Project 1 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

November 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 Brief Psychiatric Rating Scale (BPRS) as a proxy measure of disease severity. The BPRS is a widely used clinical instrument that evaluates psychiatric symptoms such as anxiety, depression, and emotional withdrawal, providing a standardized way to quantify patients' psychological condition. 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 BPRS, the dataset includes several categorical and numerical variables that provide further insight into patient characteristics and disease progression.

Our approach begins with an exploration of the data structure to identify patterns that may inform model selection and underlying assumptions. Building on these findings, we develop and compare three statistical frameworks: a multivariate model, a two-stage analysis, and a 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 Scale (`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).

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.

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 significant and steady patient dropout over time. As shown in Table 3, the number of available BPRS observations declines from 1,253 at baseline to 511 by year 6, resulting in a highly unbalanced data structure.

This attrition appears to be non-random; a t-test revealed a highly significant difference ($t = 27.137$, p-value $< 2.2e-16$) between the baseline BPRS scores of patients who completed the study (mean = 64.43) and those who dropped out (mean = 82.99). This suggests that patients with higher (worse) initial psychiatric symptoms were more likely to leave the study, a crucial factor to consider during modeling.

For the purposes of visualization and modeling with longitudinal packages (like `lme` in R), the data was converted from its original wide format (one row per patient) to a long format (one row per patient-time observation).

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%

Table 3: Patient Dropout: Non-Missing BPRS Observations

Time Point	Number of Patients
bprs0	1253
bprs1	1106
bprs2	1014
bprs3	907
bprs4	777
bprs5	652
bprs6	511

2.2 Exploratory Analysis

To determine the appropriate modeling strategy, we first examine the longitudinal trajectory of our primary outcome, the Brief Psychiatric Rating Scale (BPRS).

The first step in this process is to visualize some of the raw individual trajectories, as shown in the so called *spaghetti plot* in Figure 1. We notice a significant between-patient variability, as patients clearly differ in both baseline levels (intercepts) and rates of change (slopes).

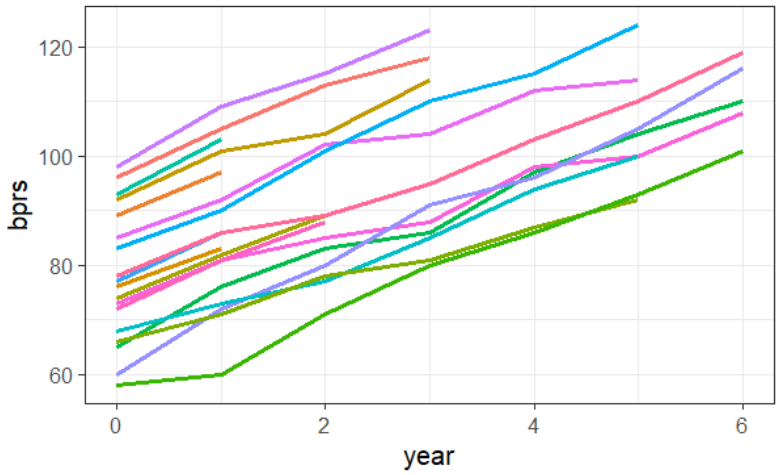


Figure 1: Spaghetti plot showing individual patient BPRS trajectories over the 7-year study period.

The next step consists in examining the mean structure to understand the cohort's average behavior. In Figure 2 we can see how the average `BPRS` score evolves over time: we can safely state that this quantity follows a linear trend over the years.

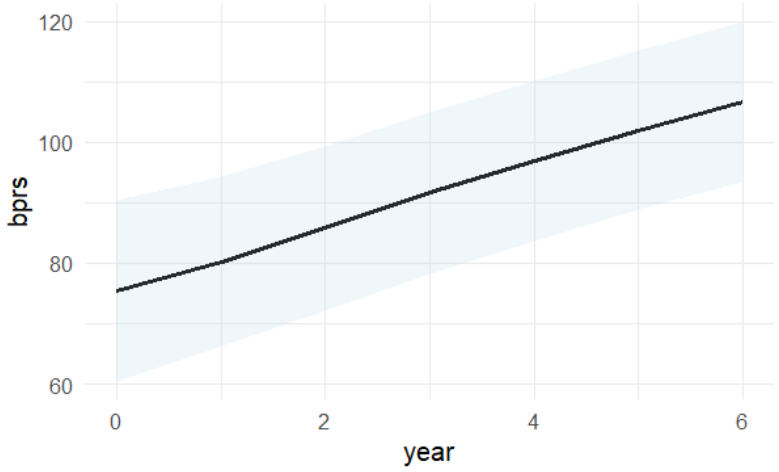


Figure 2: Average BPRS trajectory for the entire cohort from baseline to Year 6. The solid line illustrates the increasing mean trend in symptoms, while the shaded region indicates the 95 percent confidence interval around the mean.

Then we can also have a look at how the average BPRS value behaves in different groups to have an initial guess of what information will play a role into distinguishing different baseline levels and temporal slopes. In detail, it seems that variables such as `Age` and `ADL` will influence both these quantities, while variables such as `Trial`, `Job`, `BMI`, `WZC` and `CDRSB` will influence the baseline value.

Beyond the mean structure, we investigate also the variance structure. To do this, we can have a look at the plot of the variance of the `BPRS` scores over time, as shown in Figure 3.

This plot clearly suggests that the variance is not constant across the years, a condition known as heteroscedasticity. The variance is highest at baseline (Year 0) and decreases over time. This finding indicates that a simple covariance structure assuming constant variance (homoscedasticity) would be inappropriate for this data (when ignoring the longitudinal framework of our data). Imposing a constant variance assumption would force the model to estimate a single, average variance for all time points. This would result in a poor fit, as the model would significantly underestimate the variability at baseline while overestimating the variability at the later follow-up years, ultimately leading to biased standard errors.

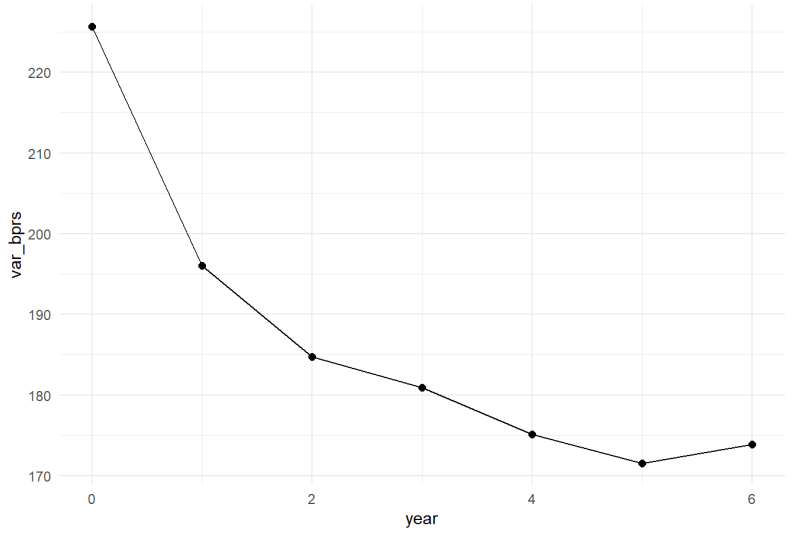


Figure 3: Change in BPRS variance over the 7-year study period. The non-constant trend, with variance decreasing from baseline, indicates the presence of heteroscedasticity.

Finally, we assess the correlation structure within patients. In Figure 4 we visualize the empirical correlation matrix for the `BPRS` measurements as a heatmap. Correlations are highest for adjacent time points (e.g., `bprs1` and `bprs2`) and they systematically decrease as the time lag between measurements increases (e.g., `bprs1` and `bprs6`). This pattern is a strong indicator of serial correlation, meaning an observation from a patient at one time point is predictive of their observation at the next.

2.3 Implications for Modeling

Taken together, these exploratory findings provide a clear directive for our modeling strategy: a simple model would fail to capture the complexity of our data. Specifically, any

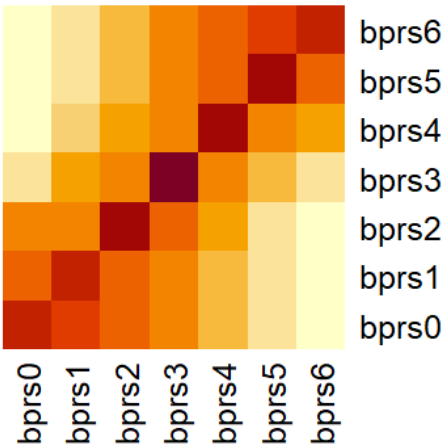


Figure 4: Correlation matrix (heatmap) for BPRS measurements from Year 0 to Year 6. Darker colors (reds) indicate stronger positive correlations, while lighter colors (yellows) indicate weaker correlations.

statistical model applied to our data must adequately account for the following three key features.

1. Informative Dropout. Patients with higher baseline scores are more likely to be missing at later time points. However, in our analysis we will not take into consideration this critical feature in an explicit way.

2. Heterogeneous Variances. The variance changes over time.

3. Serial Correlation. It is the correlation between serial measurements, which is usually a decreasing function of the time separation between these measurements.

These findings strongly suggest that a mixed-effects model combining a decaying correlation structure (like Autoregressive order 1, or AR(1)) with a heterogeneous variance model (such as `weights = varIdent()`) will be a more appropriate and parsimonious starting point than simpler alternatives.

2.4 Summary Statistics

Before fitting a complex longitudinal model, we applied a Summary Statistics Approach. This two-step method simplifies the longitudinal problem by first summarizing each subject's repeated BPRS measurements into a single value. In the second step, this single statistic is analyzed as the outcome in a classical linear model using the baseline covariates as predictors. This reduces the longitudinal data to a simple, cross-sectional analysis.

Based on the theory, we computed several summary statistics for each patient:

- Endpoints (y_{in_i}). The last available BPRS measurement.
- Increments ($y_{in_i} - y_{i1}$). The total change in BPRS from the first to the last measurement.

- Normalized AUC values (`nAUC`). The Area Under the Curve (calculated via the trapezoidal rule) normalized by the number of observed time points (`tmax`) to account for differing follow-up durations.
- Rates of Increment (`rate`). The total increment normalized by the number of observed time points (`tmax`).
- ANCOVA. A model for the endpoint (y_{in_i}) that includes the baseline BPRS (y_{i1}) as a predictor.

We then fit a linear model (`lm`) for each statistic. The results for the Normalized AUC (`nAUC`) and the Rate of Increment (`rate`) are, as expected, very similar. Both models show that trial, education level, age, income, job status, ADL, WZC, baseline amyloid (`abpet_base`) and baseline CDRSB (`cdrsb_base`) are highly significant predictors.

The models for Endpoints (`yini`) and total Increments also highlight a consistent set of predictors, including education level, income, job status, ADL, WZC, `abpet_base`, and `cdrsb_base`. The Endpoint model is also significantly associated with age, while the Increment model was not. The ANCOVA model, which corrects for the baseline BPRS (`y0`), confirms the significance of most of these same predictors.

2.5 Conclusions

This preliminary exploration has revealed several critical features of our data. We identified a significant informative dropout pattern, where patients with higher (worse) baseline BPRS scores were more likely to leave the study. This non-random missingness means simple models could be biased. Furthermore, our analysis uncovered a complex error structure: the data exhibits heteroscedasticity, with variance decreasing over time, as well as serial correlation, where measurements closer in time are more strongly related.

Our analysis using a summary statistics approach confirmed that regardless of the specific summary used, baseline covariates like trial, age, job status, ADL, WZC, and `cdrsb0` are robustly associated with the overall BPRS trajectory. While this method is straightforward and provides a good initial overview, its main drawback is that it only uses partial information, losing the detailed temporal evolution which we will capture with the multivariate and mixed-effects models.

Therefore, these findings (informative dropout, complex variance, and robust baseline predictors) collectively build a strong case for the advanced models that follow. A mixed-effects model is necessary to properly account for these complex longitudinal data structures and avoid the limitations of simpler methods.

3 Model Fitting and Explanation

In the previous section we studied the preliminary mean and covariance structures through simple methods, such as the exploratory data analysis or some summary statistics. The aim of this section is now to fit proper statistical models to our data in order to understand from a more rigorous point of view what drives the evolution of the BPRS value over time.

We begin by considering a simple multivariate model, which can provide useful insights even though it does not fully account for the longitudinal structure of our data. To

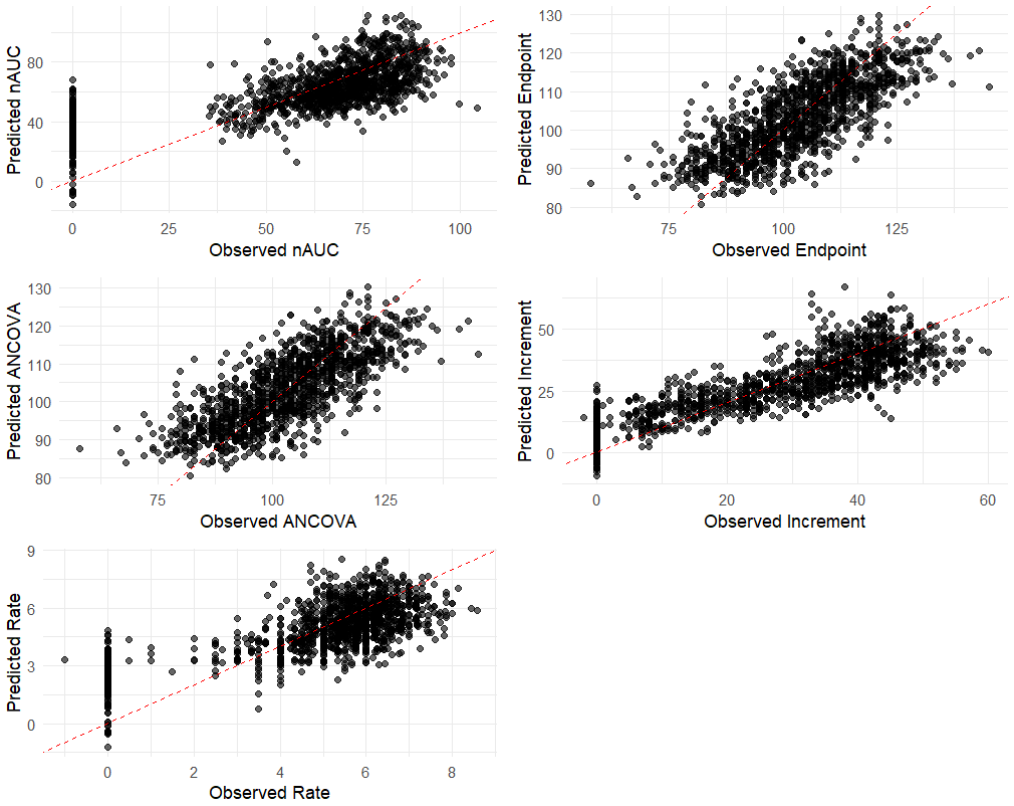


Figure 5: Predicted versus Observed Values for Summary Statistics Models. Each panel displays the predicted values from a linear model against the observed summary statistic for that patient. The dashed red line represents perfect prediction (Predicted = Observed). The clusters of points along the Y-axis at low observed values (particularly noticeable for `nAUC`, Increment, and Rate of Change) correspond to patients who only attended the first session, resulting in zero or near-zero values for these change-based summary statistics. These "zero-change" patients effectively add noise to the summary statistic models (e.g., for "Rate" or "Increment") and may bias the `lm` coefficients.

properly address this aspect, we introduce random effects into the model. Since our preliminary analysis suggests that a linear relationship with respect to time is reasonable, we will evaluate whether a linear mixed-effects model is an appropriate choice. As a first step, we conduct a two-stage analysis to gain an initial impression of the model's suitability. We then proceed to fit a linear mixed-effects model directly to the data using the functions available in R. Prior to model fitting, all relevant categorical variables are factorized as required.

3.1 Multivariate Model

We start our analysis by ignoring the random effects over time and by fitting a multivariate model to our data. This procedure, as the other procedures that we will see, requires us to specify proper mean and covariance structures.

Regarding the first choice, we would like to start by considering a model as generic as possible. From the exploratory data analysis performed in the previous section, it is clear that the BPRS value evolves linearly over time: so linearity over this variable is the first assumption that we can safely make. The next step is to study how to incorporate the additional information we have about our patients into the model. One possible approach would be to consider only the interactions identified during the exploratory data analysis. However, in order to be as rigorous as possible, we will allow both the temporal intercept and the temporal slope to depend on all the baseline information available for each patient. We will then assess which interactions prove to be significant at a later stage.

Now we need to determine the covariance structure for our model. Ideally, we would choose an unstructured covariance to avoid imposing assumptions and constraints. However, this choice of covariance structure combined with the mean structure (a full variable selection with some interaction terms) is too computationally complex. Therefore, a more constrained choice of heterogeneous AR(1) or Toeplitz seem to be appropriate here based on the exploratory analysis. Since observations become less correlated as their distance increases and variance decreases over time, a heterogeneous AR(1) structure is the most appropriate choice.

3.1.1 Multivariate Model Reduction

The model we built is not parsimonious at all; it considers numerous interactions between the time-effects and the other information available. So we would like to simplify it in order to get to a final model that is both parsimonious and well-performing.

We start reducing the mean structure by applying an automatic stepwise procedure based on the BIC value, which penalizes in an higher way complex models with respect to simple ones, trying both forward selection and backward elimination. As we get two nested models from these two procedures, we can use a likelihood-ratio test to decide which one to keep: the result leads us to pick the backwards one. We therefore achieve the following structure:

$$\begin{aligned} BPRS_{i,j} = {} & \beta_0 + \beta_{0,T} Trial_i + \beta_{0,A} Age_i + \beta_{0,B} BMI_i + \beta_{0,J} Job_i \\ & + \beta_{0,AD} ADL_i + \beta_{0,W} WZC_i + \beta_{0,C} CDRSB_i \\ & + (\beta_1 + \beta_{1,A} Age_i + \beta_{1,AD} ADL_i + \beta_{1,C} CDRSB_i) \cdot Year_{i,j} + \varepsilon_{i,j} \end{aligned}$$

where the full parameter estimates can be retrieved in Appendix A. Table 4 reports all parameter estimates and corresponding p-values, excluding `trial` for brevity.

Parameter	Estimate	p-value	Parameter	Estimate	p-value
(Intercept)	-46.34068	0.0000	job1	-3.89547	0.0005
age	1.63207	0.0000	year	5.59909	0.0000
bmi	0.14457	0.0001	age:year	0.01673	0.0040
adl_num	-0.12573	0.4401	adl_num:year	0.05765	0.0051
wzc1	1.89856	0.0000	cdrsb_base:year	-0.02276	0.0000
cdrsb_base	-0.02075	0.0123	job1:year	-0.42103	0.0077

Table 4: Estimated coefficients of multivariate model (non-trial terms)

From this, we see that the variables with the most positive relative influence on the outcome `BPRS` are `age`, `year`, and `wzcl`. For each year of age, the `BPRS` score increases by 1.632 (approximated), and for `wzcl`, each point results in an increase of 1.899 to the `BPRS`. `year` has a very significant, positive effect on the outcome with each additional year into the trial corresponding to an approximately 5.599 increase in `BPRS`. `ADL` alone is not significant; however, because the interaction effect between it and `year` is, it is included in the reduced model.

Variables that have a negative impact on the outcome are the baseline values of `CDRSB`, `job` (specifically not having one), and the interaction terms between `year` with `CDRSB` and `job`, respectively. Although not included in the above table, the `trial` parameter estimates varied in effect direction (positive and negative) and magnitude, with the highest positive and negative effect seen in trials 6 and 25, respectively. Relative to trial group 1, patients in these groups had a positive `BPRS` differential of 5.652 or a negative differential of 5.155. It is unclear if there is a geographic factor among trial groups, as this is currently unaccounted for in the given dataset.

Finally, we check if we can simplify also the covariance structure: we compare our model to other models fitted with the same (reduced) mean structure but different covariance structures, such as heterogeneous exponential or heteroskedastic covariance. After applying the proper statistical tests (LRT for nested structures, and BIC for non-nested), we conclude that the heterogeneous AR(1) choice is still the best one, and so we do not need to simplify that.

3.2 Two-Stage Analysis

The multivariate analysis that we performed in the previous section is a good starting point in the analysis of the temporal evolution of the BPRS value. However, this approach does not take into consideration the longitudinal framework of our dataset in a proper way. A well-known way to proceed to take into consideration within-patient variability is given by the linear mixed effects model: in this first section we will see an explicit two-stage way

to construct it, while in the last part of this document we will see how to build it directly from the data without passing through the intermediate stage.

3.2.1 Two-Stage Analysis: First Stage

The first step of our two-stage analysis consists in choosing a proper model that can be fitted separately for each patient to describe the data. The exploratory data analysis performed in the previous section comes in handy here: as we have seen, it clearly suggests to start with a linear model over time of the form

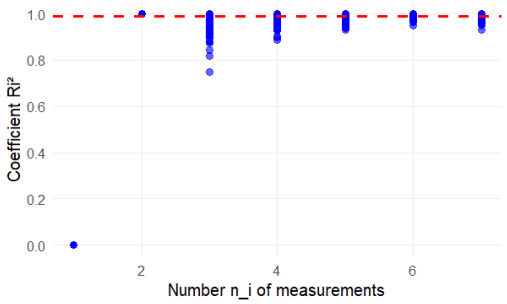
$$\begin{cases} BPRS_{i,j} = \beta_{0,i} + \beta_{1,i} Year_{i,j} + \varepsilon_{i,j} \\ \varepsilon_i \sim \mathcal{N}(0, \Sigma_i) \end{cases} \tag{1}$$

In order to proceed we need to specify a proper variance-covariance matrix for the individual errors ε_i: as a first guess, it is reasonable to stick to the classical assumption of homoskedasticity, *i.e.*, $\Sigma_i = \sigma_i^2 I_{n_i}$. We can briefly analyze the performance of these individual linear models by looking at the R_i^2 scatterplot in Figure 6a. As we can see graphically, it seems that overall the linear model is able to capture the evolution over time of the BPRS value.

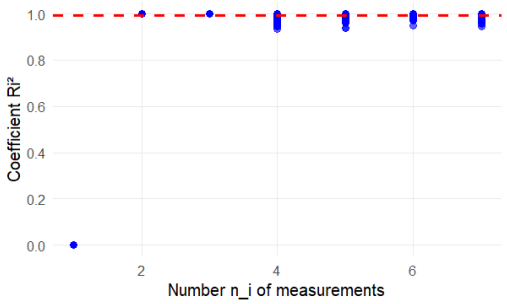
At this stage it is interesting to notice that we can justify formally why we do not choose a quadratic model over time. We start by fitting a quadratic model over time for each patient in the following form

$$\begin{cases} BPRS_{i,j} = \beta_{0,i} + \beta_{1,i} Year_{i,j} + \beta_{2,i} Year_{i,j}^2 + \varepsilon_{i,j} \\ \varepsilon_i \sim \mathcal{N}(0, \Sigma_i) \\ \Sigma_i = \sigma_i^2 I_{ni} \end{cases} \tag{2}$$

If we compare the two models (1) and (2) via an F-test, we get a p-value 1, which means that it is not statistically significant to take into consideration a quadratic term over time. So this is the reason why we use the linear model (1).



(a) Plot of the individual R^2 values of the first-stage linear model over time.



(b) Plot of the individual R^2 values of the first-stage quadratic model over time.

Figure 6: Comparison of the individual R^2 values for the first-stage linear (a) and quadratic (b) models over time.

3.2.2 Two-Stage Analysis: Second Stage

After picking an appropriate individual model for the evolution of the BPRS index, the second step of our two-stage analysis consists in studying how the patient-specific regression parameters might be related to the other information that we have through new "global" parameters. In formula, we aim to derive a model in the following form:

$$\begin{cases} \beta_i = \begin{pmatrix} Z_{i,0} & 0 \\ 0 & Z_{i,1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \\ \mathbf{b} \sim \mathcal{N}(\mathbf{0}, D) \end{cases} \tag{3}$$

where the vectors $Z_{i,0}$ and $Z_{i,1}$ contain the information on the patient i (*e.g.* Age, Education Level, Baseline Amyloid-Beta value...) that we decide are significant to explain the differences in the intercepts and temporal slopes between patients, respectively. In this case we keep the matrix D unstructured in order to capture the variability yet to examine in the individual models.

Following the idea we used in the multivariate model, we start by keeping the most general mean structure possible, *i.e.*, we include all the available baseline information on the patients in the vectors $Z_{i,0}$ and $Z_{i,1}$. Then we reduce the model by taking into consideration only the variables whose estimation is considered to be significant.

This procedure leads to the following reduced model:

$$\begin{cases} \beta_{0,i} & = \beta_0 + \beta_{0,T} Trial_i + \beta_{0,A} Age_i + \beta_{0,B} BMI_i \\ & \quad + \beta_{0,J} Job_i + \beta_{0,AD} ADL_i + \beta_{0,W} WZC_i + \beta_{0,C} CDRSB_i + b_0 \\ \beta_{1,i} & = \beta_1 + \beta_{1,A} Age_i + \beta_{1,J} Job_i + \beta_{1,AD} ADL_i \\ & \quad + \beta_{1,W} WZC_i + \beta_{1,C} CDRSB_i + b_1 \\ \mathbf{b} & \sim \mathcal{N}(\mathbf{0}, D) \end{cases} \tag{4}$$

where the full parameter estimates can be found in Appendix A. Table 5 reports a reduced set of the parameter estimates (again excluding `trial` estimates for brevity).

Older age, higher BMI and living in-residence as opposed to in-home correspond to higher `BPRS` values, while having a job and greater baseline `CDRSB` relate to lower values. The slope estimates show a similar pattern, albeit demonstrably weaker than the intercept estimates: Age and in-residence living predict slightly faster yearly increases, whereas having a job and higher baseline `CDRSB` correspond to slower progression.

3.3 Linear Mixed Effects Model

The final approach we study for our data is given by the Linear Mixed Effects Model, which combines the two stages mentioned earlier into a single model. It can be written as:

$$\begin{cases} Y_i = X_i \beta + Z_i b_i + \varepsilon_i \\ b_i \sim \mathcal{N}(0, D) \\ \varepsilon_i \sim \mathcal{N}(0, \Sigma_i) \\ b_1, \ldots, b_N, \varepsilon_1, \ldots, \varepsilon_N \text{ independent} \end{cases}$$

Variable	Coefficient
(Intercept)	-48.724^{***}
age	1.647^{***}
bmi	0.167^{***}
job1	-4.388^{***}
wzc1	1.741^{***}
cdrsb_base	-0.019^{***}

Variable	Coefficient
(Intercept)	1.257
age	0.073^{*}
job1	-1.218^{*}
wzc1	0.190^{*}
cdrsb_base	-0.020^{***}

Table 5: Fixed effects for intercept (left) and slope (right) in the twoo-stage analysis

This model represents a good approach to our problem because it extends the multivariate linear regression model to longitudinal settings. Specifically, it allows modeling the difference between subjects by incorporating random effects b_i, while simultaneously accounting for the correlation between measurements at different time points through the variance components Σ_i. Additionally, the matrix D captures the relationships among the random effects themselves, allowing for random intercepts, random slopes, or both.

The first step towards the construction of a linear mixed effects model for our data consists in studying an OLS regression: we will take into consideration a mean structure that is as general as possible, in order to avoid unwanted reductions. An interesting result is given by studying the residual structure, which can be seen in Figure 7.

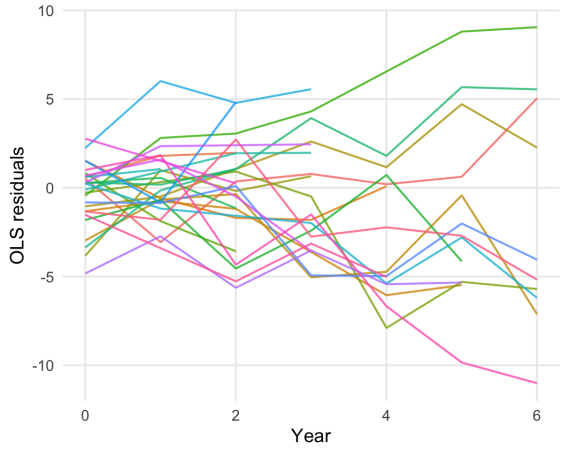


Figure 7: Trajectories of OLS residuals of 30 patients sampled randomly from the dataset. Each line represents a single patient.

We first decide to include a random intercept for each patient to take into account differences in the baseline value of the BPRS quantity. Then, Figure 7 suggests the need

of a random slope for each patient, as residuals profile do not have linear and constant evolution over time. To understand if our guess is right, we can compare the variance of the OLS residuals with the fitted variance that is predicted by the LME model: we get the results in Figure 8. This figure clearly shows that the previous assumptions were reasonable, since the lines are very close.

After the preliminary analysis on the possibility of introducing random effects, the serial correlation was also discussed, but empirical tests and theoretical discussions lead to ignore this component. As a matter of fact, the time points have a one-year interval, which means that it is not strictly necessary to account for serial correlation. In any case, we can test different serial correlation structures with various function, but none leads to significant improvements. Therefore, we keep the residual covariance structure as $\Sigma_i = \sigma^2 I_{n_i}$ since random effects explain most of the systematic variability in the data.

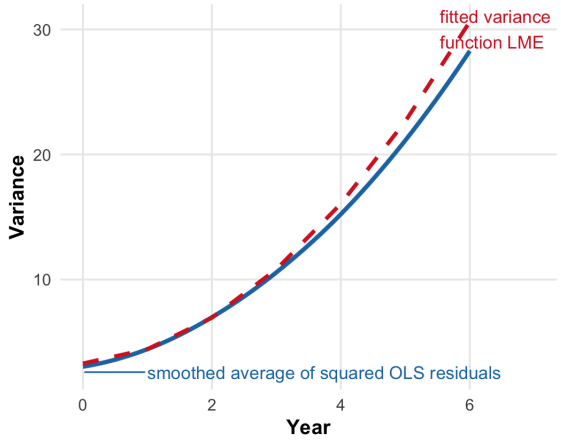


Figure 8: Comparison of smoothed average of squared OLS residuals with fitted variance function from the LME model with random slope and intercept.

At this stage of the analysis, we perform a likelihood ratio test on the random slope to compare a model with random intercept and slope to a model with only a random intercept. In this case, as we know from theoretical results, the asymptotic null distribution of the test is given by a mixture of χ^2_1 and χ^2_2 with equal weights (0.5). The result suggests that there are significant differences between patients in the evolution of BPRS over time, so the random slope is necessary (LRT: $\chi^2 = 2931.45$, $p < 0.001$). Moreover, examining the matrix D at this stage shows that the random intercept and slope are important, as both their variances are substantial ($\sigma^2_{\text{Intercept}} = 0.53$, $\sigma^2_{\text{Slope}} = 0.65$).

The last stage of model construction consists in reducing the mean structure. This process is carried out using a likelihood ratio test. We stress here that the reduction is performed by fitting the model with a mixed-effects approach using maximum likelihood (ML) rather than restricted maximum likelihood (REML), since the likelihood ratio test is not valid under REML. Combining this way of reasoning with a stepwise procedure based on the BIC value, which is useful to get initial impression of the variables that are less significative, we get to the final model in (5). Note that the full final estimates of the parameters, which are presented in Table 11 in Appendix A, are obtained by fitting the

model using the REML approach.

$$
\begin{cases}
\text{BPRS}_i & = \beta_0 + \beta_{0,T}\text{Trial}_i + \beta_{0,S}\text{Sex}_i + \beta_{0,A}\text{Age}_i + \beta_{0,E}\text{Edu}_i + \beta_{0,B}\text{Bmi}_i \\
& \quad + \beta_{0,J}\text{Job}_i + \beta_{0,AD}\text{Adl}_i + \beta_{0,W}\text{WZC}_i + \beta_{0,C}\text{CDRSB}_i \\
& \quad + (\beta_1 + \beta_{1,A}\text{Age}_i + \beta_{1,E}\text{Edu}_i + \beta_{1,J}\text{Job}_i + \beta_{1,AD}\text{Adl}_i + \beta_{1,C}\text{CDRSB}_i) \cdot \text{Year}_i \\
& \quad + b_{0i} + b_{1i}\text{Year}_i + \varepsilon_i \\
b_i & \sim \mathcal{N}(0, D) \\
\varepsilon_i & \sim \mathcal{N}(0, \Sigma) \\
\Sigma & = \sigma^2 \cdot I_{n_i}
\end{cases}
\tag{5}
$$

Variable	Coefficient	Variable	Coefficient
Intercept	-47.018^{***}	**cdrsb_base**	-0.019^{***}
age	1.635^{***}	**year**	5.674^{***}
edu2	0.112	**age:year**	0.015^{**}
edu3	0.014	**edu2:year**	0.146
edu4	0.123	**edu3:year**	0.253^{**}
bmi	0.156^{***}	**edu4:year**	0.088
job1	-3.952^{***}	**job1:year**	-0.408^{**}
adl	-0.105	**adl:year**	0.051^{**}
wzc1	1.927^{***}	**cdrsb_base:year**	-0.021^{***}

Table 6: Estimated coefficients of the LME model (non-trial terms).

In the mixed effects model (see Table 6), `year` showed the largest effective, indicating a strong annual increase in `BPRS` scores across all patients, irrespective of other factors. `age` also has a significant positive effect on `BPRS`, while the interaction between `year` and `age` is small and positive but also significant. Higher `bmi` and `wzc` (group living in residence) also had high, positive effects on `BPRS`, contributing to increases of 0.156 per point and 1.927 overall for those living in residences. Interesting, relative to the baseline for education (Higher), only the interaction between `year` and `edu3` (Lower Secondary) was significant with a 0.253 increase. All education levels were otherwise statistically insignificant in differences, as well the other interaction terms. Having a job decreased `BPRS` levels both as a baseline, and on a per-year basis.

A relevant feature, that we will analyze later in the document, is given by the matrix D, whose estimate is given in (6).

$$D = \begin{pmatrix} 0.584 & 0.237 \\ 0.237 & 0.824 \end{pmatrix} \tag{6}$$

At the end of the analysis, it is informative to examine Figure 9, where we plotted for every patient the (estimated) random effects for the intercept (on the horizontal line) coupled with the (estimated) random effects for the slope (on the vertical line) from the final model (5).

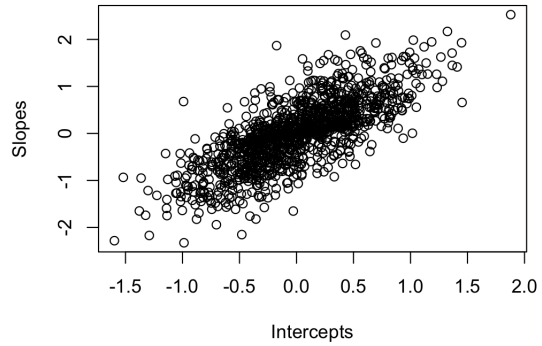


Figure 9: Scatterplot of subject-specific random effects (intercepts and slopes)

The plot clearly confirms the positive correlation between the two random effects we introduced in our model (which can also be derived by the fact that the non-diagonal element of (6) is positive). This finding is coherent with the biological interpretation available in the literature: as a matter of fact, studies (such as [3]) show that patients with higher baseline BPRS levels tend to get more worse over time compared to those with lower baseline BPRS.

In addition, from Figure 9 we see that there are two patients whose estimated random effects deviate substantially from the overall pattern: one displays exceptionally high intercept and slope estimates, while the other exhibits exceptionally low values. Such results can be valuable for identifying potential measurement errors, as outliers are often associated with unintended inaccuracies in data collection.

4 Discussion of the results

In this section, we aim to compare and analyze the models examined in the previous part of our study, beginning with the mean structure and subsequently addressing the covariance structure.

4.1 Mean Structure

Starting from the mean structure, it is useful to summarize the variables that play a key role in our models in Table 7.

Model	**Intercept**	**Temporal Slope**
Multivariate Model	Trial, Age, BMI, Job, ADL, WZC, CDRSB	Age, ADL, CDRSB
Two-Stage Analysis	Trial, Age, BMI, Job, ADL, WZC, CDRSB	Age, Job, ADL, WZC, CDRSB
Linear Mixed Effects	Trial, Sex, Age, Edu, BMI, Job, ADL, WZC, CDRSB	Age, Edu, Job, ADL, CDRSB

Table 7: Comparison between the fixed effects in various models.

First of all, we notice that all the models we fitted present more or less the same covariates, and that they are coherent with the preliminary data analysis we performed in Section 2. As a matter of fact, speaking of the intercept, we have that the variables `Trial`, `Age`, `BMI`, `Job`, `ADL`, `WZC` and `CDRSB` are identified as playing a significant role when studying the baseline levels of BPRS. Looking at the temporal slope, the variables `Age`, `ADL` and `CDRSB` are identified to be significant across all models. The only difference with the results from the exploratory data analysis is given by the `BMI` and the `ab_base` variables: as a matter of fact, the first one is never identified as significant in Section 2, while the latter one is never identified by our models.

Before looking at the differences between the models, let us comment quickly what these variables tell us from a biological point of view. To begin with, certain basic biological indicators of the patients, such as age, play a crucial role in the temporal evolution of BPRS values: this finding is entirely consistent with medical interpretation of the results. Interestingly enough, the variable that quantifies the daily activities of the patients influences both the baseline and the temporal evolution, as the variable that indicates if the patient is working or not does on the baseline. This is exactly what we hope to get when studying Alzheimer's disease, as it is reasonable that this neurodegenerative disease is highly linked to the patient's activities. In addition, it is completely reasonable that there is a significant difference in the baseline of patients that are in a retirement house: as a matter of fact, it is common that patients prefer to stay in their own homes as long as their condition is not unbearable, and this means that people in the retirement houses are usually in a worse condition.

A notable observation is that the variable `Trial` is definitely important for the baseline prediction across all the models: we can give a quick explanation of this fact. As a matter of fact, the BPRS quantity is not something that can be measured objectively using physiological indicators (such as in the case of the Amyloid-Beta level), but it based

on a scale that is interpreted by the physician. So it is frequent that a physician will interpret the scale a little bit differently than another one, leading to observer bias in the baseline values. In addition, it can also happen that different trials accept patients starting from different stages of the disease, leading to different starting levels. Interestingly enough, from our analysis we observed that the temporal evolution is not influenced by this variable: a possible explanation is given by the fact that we can safely assume that the doctors maintain a consistent way to evaluate the patients over time, and so the observer bias does not influence the temporal slope. In addition, although there is a modest correlation between the onset of the disease and its progression (see [3]), the fact that different trials examine varying stages of Alzheimer's does not appear to play a decisive role.

The CDR score emerges as a key metric both at baseline and throughout disease progression, with a straightforward biological interpretation. Specifically, the CDR reflects the severity of dementia in terms of everyday functioning, whereas the BPRS captures its psychiatric dimension, such as anxiety and depression. It is therefore reasonable to expect that, as the disease advances, patients' ability to manage daily activities declines, and their psychiatric condition deteriorates accordingly.

Another interesting observation that we can make on the mean structures is that in all the three models that we considered the baseline level of the Amyloid-Beta and Tau proteins are not included as fixed-effects. This is not what we would hope to get: as a matter of fact, these two proteins are thought to be one of the main causes behind Alzheimer's disease [1], and so we would like to get a model that takes into consideration these quantities. A possible explanation for the lack of statistical significance of these biomarkers could be the limited variability of their baseline measurements (see Table 1). This suggests that including baseline levels as predictors may require either more informative prior structures, longitudinal biomarker measurements, or alternative parameterizations to better capture their effect.

At this point we can analyze more in detail the differences in the means structures of our models. We will take as the benchmark model the linear mixed effects one, the state-of-art for this type of data.

We begin by examining the final reduced multivariate model. The main distinction from the LME approach is that the latter identifies Educational Level as a significant variable both at baseline and during temporal evolution, while Job quantity is significant in the slope. More generally, the LME model incorporates a more complex mean structure, which is a recurring feature of such analyses, whereas the multivariate model is more rigid and does not account for potential interactions between or within subjects.

We now turn to a comparison between the two-stage analysis and the linear mixed-effects (LME) model. Similar to the multivariate case, the LME model identifies educational level as a significant factor both at baseline and during temporal evolution. A noteworthy distinction arises with the variable `WZC`: while it is included in the temporal evolution in the two-stage analysis, it is not considered in the LME model. At first glance, this may seem inconsistent, given that the LME framework essentially represents a direct

way of fitting the two-stage approach to the data. Consequently, obtaining two different mean structures appears unusual. This discrepancy, however, can likely be attributed to the regression of subject-specific coefficients in the second stage, where these coefficients are estimated from models based on relatively few observations. Their associated uncertainty therefore requires more careful treatment, which may explain the differences observed relative to the LME model.

Finally, it is interesting to analyze some of the estimates we got from fitting the Linear Mixed Effects model to understand the effect of some variables on the BPRS value. For example, we have that the coefficient linked to the `Age` variable in the intercept is positive: this means that the baseline level of the BPRS quantity increases for older people. In addition, also the coefficient representing the interaction between age and the temporal slope is positive: this means that the disease tends to get worse in older patients.

Another interesting finding is that `ADL` has a negative effect on `BPRS` at baseline. This suggests that Alzheimer's tends to manifest with lower values in patients who remain more active (*i.e.*, those with higher `ADL` scores). The `WZC` variable shows a positive coefficient for group 1 (individuals living in retirement homes), indicating that these patients begin with higher `BPRS` values. This is consistent with the observation that patients generally prefer to remain at their own home for as long as possible. Another notable result is that both coefficients associated with category 1 of the `Job` variable (*i.e.*, individuals who are employed) are negative, implying that working appears to slow the progression of Alzheimer's. Finally, the CDR is linked to a slightly negative coefficient both at baseline and in the temporal slope. This may be interpreted as follows: as the disease worsens in terms of daily activity capacity, patients may not fully perceive the decline, and their psychological condition is, to some extent, influenced, or at least perceived by observers, in a relatively positive way.

4.2 Covariance Structure

The last point of our analysis concerns the covariance structure among the models. It is particularly interesting to compare the multivariate covariance structure and the one that comes from the LME model: as a matter of fact, in the first case we decided to keep an heterogeneous AR(1) structure, which models explicitly the serial correlation that we found in the preliminary data analysis. However, when fitting the LME, we kept an unstructured error matrix D for the random effects, while we found out that it was not necessary to include serial correlation in the errors ε_i.

This finding first seems to be inconsistent with the exploratory data analysis. However, we note that in this stage we keep into consideration explicitly the longitudinal structure of the data. So, the fact that serial correlation is now excluded means that adding the random effects is sufficient to explain this variability.

We can finally interpret the estimated matrix D for the LME model, whose estimation is explicitly reported in (6). It is interesting to notice that the non-diagonal elements are significantly different from 0: this means that there is (positive) correlation between the baseline value and the temporal evolution of BPRS. As we have already noticed in the

previous section, this has a clear interpretation: the disease gets worse when the starting point is worse (see [3]).

5 Conclusion and further studies

To summarize the results of our analysis, we can start by noting that the clinical center, the age, the fact of living in a nursing home, the body mass index, the employment status, and the CDR-SB significantly influence the outcome variable at the baseline. Moreover, it is evident that daily activities, CDR-SB, and age affect the progression of the disease over time. On the other hand, no evidence was found for an effect of income, education, and sex. For a detailed understanding of the direction of these covariates, refer to Section 4.

As previously mentioned, the impact of the variables `CDRSB`, `ABPET`, and `TAU` is not fully explored here; available within the dataset is the longitudinal observations of these variables, however, only the baseline values are included in the exploration and modeling of this report. Including the full scope of these variables may unlock additional insights into the evolution of `BPRS` over time. This is especially pertinent considering that their are existing studies in the field that link these variables to cognitive function (specifically in relation to Alzheimer's disease).

It may also be interesting for future research to apply specific techniques capable of handling informative dropout since, as previously mentioned, it may lead to biased estimates.

Acknowledgements

AI was used throughout this project to enhance productivity and accuracy. It helped refine and review text, assisted with R 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.
The two-stage analysis was performed with the help of the code in [4].

6 References

References

[1] Michiel Bertsch, Bruno Franchi, Maria Carla Tesi, and Veronica Tora. "The role of Aβ 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 1 LDA - Code*. https://github.com/dcapelli02/Project-1-LDA. Accessed: 20 November 2025. 2025.

[3] R. E. Kennedy, G. R. Cutter, G. Wang, and L. S. Schneider. "Using baseline cognitive severity for enriching Alzheimer's disease clinical trials: How does Mini-Mental State Examination predict rate of change?" In: *Alzheimer's & Dementia (New York, N. Y.)* 1.1 (2015), pp. 46–52. DOI: 10.1016/j.trci.2015.03.001. URL: https://doi.org/10.1016/j.trci.2015.03.001.

[4] Wolfgang Viechtbauer. *Two-Stage Analysis versus Linear Mixed-Effects Models for Longitudinal Data*. https://www.metafor-project.org/doku.php/tips:two_stage_analysis. Accessed: 20 November 2025. 2022.

A Additional Tables

Table 8: Estimated coefficients of the Multivariate model, with significance indicated by asterisks.

Variable	Coefficient	Variable	Coefficient
(Intercept)	-46.466^{***}	**trial19**	-1.209^{**}
trial2	4.232^{***}	**trial20**	-4.362^{***}
trial3	-2.363^{***}	**trial21**	0.184
trial4	3.490^{***}	**trial22**	1.969^{***}
trial5	3.890^{***}	**trial23**	-2.421^{***}
trial6	6.175^{***}	**trial24**	-2.305^{***}
trial7	1.879^{***}	**trial25**	-4.591^{***}
trial8	4.573^{***}	**age**	1.632^{***}
trial9	-0.231	**bmi**	0.145^{***}
trial10	0.194	**adl**	-0.126
trial11	-3.269^{***}	**wzc1**	1.899^{***}
trial12	-2.484^{***}	**cdrsb_base**	-0.021^{*}
trial13	0.481	**job1**	-3.895^{***}
trial14	-1.492^{**}	**year**	5.657^{***}
trial15	-3.116^{***}	**age:year**	0.017^{**}
trial16	-2.515^{***}	**adl:year**	0.058^{**}
trial17	3.627^{***}	**cdrsb_base:year**	-0.023^{***}
trial18	1.669^{**}	**job1:year**	-0.421^{**}

Table 9: Fixed effects for the intercept in the two-stage model, with significance indicated by asterisks.

Variable	Coefficient	Variable	Coefficient
(Intercept)	-48.724^{***}	**trial16**	-2.813^{***}
trial2	3.347^{***}	**trial17**	3.007^{***}
trial3	-3.092^{***}	**trial18**	1.209^{*}
trial4	2.504^{***}	**trial19**	-1.586^{***}
trial5	3.517^{***}	**trial20**	-4.618^{***}
trial6	5.652^{***}	**trial21**	-0.103
trial7	1.853^{***}	**trial22**	1.663^{***}
trial8	4.026^{***}	**trial23**	-3.225^{***}
trial9	-0.436	**trial24**	-3.272^{***}
trial10	-0.531	**trial25**	-5.155^{***}
trial11	-3.461^{***}	**age**	1.647^{***}
trial12	-3.128^{***}	**bmi**	0.167^{***}
trial13	0.179	**job1**	-4.338^{***}
trial14	-1.747^{***}	**wzc1**	1.741^{***}
trial15	-3.675^{***}	**cdrsb_base**	-0.019^{**}

Table 10: Fixed effects for the slope in the two-stage model, with significance indicated by asterisks.

Variable	**Coefficient**
(Intercept)	1.257
age	0.073^{*}
job1	-1.215^{*}
wzc1	0.190^{*}
cdrsb_base	-0.020^{***}

Table 11: Estimated coefficients of the LME model, 3 decimal places, with significance indicated by asterisks.

Variable	Coefficient	Variable	Coefficient
Intercept	-47.018^{***}	**trial22**	2.055^{***}
trial2	3.861^{***}	**trial23**	-2.733^{***}
trial3	-2.704^{***}	**trial24**	-2.664^{***}
trial4	3.021^{***}	**trial25**	-4.999^{***}
trial5	3.669^{***}	**age**	1.635^{***}
trial6	6.070^{***}	**edu2**	0.112
trial7	1.990^{***}	**edu3**	0.014
trial8	4.526^{***}	**edu4**	0.123
trial9	-0.270	**bmi**	0.156^{***}
trial10	-0.167	**job1**	-3.952^{***}
trial11	-3.043^{***}	**adl**	-0.105
trial12	-2.628^{***}	**wzc1**	1.927^{***}
trial13	0.294	**cdrsb_base**	-0.019^{***}
trial14	-1.706^{***}	**year**	5.674^{***}
trial15	-3.473^{***}	**age:year**	0.015^{**}
trial16	-2.581^{***}	**edu2:year**	0.146
trial17	3.538^{***}	**edu3:year**	0.253^{**}
trial18	1.693^{***}	**edu4:year**	0.088
trial19	-1.294^{***}	**job1:year**	-0.408^{**}
trial20	-4.350^{***}	**adl:year**	0.051^{**}
trial21	-0.019	**cdrsb_base:year**	-0.021^{***}