# Project 2 Longitudinal Data Analysis

*Daniele Capelli (r1084933)*

*Vittorio Carfagno (r1085685)*

*Guillem Olivart Garrofé (r1085574)*

*Theodore Kristoffer Wood (r0817082)*

KU Leuven

Master of Statistics

Geert Molenberghs
Geert Verbeke

December 2025

# Contents

# 1 Introduction

Alzheimer's disease (AD) is one of the most prevalent forms of late-life dementia, representing a major public health challenge with profound social and economic consequences [1]. Characterized by progressive cognitive decline and psychiatric symptoms, AD significantly impacts patients' daily functioning and quality of life, while also placing a considerable burden on caregivers and healthcare systems.

The aim of this report is to investigate the temporal evolution of Alzheimer's disease by employing the Clinical Dementia Rating Sum of Boxes (`CDRSB`) as a proxy measure of disease severity. The `CDRSB` is a scale used to assess the severity of cognitive impairment, especially in Alzheimer's disease. It evaluates 6 domains: memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. The analysis provided in this document is based on data from a clinical trial in which elderly patients were followed across seven observation periods. In addition to `CDRSB`, the dataset includes several categorical and numerical variables that provide further insight into patient characteristics and disease progression.

Our approach begins by taking into account `sex`, `age`, and `BMI`, and then examines the importance of `ADL`, `ABPET`, and `TAUPET`. Moreover, we decided to dichotomize the variable `CDRSB` obtaining two different categories, one for patients with a `CDRSB` score less than or equal to 10, and another for patients with a score greater than 10. Then we develop and compare two statistical frameworks: a marginal model and a generalized linear mixed-effects (random-effects) model. By contrasting the results across these methods, we aim to clarify how different modeling strategies capture the complexity of the data and whether they yield consistent or divergent conclusions regarding the progression of Alzheimer's disease. The code used to get our results can be retrieved in [2].

# 2 Preliminary Data Exploration

## 2.1 Dataset Overview

The analysis is based on a comprehensive longitudinal dataset tracking 1,253 unique patients over a seven-year period (baseline and six annual follow-ups). The dataset captures baseline demographic and clinical information (e.g., `age`, `sex`, `edu`, `bmi`...) as well as annual measurements for primary outcomes like the Brief Psychiatric Rating Scale (`bprs0`...`bprs6`), Clinical Dementia Rating Sum of Boxes (`cdrsb0`...`cdrsb6`), and PET scan results of Amyloid-Beta protein (`abpet0`...`abpet6`) and Tau protein (`taupet0`...`taupet6`).

**`patid`:** A unique identifier for each patient.

**`trial`:** A categorical variable (1-25) indicating the clinical trial or center.

**`sex`:** Categorical (0 = Male, 1 = Female).

**`age`:** Patient's age at baseline.

**`edu`:** Categorical education level (1 = Primary, 2 = Lower Secondary, 3 = Upper Secondary, 4 = Higher).

**bmi:** Body Mass Index at baseline.

**inkomen:** Income at baseline.

**job:** Categorical (0 = No Job, 1 = Job).

**adl:** Activities of Daily Living score at baseline.

**wzc:** Categorical residence (0 = Home, 1 = Residence).

**bprs0...bprs6:** Brief Psychiatric Rating Scale (annual measurements).

**cdrsb0...cdrsb6:** Clinical Dementia Rating Scale Sum of Boxes (annual measurements).

**CDRSB_cat:** Binarized version of CDRSB:

$$\text{CDRSB\_cat} = \begin{cases} 0 & \text{se CDRSB} \leq 10, \\ 1 & \text{se CDRSB} > 10. \end{cases}$$

**abpet0...abpet6:** Amyloid-Beta PET scan results (annual measurements).

**taupet0...taupet6:** Tau PET scan results (annual measurements).

The baseline characteristics of the cohort are detailed in Table 1 and Table 2. In our analysis we will take into consideration standardized values of the `Age`, `BMI`, `ABPET` and `TAUPET` variables.

Table 1: Descriptive Statistics for Baseline Numeric Variables

| Statistic | age | bmi | inkomen | adl | cdrsb0 | abpet0 | taupet0 |
|---|---|---|---|---|---|---|---|
| mean | 72.45 | 25.76 | 2283.80 | 6.86 | 6.73 | 2.32 | 1.92 |
| std | 7.33 | 2.15 | 548.12 | 3.12 | 7.17 | 0.45 | 0.12 |
| min | 46.00 | 19.80 | 1000.00 | 0.00 | 1.00 | 2.00 | 1.90 |
| 25% | 67.00 | 24.20 | 1900.00 | 5.00 | 1.00 | 2.00 | 1.90 |
| 50% | 72.00 | 25.70 | 2300.00 | 6.00 | 2.00 | 2.00 | 1.90 |
| 75% | 77.00 | 27.10 | 2700.00 | 8.00 | 13.00 | 3.00 | 1.90 |
| max | 94.00 | 33.70 | 3800.00 | 20.00 | 19.00 | 3.00 | 2.80 |

A critical feature of this study is the patient dropout over time. As shown in Table 3, the number of available `CDRSB` observations declines from 1,253 at baseline to 511 by year 6, resulting in a highly unbalanced data structure.

Dropout was discussed, and a Pearson's Chi-squared test indicated no evidence of informative dropout ($\chi^2 = 1.9588$, $p$-value $= 0.1616$), suggesting that dropout was not associated with the response variable.

Table 2: Distribution of Baseline Categorical Variables

| Variable | Category | Proportion |
|---|---|---|
| Sex | Female | 50.8% |
| | Male | 49.2% |
| Education | Higher | 37.2% |
| | Upper Secondary | 30.7% |
| | Lower Secondary | 19.6% |
| | Primary | 12.5% |
| Job | No Job | 91.7% |
| | Job | 8.3% |
| Residence (WZC) | Home | 61.1% |
| | Residence | 38.9% |

## 2.2 Exploratory Data Analysis

To guide the choice of an appropriate modeling strategy, we first examine the longitudinal trajectory of our primary outcome, the Clinical Dementia Rating Sum of Boxes (`CDRSB`). Since the outcome is binary (coded as 0/1), it is informative to consider the proportion of observations equal to 1 over time, as depicted in Figure 1. This proportion shows a clear upward trend and appears to follow the characteristic S-shaped curve commonly associated with logistic and probit functions.

The second stage of our preliminary analysis investigates how the covariates of interest (`Sex`, `Age`, `BMI`, `ADL`, `AB`, and `TAU`) influence the progression of `CDRSB` over time. Figure 2 illustrates the temporal patterns of the proportion of outcome 1 across the different covariate groups. Clear differences emerge among the groups defined by `TAU`, `Age`, and `BMI`, while modest ones appear linked to the variable `ADL`. Regarding `Sex`, a modest divergence appears at the beginning, though it is not particularly pronounced. In contrast, the trajectories seem unaffected by `AB`.

Table 3: Patient Dropout: Non-Missing CDRSB Observations

| Time Point | Number of Patients |
|---|---|
| cdrsb0 | 1253 |
| cdrsb1 | 1106 |
| cdrsb2 | 1014 |
| cdrsb3 | 907 |
| cdrsb4 | 777 |
| cdrsb5 | 652 |
| cdrsb6 | 511 |

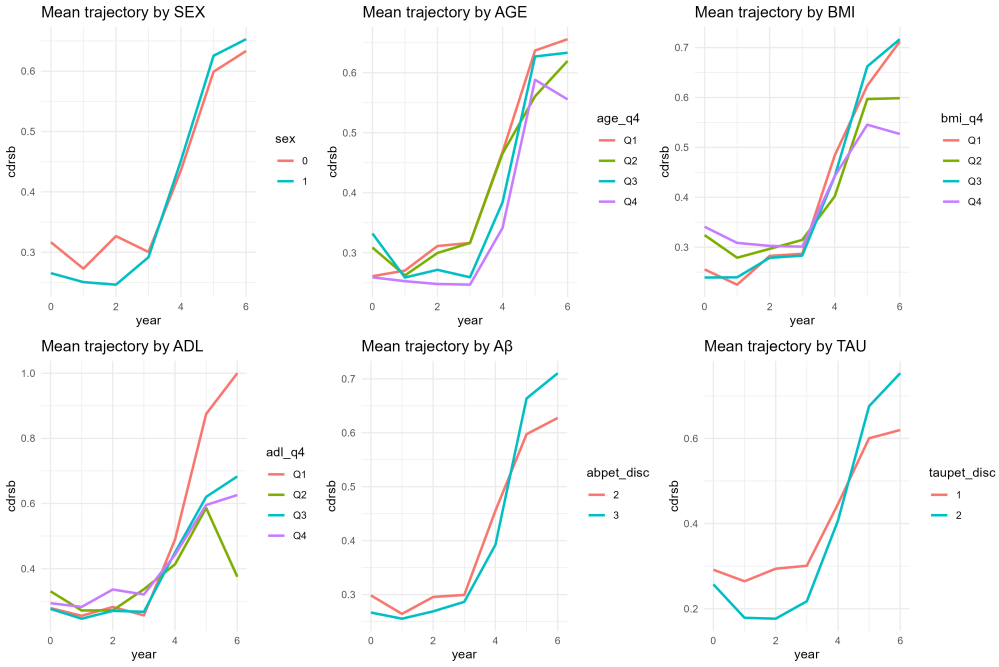


Figure 2: Fraction of outcome 1 over the total observed patients over time in correspondence of different groups. Here we group `Age`, `BMI` and `ADL` in their quartiles and we dichotomize the values of `AB` and `TAU`.

In addition to examining the mean structure, we also explore the variance structure. To this end, we consider the plot of the empirical variance of `CDRSB` over time, presented in Figure 3. Now, from a Bernoulli model with mean (so probability of success) $\pi$, the variance is simply given by $\pi(1-\pi)$. So, the variance is maximal when the probability

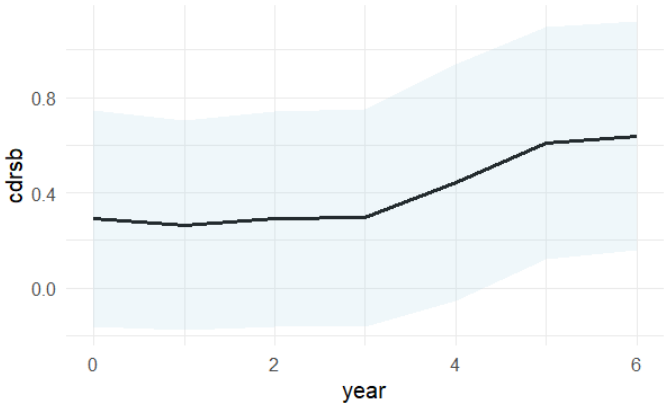


Figure 1: Temporal evolution of the number of outcomes 1 over the total observed patients at that time. The shaded region represents the 95 percent confidence interval around this value.

reaches value 0.5. In our case, the probability seems to reach its maximal value at year 4, coherently with the fact that in Figure 1 the 0.5 value in the average outcome is reached at that time. However, our attention is captured by the fact that in the first part of the study (*i.e.*, in years 0-3) the variance decreases and then increases. This suggests that a simple logit function linear over time could not be able to capture the evolution of the `CDRSB` outcome, and so non-linearity and/or random effects will be studied.

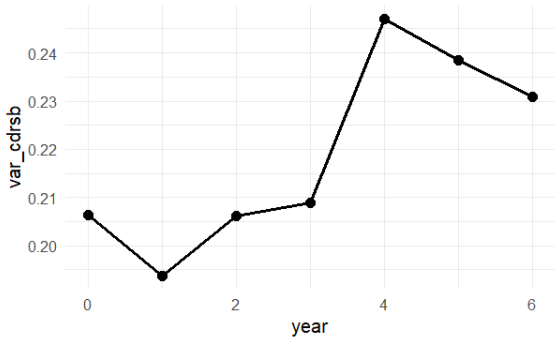


Figure 3: Change in `CDRSB` variance over the 7-year study period.

We finally turn to the empirical correlation structure, as reported in (1). Interestingly enough, the correlation between successive measurements diminishes over time, eventually becoming negative after year 4 ($t > 4$). When studying binary outcomes, this finding provides key insights into the underlying probability dynamics. The initial decrease in positive correlation reflects that the relative fraction of the prevalent outcome at the beginning of the study (outcome 0, in this case) is progressively declining. Crucially, the change in the correlation sign (at $t = 4$) signifies a complete inversion of the prevalent outcome: the outcome that was initially less frequent (outcome 1) has now become the majority state or the most probable event. In addition, after this change in prevalence, the correlation between measurements starts to increase as the time span decreases. This confirms that the probability of Outcome 1 is steadily rising in the second phase of the

study (and overall rising in the course of the study), reinforcing its status as the new predominant outcome and leading to higher correlation among proximate observations.

$$\begin{pmatrix} 1.00 & 0.58 & 0.39 & 0.05 & -0.40 & -0.66 & -0.81 \\ 0.58 & 1.00 & 0.35 & 0.07 & -0.31 & -0.54 & -0.68 \\ 0.39 & 0.35 & 1.00 & 0.05 & -0.09 & -0.33 & -0.47 \\ 0.05 & 0.07 & 0.05 & 1.00 & 0.06 & 0.06 & 0.02 \\ -0.40 & -0.31 & -0.09 & 0.06 & 1.00 & 0.37 & 0.39 \\ -0.66 & -0.54 & -0.33 & 0.06 & 0.37 & 1.00 & 0.60 \\ -0.81 & -0.68 & -0.47 & 0.02 & 0.39 & 0.60 & 1.00 \end{pmatrix} \tag{1}$$

# 3 Methods

After the first insight on our data, we want to apply more accurate statistical tools to model the evolution of the `CDRSB` quantity. We will look at two different approaches. First, we will fit a marginal model to the data in order to get a first insight on how the probability curve evolves over time. Second, we will try to better capture the longitudinal structure of our dataset by including random effects in the model.

## 3.1 Marginal Models

First we apply a marginal model to quantify the population-averaged association between covariates and the probability of severe dementia. This approach specifies the marginal mean, $\mu_{ij} = \mathbb{E}[Y_{ij}] = \mathbb{P}(Y_{ij} = 1) = \pi_{ij}$, directly as a function of the covariates, without conditioning on individual random effects. While the primary interest lies in the mean structure, the dependence among repeated observations is explicitly modeled as a nuisance parameter to yield robust standard errors and valid test statistics.

We start by fitting a full Generalized Estimating Equations (GEE) model. First of all, we need to specify a proper link function between the probabilities $\eta_{ij}$ and the covariates. Different choices are possible (such as the logit or the probit functions), but following the easy interpretability in terms of (log)odd-ratios of the coefficients we decide to use the logit function

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \eta_{ij}$$

where $\eta_{ij}$ is a function of the form $\beta^T x_{ij}$ and $x_{ij}$ includes possible interactions of the patient's covariates. In order to start with a simple model, we decide to include in $\eta_{ij}$ a linear term over time and all the possible interactions over time and `Sex`, `Age`, `BMI`, `ADL`, `AB` and `TAU`.

The second step consists in specifying the working correlation structure: as known from the theory, this working assumption is allowed to be incorrect, since both the point estimates and the standard errors remain asymptotically valid. We decide to begin with an unstructured working correlation to be as general as possible.

After fitting the model with the `GENMOD` procedure in SAS, we can have a look at the predicted outcome in Figure 4. Clearly from this plot we notice that there is still a lot of variability to be captured. In order to improve our model, we decide to change the structure of $\eta_{ij}$. More in detail, we try to use higher order terms over the temporal variable.

We start by introducing a quadratic model in $\eta_{ij}$: in particular, we take into consideration the interaction of `Time` and $\texttt{Time}^2$ with the other covariates available in the dataset. We fit again the model using the `GENMOD` procedure in SAS with an unstructured working correlation structure. In order to compare the linear and the quadratic model, we apply a procedure based on the QIC criterion (Quasi-Likelihood Information Criterion): since the QIC value of the second model is significantly lower than the previous one, we decide to keep the quadratic approach. An informal point towards our decision is given in Figure 5, where we see that the predicted average values in this second model are more towards the measured ones.

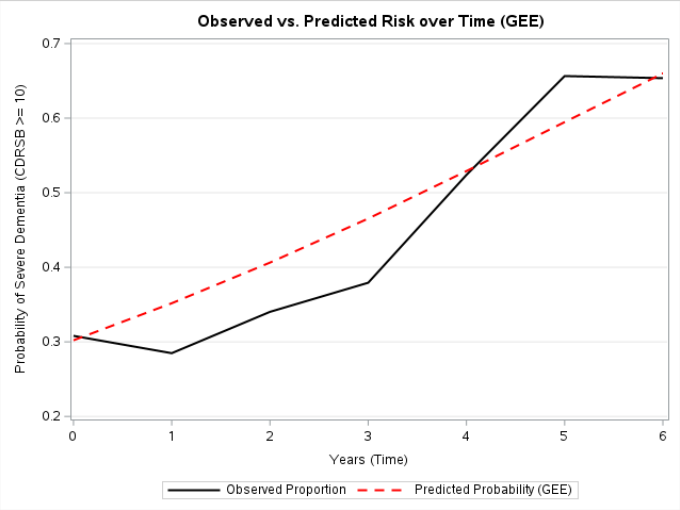


Figure 4: Observed vs. Predicted risk of severe dementia over 6 years for the Full Linear GEE Model. The linear GEE model predicts a constant, gradual increase, which does not match the rapid non-linear decline seen in the actual patient data.

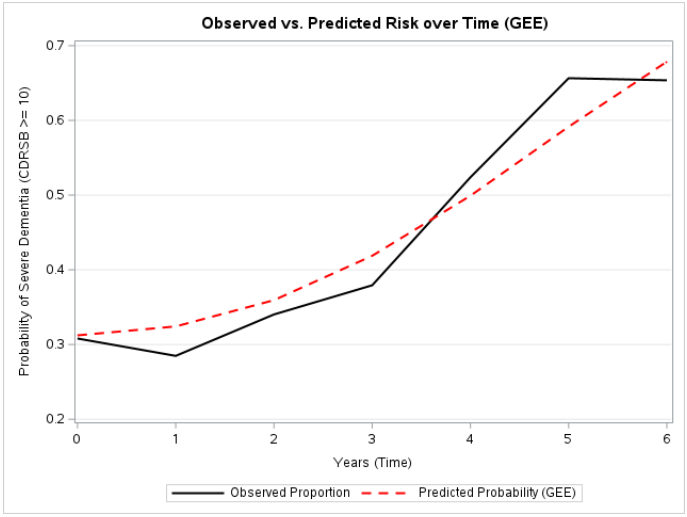


Figure 5: Observed vs. Predicted risk of severe dementia over 6 years for the Full Quadratic GEE Model.

We can also explore a cubic (over time) GEE model: by looking at the QIC value, we found a small improvement with respect to the quadratic one. However, higher-order polynomials mainly improve predictive accuracy by flexibly capturing local fluctuations in the data. Since our goal is not to produce highly accurate forecasts but rather to describe the general trajectory of Alzheimer's disease progression, interpretability becomes more important than marginal gains in fit. For this reason, we choose to keep the quadratic GEE model, which provides a more clear and meaningful representation of the underlying trend over time.

At this point we go back to the quadratic GEE model, and we study whether our assumption for the working correlation matrix is reasonable or not. By comparing the empirically corrected and the model-based standard errors of this quadratic GEE model, they are showed to be very similar, suggesting that the chosen working correlation structure is consistent with the true underlying structure. Moreover, additional models with alternative working correlation structures (AR(1), exchangeable, and independence) were tested as a sensitivity analysis for the regression parameters. These models indicate that the results are robust and do not depend on an arbitrary choice of correlation structure, as the estimates change very little across the different specifications.

Finally, after choosing our correlation structure, we try to reduce the mean structure in order to get to a parsimonious model. As stated before, we started with a full quadratic model containing all main effects, their interactions with time (Slope) and also their interactions with time$^2$. To proceed we apply a manual backward elimination procedure based on an Information Criteria approach. In each step, we remove the variables with the highest Wald $p$-values (indicatively the ones significantly greater than the significance level $\alpha = 0.05$) and we keep the reduced model if this leads to a decrease in the QIC value. We strictly adhere to the *hierarchy principle*: a main effect (Intercept) cannot be removed if its corresponding interaction with Time (Slope) remains significant.

The final marginal model is specified in (2), where $Y_{ij}$ denotes the binary outcome for patient $i$ at time $j$, $\pi_{ij} = \mathbb{P}(Y_{ij} = 1)$, and the parameters estimates are reported in Table 4. Note that the variables $AGE$ and $BMI$ represent the standardized versions of Age and BMI, respectively, to facilitate comparison of the estimates.

$$\begin{cases} Y_{ij} \sim \text{Bernoulli}(\pi_{ij}) \\ \text{Logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \eta_{ij} \\ \eta_{ij} = \beta_0 + \beta_{0,S} Sex_i + \beta_{0,B} BMI_i + \beta_{0,AD} ADL_{ij} \\ \qquad +(\beta_1 + \beta_{1,B} BMI_i + \beta_{1,AD} ADL_{ij}) \cdot Year_{ij} \\ \qquad +(\beta_2 + \beta_{2,AD} ADL_{ij}) \cdot Year_{ij}^2 \end{cases} \tag{2}$$

| **Parameter** | **Estimate** | **Std Error** | **P-value** |
|---|---|---|---|
| Intercept | -0.8983 | 0.1441 | <.0001 |
| TIME | -0.1051 | 0.1049 | 0.3164 |
| TIME × TIME | 0.0779 | 0.0179 | <.0001 |
| Sex (0) | 0.0542 | 0.0454 | 0.2322 |
| BMI | 0.1093 | 0.0600 | 0.0684 |
| ADL | 0.0114 | 0.0188 | 0.5434 |
| TIME × BMI | -0.0412 | 0.0206 | 0.0457 |
| TIME × ADL | 0.0167 | 0.0129 | 0.1972 |
| TIME × TIME × ADL | -0.0047 | 0.0020 | 0.0178 |

Table 4: Parameter estimates, Robust Standard Errors and P-Values for the reduced quadratic GEE model.

The final reduced model retains `Time`, `Time*Time`, `Sex`, `BMI` and `Adl`. Specifically, we found significant interactions between `Time` and `BMI`, as well as `Time` and `adl` and `Time*Time` and `adl`. Note that `Abpet` and `Taupet` were eliminated from the model. Our initial model failed to recognize the importance of daily functioning (`adl`) because it assumed patients decline at a steady, straight-line pace. In reality, patients with poor function do not decline steadily; they accelerate into severity, like falling off a cliff. Once we updated the model to capture this 'curved' acceleration, `adl` immediately became the strongest predictor because it effectively identifies who is about to crash. Consequently, the biomarker `Taupet` (which was significant in the simpler model) lost its significance;

it was merely acting as a rough placeholder until the model could correctly identify the rapid functional decline driven by `adl`.

We also discussed the potential need for some extensions of the GEE model, such as Alternating Logistic Regression (ALR) or Linearization Based Method (LBM). However, since there is no scientific interest in the association structure, we were satisfied with the classic one.

### 3.2 Random Effects Model

The classical step after a marginal model in a longitudinal context consists in studying whether the introduction of random effects can improve the statistical ability of capturing the real structure behind the data. In particular, we will consider a model in the form

$$\begin{cases} Y_{ij} \sim Bernoulli(\pi_{ij}) \\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \eta_{ij} \\ \eta_{ij} = f(x_{ij}, b) \\ b \sim \mathcal{N}(0, D) \end{cases}$$

where $\pi_{ij} = \mathbb{E}[Y_{ij}] = \mathbb{P}(Y_{ij} = 1)$, $f(x_{ij}, b)$ is a function, usually taken linear in the parameters, of the covariates in our dataset and of the random effects $b$. This latter quantity models the between-subject variability. Notice that different choices can be made on the link function between the mean $\pi_{ij}$ and the predictor $\eta_{ij}$ (*e.g.* probit, logistic ...), but here we decide to stick to the classical choice of a logistic regression due to the simple interpretability of the coefficients.

In order to proceed, we need to make a series of modeling choices. The first one regards the form of the linear predictor $\eta_{ij}$. We can look at different choices, but for simplicity we decide to start from a linear function over all the possible first order interactions over time and the other covariates in our dataset, *i.e.*,

$$\begin{aligned} \eta_{ij} = {} & \beta_0 + \beta_{0,A} Age_i + \beta_{0,B} BMI_i + \beta_{0,S} Sex_i + \beta_{0,AD} ADL_i \\ & + \beta_{0,AB} AB_{ij} + \beta_{0,\tau} TAU_{ij} + b \\ & + (\beta_1 + \beta_{1,A} Age_i + \beta_{1,B} BMI_i + \beta_{1,S} Sex_i + \beta_{1,AD} ADL_i \\ & + \beta_{1,AB} AB_{ij} + \beta_{1,\tau} TAU_{ij}) \cdot Year_{ij} \end{aligned}$$

Later on we will examine whether it is appropriate or not to extend the dependence of $\eta_{ij}$ to higher temporal terms.

The second choice we are left with consists into finding an appropriate structure for the random effects $b$: we decide to start by including random effects both in the intercept and over time to be as general as possible. As a matter of fact, as we have seen in the exploratory data analysis, the empirical variance of the `CDRSB` quantity does not exactly behave like in a classical Binomial model. In addition, we keep the matrix $D$ unstructured, as it is usually done in this situation.

Given these premises, we can start fitting our model by using the `GLIMMIX` procedure available in SAS. As known from the theory behind the Generalized Linear Mixed-Effects

models, we need to select a proper way to approximate the likelihood: in Table 5 we report the estimates from PQL, Laplace and Adaptive Gaussian quadrature with 3, 10 and 30 points. From theoretical results we know that the last approach is the most appropriate one: we decide to use 10 quadrature points in the following of our analysis to improve the computational cost of our model.

Table 5: Coefficient Estimates and Standard Errors from the `GLIMMIX` procedure and different approximation methods. Here "Gauss $n$" stands for the Adaptive Gaussian Quadrature approximation with $n$ interpolation points.

| **Parameter** | **PQL** | **Laplace** | **Gauss 3** | **Gauss 10** | **Gauss 30** |
|---|---|---|---|---|---|
| Intercept | -1.5923<br>(0.2932) | -3.2727<br>(0.5087) | -2.8927<br>(0.4967) | -3.0014<br>(0.5246) | -3.0122<br>(0.5256) |
| TIME | 0.3542<br>(0.1060) | 0.8108<br>(0.1786) | 0.7028<br>(0.1761) | 0.7294<br>(0.1855) | 0.7323<br>(0.1859) |
| AGE_STD | 0.2978<br>(0.1708) | 0.2170<br>(0.2676) | 0.3227<br>(0.2683) | 0.3234<br>(0.2819) | 0.3213<br>(0.2824) |
| BMI_STD | 0.1483<br>(0.09151) | 0.1998<br>(0.1547) | 0.2162<br>(0.1524) | 0.2265<br>(0.1621) | 0.2270<br>(0.1626) |
| sex | 0.2276<br>(0.1905) | 0.3930<br>(0.3226) | 0.3724<br>(0.3174) | 0.3933<br>(0.3373) | 0.3953<br>(0.3381) |
| adl | 0.01157<br>(0.04117) | 0.06259<br>(0.06909) | 0.04157<br>(0.06803) | 0.04294<br>(0.07201) | 0.04338<br>(0.07214) |
| ABPET_STD | -0.1151<br>(0.1159) | -0.1325<br>(0.1679) | -0.1346<br>(0.1717) | -0.1346<br>(0.1776) | -0.1329<br>(0.1777) |
| TAUPET_STD | -0.1134<br>(0.09132) | -0.1090<br>(0.139) | -0.1142<br>(0.1210) | -0.1092<br>(0.1251) | -0.1084<br>(0.1256) |
| TIME_BMI | -0.05616<br>(0.02998) | -0.07371<br>(0.04935) | -0.07934<br>(0.04907) | -0.08244<br>(0.05196) | -0.08257<br>(0.05209) |
| TIME_TAU | 0.04478<br>(0.02339) | 0.04596<br>(0.03017) | 0.04852<br>(0.03120) | 0.04727<br>(0.03212) | 0.04706<br>(0.03222) |
| TIME_SEX | 0.05360<br>(0.06235) | 0.1035<br>(0.1028) | 0.09699<br>(0.1020) | 0.1034<br>(0.1080) | 0.1041<br>(0.1082) |
| TIME_AGE | -0.1204<br>(0.05465) | -0.08965<br>(0.08408) | -0.1245<br>(0.08491) | -0.1239<br>(0.08889) | -0.1232<br>(0.08904) |
| TIME_AB | 0.03522<br>(0.03647) | 0.03484<br>(0.05118) | 0.03720<br>(0.05287) | 0.03602<br>(0.05437) | 0.03632<br>(0.05437) |
| TIME_ADL | -0.00462<br>(0.01307) | -0.02062<br>(0.02170) | -0.01450<br>(0.02153) | -0.01496<br>(0.02272) | -0.01511<br>(0.02276) |

In Figure 6 we see a comparison between the average of the predicted outcome 1 over time with respect to the observed one: it seems that our model is able to capture the behavior of this quantity. However we would like to test from a more formal point of view

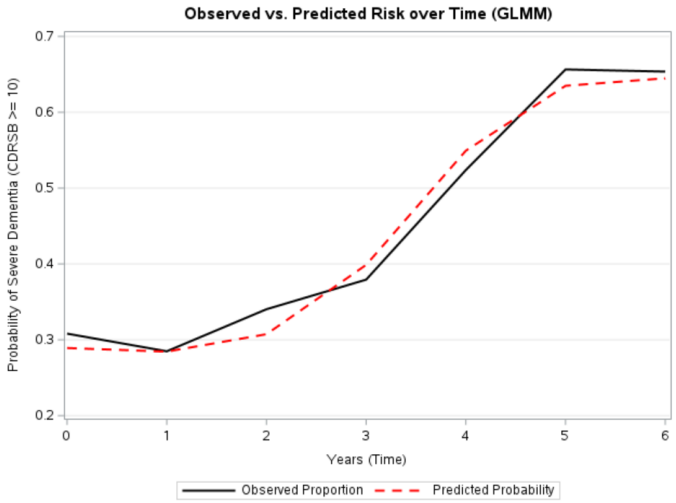


Figure 6: Predicted proportion of outcome 1 in the `CDRSB` value in the (complete) GLM model with random intercept and slope. The predictions were done using the `GLIMMIX` procedure in SAS.

our modeling decisions. First of all, we test whether the adding of higher order terms over time is needed. To do that, we fit a quadratic model over time for $\eta_{ij}$ and we compare the BIC values of the quadratic and the linear model. As we get that the BIC value in the quadratic approach is significantly higher than the linear one, we decide to stick to our previous assumption.

Next, we test whether the random effects structure can be reduced: as we know that there are no explicit tests available in the literature, we apply again an Information Criterion approach. In detail, we compare the BIC values of our fitted model with a model with only random intercepts: we get a significant increase in the BIC value of the second model. So we consider again a random component over time.

At this point, after having discussed the temporal form of $\eta_{ij}$ and the random effects $b$, we turn our attention to simplifying the structure of the linear predictor $\eta_{ij}$. As a matter of fact, many of the estimates are not significant and we want to get to a model that is as parsimonious as possible. We apply a manual backward selection procedure based on the BIC value. More in detail, at each step we remove the variables whose detected $p$-value from the Wald test is (significantly) higher than 0.5, and we keep the reduced model only if it leads to a decrease in the BIC value. As done in the marginal model, we stick to the hierarchical principle described above.

The final reduced model retains `Time`, `BMI`, `TAU`, `Time*BMI` and `Time*TAU`. The estimates and their standard errors can be retrieved in Table 6. We can briefly analyze the results we got, starting from the coefficients in the $\eta$ quantity. The first thing that captures our attention is that, except for the intercept and the `TIME` slope, the estimates related to `BMI` and `TAU` are clearly not significant (their $p$-values span from 0.13 to 0.19, but we kept them because of the BIC criterion we explained before). Apparently, when we introduce the random effects over time, these are sufficient to explain the differences

| **Parameter** | **Estimate** | **Std Error** | $p$**-value** |
|---|---|---|---|
| Intercept | -2.5656 | 0.1789 | <.0001 |
| TIME | 1.6285 | 0.6197 | 0.0087 |
| BMI_STD | 0.2120 | 0.1610 | 0.1879 |
| TAU_STD | -0.1218 | 0.1199 | 0.3096 |
| TIME*BMI_STD | -0.03573 | 0.02398 | 0.1363 |
| TIME*TAU_STD | 0.04353 | 0.03961 | 0.1551 |

Table 6: Estimated values of the parameters of the reduced random effects model. The estimates were obtained using the `GLIMMIX` procedure in SAS and an Adaptive Gaussian Quadrature approach with 10 interpolating points.

between the evolution of the `CDRSB` value in different patients. So it is not necessary (and it seems also somehow wasteful) to correct for the information that we have over patients. In particular, this result means that the level of daily activities at the baseline, and the Amyloid-Beta and Tau protein quantities do not play a role into capturing the evolution of the `CDRSB` when we introduce a random effect structure.

We can also study more in detail the results for the random effects in Table 7. We observe that they present a marked volatility: as a matter of fact, the estimated variance of the random intercept is around 21, while the estimated variance of the random effects over time is around 2. This means that the subject-specific bheaviour of the `CDRSB` quantity can be significantly different from the mean profile. In addition, it is interesting to notice that there is an high correlation in the random intercepts and slopes: as a matter of fact, we can compute the correlation from our output, which is around -1.

| **Parameter** | **Estimate** | **Std Error** |
|---|---|---|
| Rnd Intercept (Variance) | 21.1667 | 2.0998 |
| Rnd Effect (Time) | 2.1287 | 0.2078 |
| Covariance | -6.6450 | 0.6475 |

Table 7: Estimated values of the variances and the covariance in the random effects. The estimates were obtained using the `GLIMMIX` procedure in SAS and an Adaptive Gaussian Quadrature approach with 10 interpolating points.

### 3.2.1 Empirical Bayes Estimates

Finally, it is useful to examine the Empirical Bayes estimates of the random effects. The histograms in Figure 7 show that the estimated intercepts and slopes do not follow a clear normal shape and even appear somewhat multimodal. This is not unexpected: EB estimates are conditional quantities that combine the prior with the information available for each individual, and thus they are not a direct sample from the assumed distribution of the random effects. Their empirical distribution can therefore deviate from normality.

In this case, the observed multimodality is likely related to a limited explanatory power of the fixed effects. Some important covariates may be missing, which would otherwise account for differences in baseline levels or trajectories. As a result, the model attributes this unexplained structure to the random effects, leading the EB estimates to cluster into the groups visible in the histograms.

A last quick observation can be made by looking at the scatterplot of the random effects in Figure 8: this shows a clear negative association between the estimated intercepts and slopes, which is coeherent with what we derived from our model.

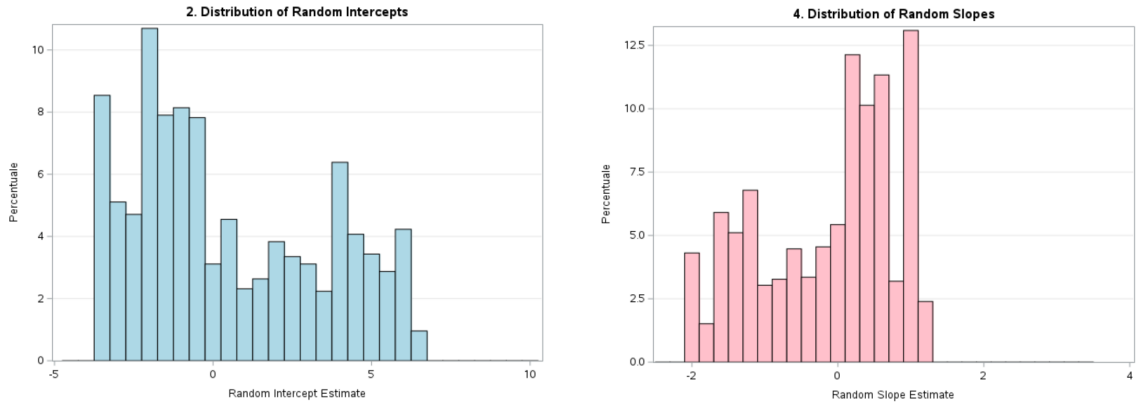


Figure 7: Histograms of random intercepts (left) and random slopes (right).

# 4 Conclusion and further studies

At the end of our analysis we can conclude that for the GEE model, which focuses on the population average, we observe a non-significant but close effects for `BMI` at baseline and a significant effect of `BMI` over time, but not much else, since `TIME` and `BMI` over time explain most of the variability. An additional noteworthy finding is that the variable `ADL`, which measures the number of daily activities performed by patients at baseline, appears to influence the evolution of the probability of `CDRSB` $= 1$. This suggests that daily activities may play a role in cognitive decline: while the initial level of `ADL` does not seem to affect the baseline probability of being in a worse cognitive condition, it becomes relevant throughout the disease progression. In particular, ADL acts as a critical modulator of disease trajectory, governing the acceleration of clinical decline. Specifically, lower functional independence predicts a precipitous, non-linear deterioration into severe dementia.

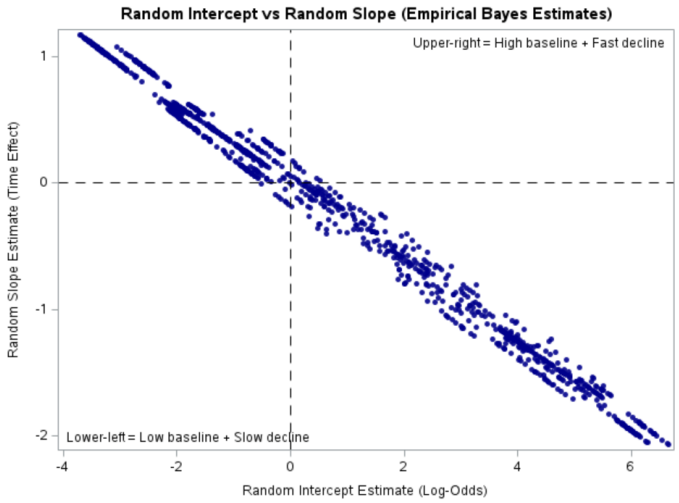


Figure 8: Scatterplot of the Empirical Bayes estimates of the random intercepts and slopes in the final reduced GLM model. For any patient, the random intercept is on the horizontal axis and the random slope on the vertical one.

In contrast with the marginal model, the random-effects mixed model revealed that subject-specific heterogeneity was the dominant feature of the data. By explicitly modeling individual trajectories, the temporal dynamics (`TIME`) and the subject-specific random effects absorbed the vast majority of the outcome variability. Consequently, the predictive power of static covariates was overshadowed by the fact that every patient declines in their own unique way, leaving little residual variance for them to explain. Regarding `Taupet` and `Abpet`, both variables show little variability across subjects. This limited variation makes it difficult for them to have a measurable impact on the outcome in our models, which may explain why they do not appear as significant predictors. Essentially, their lack of statistical significance is likely due to the fact that their contribution is covered by the stronger effects of `TIME` and the subject-specific random effects.

Beyond the biological interpretation of our findings, it is also valuable to compare the performance of the statistical models applied to our dataset. By predicting the logistic curve for each patient and then averaging across individuals, we can study the overall temporal evolution. We observe that the random effects model provides a closer fit to the average fractions than the marginal model. This result is consistent with expectations in longitudinal analyses, where random effects often offer a more flexible and efficient framework for capturing subject-specific variability.

For future studies, a valuable approach would be to explore the use of spline models, which could capture non-linear trends over time with greater flexibility than polynomial terms. Additionally, experimenting with alternative measurement scales for the protein markers might reveal subtle effects that were not detectable with the current approach.

## Acknowledgements

AI was used throughout this project to enhance productivity and accuracy. It helped refine and review text, assisted with R ans SAS programming by assisting with data analysis, debugging, and creating complex visualizations, and generated well-formatted LaTeX tables from statistical results. Human oversight remained key, with students reviewing all outputs to ensure scientific accuracy, making AI a productivity booster rather than an independent agent.

# 5 References

## References

[1] Michiel Bertsch, Bruno Franchi, Maria Carla Tesi, and Veronica Tora. "The role of A$\beta$ and Tau proteins in Alzheimer's disease: a mathematical model on graphs". In: *Journal of Mathematical Biology* 87.49 (2023). DOI: 10.1007/s00285-023-01985-7. URL: https://link.springer.com/article/10.1007/s00285-023-01985-7.

[2] Capelli, Carfagno, Garrofé, and Wood. *Project 2 LDA - Code*. https://github.com/dcapelli02/Project-2-LDA. Accessed: 20 November 2025. 2025.