEE framework for the analysis of longitudinal ordinal missing data and related issues. Statistical Modelling, 19, 174 - 193.

De Jay, N., Papillon-Cavanagh, S., Olsen, C., El-Hachem, N., Bontempi, G., & Haibe-Kains, B. (2013). mRMRe: an R package for parallelized mRMR ensemble feature selection. Bioinformatics (Oxford, England), 29(18), 2365–2368. https://doi.org/10.1093/bioinformatics/btt383

Diggle, P. (2002). Analysis of longitudinal data. Oxford university press.

Diwate, R. B., Ghosh, R., Jha, R., Sagar, I., & Singh, S. K. (2021, September). Dementia Prediction Using OASIS Data for Alzheimer's Research. In 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV) (pp. 1-7). IEEE.

Fahrmeir, L., Tutz, G. (1994). Models for multicategorical responses: multivariate extensions of generalized linear models. In: Multivariate Statistical Modelling Based on Generalized Linear Models. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4899-0010-4_3

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. Behavior research methods, 50(5), 2016–2034. https://doi.org/10.3758/s13428-017-0971-x

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, & Rubin, D.B. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2004). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Goeman J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. Biometrical journal. Biometrische Zeitschrift, 52(1), 70–84. https://doi.org/10.1002/bimj.200900028

Greene, W. H., & Hensher, D. A. (2010). Modeling Ordered Choices: A Primer. Cambridge: Cambridge University Press.

Groll, A., & Tutz, G. (2012). Variable selection for generalized linear mixed models by L1-penalized estimation. Statistics and Computing, 24, 137 - 154.

Haario, H., Saksman, E., & Tamminen, J. (2001). An Adaptive Metropolis Algorithm. Bernoulli, 7(2), 223–242. https://doi.org/10.2307/3318737

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. Journal of Statistical Software, 33(2), 1–22. https://doi.org/10.18637/jss.v033.i02

Højsgaard, S. ., Halekoh, U., & Yan, J. (2005). The R Package geepack for Generalized Estimating Equations. Journal of Statistical Software, 15(2), 1–11. https://doi.org/10.18637/jss.v015.i02

Handels, R., Jönsson, L., Garcia-Ptacek, S., Eriksdotter, M., & Wimo, A. (2020). Controlling for selective dropout in longitudinal dementia data: Application to the SveDem registry. Alzheimer's & dementia : the journal of the Alzheimer's Association, 16(5), 789–796. https://doi.org/10.1002/alz.12050

Hardan, L., Filtchev, D., Kassem, R., Bourgi, R., Lukomska-Szymanska, M., Tarhini, H., Salloum-Yared, F., Mancino, D., Kharouf, N., & Haikel, Y. (2021). COVID-19 and Alzheimer's Disease: A Literature Review. Medicina (Kaunas, Lithuania), 57(11), 1159. https://doi.org/10.3390/medicina57111159

Hastie, T., Taylor, J., Tibshirani, R., & Walther, G. (2007). Forward stagewise regression and the monotone lasso.

Heagerty, P. J., & Zeger, S. L. (1996). Marginal Regression Models for Clustered Ordinal Measurements. Journal of the American Statistical Association, 91(435), 1024–1036. https://doi.org/10.1080/01621459.1996.10476973

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. Biometrics, 50(4), 933–944.

Hornung, R. Ordinal Forests. J Classif 37, 4–17 (2020). https://doi.org/10.1007/s00357-018-9302-x

Jiang, Y., Man, Q., Liu, Z., Wang, Y., Suo, C., Jin, L., Dong, Q., Cui, M., & Chen, X. (2020). Temporal trends in the mortality rate of Alzheimer's disease and other dementias attributable to smoking, 1990-2017. Environmental research, 184, 109183. https://doi.org/10.1016/j.envres.2020.109183

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models, Computational Statistics & Data Analysis, Elsevier, vol. 52(12), pages 5066-5074, August.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995) Continuous Univariate Distributions, Vol. 2. J. Wiley & Sons, New York, Chap. 28.

Molenberghs, G. and Lesaffre, E. (1994) Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. J. Am. Stat. Assoc. 89, 633–644.

Koenig, L. N., Day, G. S., Salter, A., Keefe, S., Marple, L. M., Long, J., LaMontagne, P., Massoumzadeh, P., Snider, B. J., Kanthamneni, M., Raji, C. A., Ghoshal, N., Gordon, B. A., Miller-Thomas, M., Morris, J. C., Shimony, J. S., Benzinger, T. L. S., & Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Dominantly Inherited Alzheimer Network (DIAN) (2020). Select Atrophied Regions in Alzheimer disease (SARA): An improved volumetric model for identifying Alzheimer disease dementia. NeuroImage. Clinical, 26, 102248. https://doi.org/10.1016/j.nicl.2020.102248

LaMontagne, P. J., Benzinger, T. L., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., . . . Vlassenko, A. J. M. (2019). OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease.

Lee, K., & Daniels, M. J. (2007). A class of markov models for longitudinal ordinal data. Biometrics, 63(4), 1060–1067. https://doi.org/10.1111/j.1541-0420.2007.00800.x

Leszek, J., Mikhaylenko, E. V., Belousov, D. M., Koutsouraki, E., Szczechowiak, K., Kobusiak-Prokopowicz, M., Mysiak, A., Diniz, B. S., Somasundaram, S. G., Kirkland, C. E., & Aliev, G. (2021). The Links between Cardiovascular Diseases and Alzheimer's Disease. Current neuropharmacology, 19(2), 152–169. https://doi.org/10.2174/1570159X18666200729093724

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. J. Multivar. Anal., 100, 1989-2001.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. Biometrika, 73(1), 13–22. https://doi.org/10.2307/2336267

Lin, X. and Breslow, N.E. (1996) Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion. Journal of the American Statistical Association, 91, 1007-1016. http://dx.doi.org/10.1080/01621459.1996.10476971

Lipsitz, S. R., Kim, K., & Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. Statistics in medicine, 13(11), 1149–1163. https://doi.org/10.1002/sim.4780131106

Lumley, T. (1996). Generalized Estimating Equations for Ordinal Data: A Note on Working Correlation Structures. Biometrics, 52(1), 354–361. https://doi.org/10.2307/2533173

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). 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. https://doi.org/10.1162/jocn.2007.19.9.1498

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. Journal of cognitive neuroscience, 22(12), 2677–2684. https://doi.org/10.1162/jocn.2009.21407

Markoulidakis, I., Kopsiaftis, G., Rallis, I., & Georgoulas, I. (2021). Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem. Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference.

Miller, M.E., Davis, C.S. and Landis, J.R. (1993) The Analysis of Longitudinal Polytomous Data: Generalized Estimating Equations and Connections with Weighted Least Squares. Biometrics, 49, 1033-1044. https://doi.org/10.2307/2532245

Morris J. C. (1993). The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology, 43(11), 2412–2414. https://doi.org/10.1212/wnl.43.11.2412-a

Naidu, C., Kumar, D., Maheswari, N., Sivagami, M., & Li, G. (2019). Prediction of alzheimer's disease using oasis dataset.

Nestor, S. M., Rupsingh, R., Borrie, M., Smith, M., Accomazzi, V., Wells, J. L., Fogarty, J., Bartha, R., & Alzheimer's Disease Neuroimaging Initiative (2008). Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. Brain : a journal of neurology, 131(Pt 9), 2443–2454. https://doi.org/10.1093/brain/awn146

Nguyen, M. H. (2019). Impacts of unbalanced test data on the evaluation of classification methods. International Journal of Advanced Computer Science and Applications, 10(3).

Ni, X., Zhang, D., and Zhang, H. H. (2010). Variable selection for semiparametricmixed models in longitudinal studies. Biometrics 66 79–88

NIH National Institute on Aging, (2024). What causes Alzheimer's Disease? https://www.nia.nih.gov/health/alzheimers-causes-and-risk-factors/what-causes-alzheimers-disease

O'Bryant, S. E., Humphreys, J. D., Smith, G. E., Ivnik, R. J., Graff-Radford, N. R., Petersen, R. C., & Lucas, J. A. (2008). Detecting dementia with the mini-mental state examination in highly educated individuals. Archives of neurology, 65(7), 963–967. https://doi.org/10.1001/archneur.65.7.963

Parsons NR, Edmondson RN, Gilmour SG (2006). “A Generalized Estimating Equation Method for Fitting Autocorrelated Ordinal Score Data with an Application in Horticultural Research.” Journal of the Royal Statistical Society C, 55(4), 507–524.

Pellagatti, M., Masci, C., Ieva, F., & Paganoni, A.M. (2021). Generalized mixed‐effects random forest: A flexible approach to predict university student dropout. Statistical Analysis and Data Mining: The ASA Data Science Journal, 14, 241 - 257.

Profyri, E., Leung, P., Huntley, J., & Orgeta, V. (2022). Effectiveness of treatments for people living with severe dementia: A systematic review and meta-analysis of randomised controlled clinical trials. Ageing research reviews, 82, 101758. https://doi.org/10.1016/j.arr.2022.101758

Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC bioinformatics, 18(1), 9. https://doi.org/10.1186/s12859-016-1423-9

Sacchi, L., Arcaro, M., Carandini, T., Pietroboni, A. M., Fumagalli, G. G., Fenoglio, C., Serpente, M., Sorrentino, F., Visconte, C., Pintus, M., Conte, G., Contarino, V. E., Scarpini, E., Triulzi, F., Galimberti, D., & Arighi, A. (2023). Association between enlarged perivascular spaces and cerebrospinal fluid aquaporin-4 and tau levels: report from a memory clinic. Frontiers in aging neuroscience, 15, 1191714. https://doi.org/10.3389/fnagi.2023.1191714

Salehipour, A., Dolatshahi, M., Haghshomar, M., & Amin, J. (2023). The Role of Thyroid Dysfunction in Alzheimer's Disease: A Systematic Review and Meta-Analysis. The journal of prevention of Alzheimer's disease, 10(2), 276–286. https://doi.org/10.14283/jpad.2023.20

SAS (2008) Statistical Analysis Systems Institute. Version 9.1, SAS Institute Inc., Cary, North Carolina, USA.

Schelldorfer, J., Meier, L., & Bühlmann, P. (2014). GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using ℓ1-Penalization. Journal of Computational and Graphical Statistics, 23(2), 460–477. https://doi.org/10.1080/10618600.2013.773239

SCHELLDORFER, J., BÜHLMANN, P., & VAN DE GEER, S. (2011). Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ-1 Penalization. Scandinavian Journal of Statistics, 38(2), 197–214. http://www.jstor.org/stable/23015490

Simon, J. R. (1969). Reactions toward the source of stimulation. Journal of Experimental Psychology, 81(1), 174–176. https://doi.org/10.1037/h0027448

Skrondal, A., & Rabe-Hesketh, S. (2004). Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780203489437

Stan Development Team (2024). "RStan: the R interface to Stan." R package version 2.32.6, https://mc-stan.org/.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Tahami Monfared, A. A., Byrnes, M. J., White, L. A., & Zhang, Q. (2022). Alzheimer's Disease: Epidemiology and Clinical Progression. Neurology and therapy, 11(2), 553–569. https://doi.org/10.1007/s40120-022-00338-8

Taylor, J. E., Rousselet, G. A., Scheepers, C. and Sereno, S. C. (2023) Rating norms should be calculated from cumulative link mixed effects models. Behavior Research Methods, 55, pp. 2175-2196. (doi: 10.3758/s13428-022-01814-7) (PMID:36103049) (PMCID:PMC10439063)

Thomas R. Ten Have. (1996). A Mixed Effects Model for Multivariate Ordinal Response Data Including Correlated Discrete Failure Times with Ordinal Responses. Biometrics, 52(2), 473–491. https://doi.org/10.2307/2532888

Touloumis, A. (2015). R Package multgee: A Generalized Estimating Equations Solver for Multinomial Responses. Journal of Statistical Software, 64(8), 1–14. https://doi.org/10.18637/jss.v064.i08

Touloumis, A., Agresti, A., & Kateri, M. (2013). GEE for multinomial responses using a local odds ratios parameterization. Biometrics, 69(3), 633–640. https://doi.org/10.1111/biom.12054

Tutz, G. & Hennevogl, W. (1996). "Random effects in ordinal regression models," Computational Statistics & Data Analysis, Elsevier, vol. 22(5), pages 537-557, September.

T.U.O.Q. (2018), Queensland Brain Institute, “What causes dementia?” https://qbi.uq.edu.au/dementia/dementia-causes-and-treatment

Tylee, D. S., Kikinis, Z., Quinn, T. P., Antshel, K. M., Fremont, W., Tahir, M. A., Zhu, A., Gong, X., Glatt, S. J., Coman, I. L., Shenton, M. E., Kates, W. R., & Makris, N. (2017). Machine-learning classification of 22q11.2 deletion syndrome: A diffusion tensor imaging study. NeuroImage. Clinical, 15, 832–842. https://doi.org/10.1016/j.nicl.2017.04.029

Tzeng, R. C., Yang, Y. W., Hsu, K. C., Chang, H. T., & Chiu, P. Y. (2022). Sum of boxes of the clinical dementia rating scale highly predicts conversion or reversion in predementia stages. Frontiers in aging neuroscience, 14, 1021792. https://doi.org/10.3389/fnagi.2022.1021792

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in medicine, 18(6), 681–694. https://doi.org/10.1002/(sici)1097-0258(19990330)18:6<681::aid-sim71>3.0.co;2-r

Wang, M., Song, L., Wang, X. (2010). Bridge estimation for generalized linear models with a divergingnumber of parameters. Statistics and Probability Letters 80:1584–1596.

Wechsler, D. (1987). WMS-R: Wechsler Memory Scale-Revised: Manual. San Antonio, TX: Psychological Corporation.

Yan, J., Højsgaard, S., Halekoh, U. (2012a). geepack: Generalized estimating equation package, http://CRAN.R-project.org/package=geepack, R package version 1.1-6.

Yeung, L. K., Alschuler, D. M., Wall, M., Luttmann-Gibson, H., Copeland, T., Hale, C., Sloan, R. P., Sesso, H. D., Manson, J. E., & Brickman, A. M. (2023). Multivitamin Supplementation Improves Memory in Older Adults: A Randomized Clinical Trial. The American journal of clinical nutrition, 118(1), 273–282. https://doi.org/10.1016/j.ajcnut.2023.05.011

Zhang, J., Liu, M., Le An, Gao, Y., & Shen, D. (2017). Alzheimer's Disease Diagnosis Using Landmark-Based Features From Longitudinal Structural MR Images. IEEE journal of biomedical and health informatics, 21(6), 1607–1616. https://doi.org/10.1109/JBHI.2017.2704614

APPENDICES

A. R Code for Data Cloning (Slightly Modified Version of OMERF)

The following R scripts were taken from G Bergonzoli's (2024) work and were slightly revised to be applied to the random intercept model. The offset variable in the clmm model formulation, which is used to estimate the random effect part in the algorithm, seems not to be working. Therefore, to get the benefit of offset variable in the formula, the formula was revisited by utilizing the "stats" package in R. Besides, parameters used in the function are defined as follows:

```
#y = response variable in the dataset
#cov = all covariates except response, time and grouping variable
#group = grouping variable
#xnam = all covariates (fixed variable) to be used in the random forest algorithm
#znam = NULL
#bizero = NULL
#itmax = 100
#toll = 0.05
OMERF_modified= function (y, cov, group, xnam=NULL, znam=NULL, bizero=NULL,
              itmax=100, toll=0.05) {

 ########################################
 ######        STEP 1: Initialization  ######
 ########################################

 N <- length(y) #number of observations
 n=length(levels(group)) #number of groups
```

```
if(is.null(znam)) {
  q = length(znam) + 1
} else {
  q = 1
}

#q <- length(znam)+1          #number of random covariates + random intercept
Zi=NULL

z.not.null=!(is.null(znam))
if(z.not.null) {
  Zi = cov[NULL]
} else {
  Zi = cov[znam]
}
#check whether there are covariates included in the random effects
#if (z.not.null) Zi=cov[znam] #random effects covariates
Zi.int= cbind(rep(1,N),Zi) #random intercept + random effects covariates

#Initialize (if NULL) bi to 0
if( is.null(bizero) ){
  bi <- NULL
  for(i in 1:n) bi=cbind(bi,rep(0,q))
}
if( !is.null(bizero) ) bi=bizero
lev=levels(group) #group names
bi=data.frame(bi)
names(bi)=lev
all.bi=list()  #b_i of each iteration
all.bi[[1]]=bi
```

```
  #if xnam is NULL assume that all cov variables are to be used
  if(is.null(xnam)) xnam=names(cov)

  #group must be a factor, otherwise it gives an error
  if(!is.factor(group)) stop('The "group" argument must be a factor')

  forest.formula=as.formula(paste("target ~ ", paste(xnam, collapse= "+")))

  if(z.not.null) {
    clmm.formula=as.formula(paste("y ~ stats::offset(eta.est.rf) + ( 1+", paste(znam,
collapse= "+"), " | group )"))
  } else {
    clmm.formula=as.formula(paste("y ~ stats::offset(eta.est.rf) + ( 1 | group)"))
  }

  ######################################################
  ####  STEP 2: OF for eta_ijc initialization ######
  ######################################################

  library(ordinal)
  library(ordinalForest)
  orf.data= cbind(y,cov[xnam])
  orf_init=ordfor(depvar = "y",perffunction = "probability", orf.data)
  preds=predict(orf_init, newdata=subset(orf.data, select = -y))
  levy=levels(y)
  preds.orf=preds$classprobs
  preds.orf.cum <- t(apply(preds.orf, 1, cumsum))
  colnames(preds.orf.cum) <- levy
  true_y <- as.character(y)
```

```
 preds.orf.cum <- preds.orf.cum[cbind(1:nrow(preds.orf.cum), match(true_y, levy))]
 preds.orf.cum <- round(preds.orf.cum,5)
 preds.orf.cum[preds.orf.cum==0]= 0.00001
 preds.orf.cum[preds.orf.cum==1]= 0.99999
 eta.est.orf = qlogis(preds.orf.cum)
 eta.est.rf = eta.est.orf

 ########################################################
 ######        STEP 3-4: Iterative model estimation  ######
 ########################################################
 library(randomForest)
 it=1
 converged=FALSE
 while(!converged && it<itmax) {

  clmm.data = data.frame(y,cov[znam],group,eta.est.rf)
  clmm.fit= clmm(clmm.formula,link='logit', data=clmm.data,nAGQ = 4 ,Hess=TRUE, control=clmm.control(maxLineIter = 500, maxIter=1000, grtol=1e-3))

  #I want to maintain the order of the elements of bi
  select=c("(Intercept)",znam) #all bi to be extracted
  clmm.bi=ranef(clmm.fit)$group[select]

  bi.old=bi
  bi=data.frame(t(clmm.bi))

  #random forest
  target=rep(0,N) #target=eta-Z%*%b
```

```
for (i in 1:N) {
  b.temp=as.matrix(bi[group[i]], nrow=q, ncol=1)
  z.temp=as.matrix(Zi.int[i,], nrow=1, ncol=q)
  target[i]= eta.est.orf[i] + z.temp%*%b.temp
}
forest.data=cbind(target, cov[xnam])
forest=randomForest(forest.formula, forest.data, importance=TRUE)
eta.est.rf=forest$predicted

#convergence of b_i
names(bi)=lev
diff.t=abs(bi.old-bi)
n.diff=max(diff.t) #use the infinite norm (max)
n.old=max(abs(bi.old))
#converged= n.diff/n.old <toll
#if(it==1) {converged=FALSE} else {converged= n.diff/n.old <toll}
if(n.old==0) {converged=FALSE} else {converged= n.diff/n.old <toll}

it=it+1
all.bi[[it]]=bi

# #convergence of bi
# bi.old=bi
# bi=data.frame(t(clmm.bi))
# names(bi)=lev
# diff.t=abs(bi.old-bi)
# n.diff=max(diff.t) #use the infinite norm (max)
# ind=which(diff.t==n.diff, arr.ind=T)
# n.old=abs(bi.old[ind])
# converged= n.diff/n.old <toll
```

```
  #
  # it=it+1
  # all.bi[[it]]=bi

 }
 ################################################
 ######        STEP 5: Output preparation  ###########
 ################################################

 #if there is no convergence it gives an error message
 if(!converged) {
  warning('Maximum number of iterations exceeded, no convergence achieved')
 }

 result=list(clmm.fit,forest,target,bi,it,converged,all.bi,xnam)
 names(result)=c('clmm.model', 'forest.model','rf.target', 'rand.coef', 'n.iteration',
              'converged','all.rand.coef','forest.var')
 class(result)='omerf'
 result
}

summary.omerf=function(om) {
 print('Mixed effects model') #summary of the mixed effects model
 print(summary(om$clmm.model))
 str=ifelse(om$converged, 'Converged', 'Did not converge')
 print(paste(str , 'after', om$n.iteration, 'iterations')) #says whether there is
convergence
}

fitted.omerf=function(om, group, type='response') {
```

```
 #the correct fitted values are already those of the clmm function,
 #as they incorporate the fitted values of the random forest into the model
 mu_avg=om$clmm.model$fitted.values #for an avg random effect
 f.x_ij=om$forest.model$predicted
 y_train=as.numeric(om$clmm.model$y)

 ranef=rowSums(ranef(om$clmm.model)$gr)
 mu=rep(0,length(mu_avg))
 eta=rep(0,length(mu_avg))
 for (j in 1:length(mu_avg)) {
  i=group[j]
  c=y_train[j]-min(y_train)+1
  if(c==1) {
   eta[j]=om$clmm.model$Theta[c] - ranef[i] - f.x_ij[j]
   mu[j]=plogis(eta[j])
  } else if(c==max(y_train)-min(y_train)+1) {
   eta[j]=qlogis(0.99999999)
   mu[j]=1 - plogis(om$clmm.model$Theta[c-1] - ranef[i] - f.x_ij[j]) # 1 -
P(yij<=c-1)
  } else {
   eta[j]=om$clmm.model$Theta[c] - ranef[i] - f.x_ij[j]
   mu[j]=plogis(eta[j]) -
    plogis(om$clmm.model$Theta[c-1] - ranef[i] - f.x_ij[j]) # P(yij<=c) - P(yij<=c-
1)
  }
 }
 #error if the answer type is not one of the following
 allowed=c('response', 'mu', 'eta')
 msg=paste('Possible choices are ', allowed[1],', ' ,allowed[2],' and ',
       allowed[3], sep='')
```

```
  if(sum(type==allowed)==0) stop('Type of prediction not available:',msg)
  if(type=='mu') ans=mu # P(yij=c)
  if(type=='response') ans=as.factor(y_train)
  if(type=='eta') ans=eta
  ans
}

predict.omerf=function(om, y_train, newdata, group, type='response',
                       predict.all=FALSE) {
  forest.data=newdata[om$forest.var]
  f.x_ij=predict(om$forest.model,forest.data,predict.all=predict.all)

  preds <- as.data.frame(matrix(0, dim(newdata)[1], length(levels(y_train))))
  eta <- as.data.frame(matrix(0, dim(newdata)[1], length(levels(y_train))))
  ranef=rowSums(ranef(om$clmm.model)$group)
  lev=levels(y_train)
  names(preds) = lev
  names(eta) = lev
  for (c in 1:length(lev)) {
    for (j in 1:dim(newdata)[1]) {
      i=group[j]
      if(c==1) {
        eta[j,c]=om$clmm.model$Theta[c] - ranef[i] - f.x_ij[j]
        preds[j,c]=plogis(eta[j,c])
      } else if(c==length(lev)) {
        eta[j,c]=qlogis(0.99999999)
        preds[j,c]=1 - plogis(om$clmm.model$Theta[c-1] - ranef[i] - f.x_ij[j]) # 1 -
P(yij<=c-1)
      } else {
        eta[j,c]=om$clmm.model$Theta[c] - ranef[i] - f.x_ij[j]
```

```
      preds[j,c]=plogis(eta[j,c]) -
       plogis(om$clmm.model$Theta[c-1] - ranef[i] - f.x_ij[j]) # P(yij<=c) -
P(yij<=c-1)
    }
   }
  }
  max_mu <- apply(preds, 1, max)
  pred_val <- names(preds)[apply(preds, 1, which.max)]


  #error if the answer type is not one of the following
  allowed=c('response', 'mu', 'eta')
  msg=paste('Possible choices are ', allowed[1],', ' ,allowed[2],' and ',
          allowed[3], sep='')
  if(sum(type==allowed)==0) stop('Type of prediction not available:',msg)

  if(type=='eta') ans=eta # eta for the predicted class
  if(type=='mu') ans=preds # P(yij=c) c=1,...,C
  if(type=='response') ans=pred_val
  ans
}

ranef.omerf=function(om, group, znam=NULL) {
  ranef=ranef(om$clmm.model)$gr
  n = (length(znam)+1)*length(levels(group))
  step = length(levels(group))
  par(mfrow=c(1,length(znam)+1))
  for (i in 1:(length(znam)+1)) {
    condVar=rep(0,step)
    if(is.null(znam)) { condVar_temp = om$clmm.model$condVar
```

```
   } else { condVar_temp = om$clmm.model$condVar[((i - 1) * step + 1):(i * step),((i - 1) * step + 1):(i * step)] }
   for (l in 1:step) {
    if(is.null(znam)) {
      condVar[l] <- condVar_temp[l]
    } else {condVar[l] <- condVar_temp[l,l]}
   }
   condVar_sqrt <- sqrt(condVar)
   ci <- ranef[,i] + qnorm(0.975) * condVar_sqrt %o% c(-1, 1)
   ord.re <- order(ranef[,i])
   lev <- levels(group)
   ci <- ci[order(ranef[,i]),]
   plot(1:length(lev), ranef[ord.re,i], axes=FALSE, ylim=range(ci),
       xlab="random variable", ylab="random effect")
   axis(1, at=1:length(lev), labels = lev[ord.re]) #,las=2
   axis(2)
   for(k in 1:length(lev)) segments(k, ci[k,1], k, ci[k, 2])
   abline(h = 0, lty=2)
  }
}
```