Longitudinal Data Analysis

Project 1

Master of Statistics and Data Science
Hasselt University
2024-2025

Group 7:
Luong Vuong (2365900)
Neamin Berhe (2159969)
Thanh Chuong Cao (2262621)
Trang Tu Pham Hien (2261972)

Lecturers:
Prof. Dr. Geert Molenberghs
Prof. Dr. Geert Verbeke

November 12, 2024

Abstract

Background: Hemodialysis patients commonly have low hemoglobin (Hgb) levels, which is adversely related to an increased odds of death. Hemoglobin concentrations in these patients are shown to be associated with factors like Erythropoietin dose, iron status, age, and sex.
Objectives: We aimed to study the evolution of Hgb over time in hemodialysis patients and evaluate the relationship between the evolution and Erythropoietin dose, iron deficiency status, age, and sex.
Methodology: Our dataset included longitudinal Hgb measurements for patients with renal deficiency receiving hemodialysis, in which Hgb was measured monthly. We used graphical techinques to explore the mean, variance and correlation structure of the data. We described Hgb levels over time using simple techniques including analysis at each time point, analysis of increments, and analysis of area under the curves. Changes of Hgb over time and effect of EPO dose and other factors were modeled using three approaches: multivariate analysis and two-stage analysis and linear mixed model, in which different mean and variance structures were explored to find the most parsimonious models.
Results: The exploratory analysis suggested that there existed both within-individual and between-individual variability in Hgb levels. The multivariate model and linear mixed model gave similar findings, in which baseline Hgb level were significantly higher in patients with older age, male sex, higher EPO dose, and presence of iron deficiency. Hgb decreased slightly over time, this effect was more pronounced in patients with iron deficiency, but was reversed in patients receiving higher dose of EPO. In the two-stage analysis, age was found to have a modest positive association with baseline Hgb levels, and as age increased, the rate of change in Hgb over time decreased slightly.
Conclusions: we observed that in hemodialysis patients, Hgb levels decreased slightly over time. EPO mitigated this decline as higher EPO doses led to increased Hgb levels over time. Patients with higher EPO doses also had higher baseline Hgb, as did older and male patients. Iron deficiency was associated with a faster decline in Hgb levels over time.
Key Words: Linear-mixed effects model; Two-stage analysis, Multivariate-model, Hemoglobin, Hemodialysis.

1 Introduction

Anemia is a common manifestation and/or complication in patients with kidney failure undergoing hemodialysis. In these patients, hemoglobin (Hgb) can be very low [1], which is associated with a significant increase in the odds of death [2]. These patients are often treated with Erythropoietin (EPO), which is a hormone stimulating bone marrow to produce red blood cells, and intravenous iron to improve the Hgb level [3].

Besides EPO and iron status, demographic factors such as age and sex are found to be significantly associated with Hgb concentrations [2]. In this study, we aim to investigate the evolution of Hgb (Hgb) over time in hemodialysis patients and examine how this evolution is influenced by EPO dose, iron deficiency status, age, and sex.

2 Data Description

The dataset included Hgb concentrations measured longitudinally every month in 3823 patients with renal deficiency receiving hemodialysis. These patients were followed up for a maximum of 6 months. The longitudinal Erythropoietin dose for these patients was also collected, in which the EPO for the next month was decided by the Hgb level of the current month. Besides Hgb concentrations and EPO dose, several other patient characteristics were available. The study design producing the dataset is balanced but the resulting dataset is unbalanced due to missingness. The variables in the dataset were patient id, a month at which the measurement was taken (1 to 6), the age of the patient at baseline (year), the sex of the patient, the Hgb level (g/dl) at each month, a dose of EPO to be administered during the following month (IU/kg/week, IU: international Unit), and an indicator for iron deficiency status at each month.

3 Methods

3.1 Exploratory Data Analyses

Before any further investigation, exploratory analyses were conducted on the data to discover patterns, data structure and spot any existing anomalies and to obtain clear insight on conceivable implications before model building and hypothesis testing. This study used individual-specific profiles to explore the variability between and within patients. Graphical methods were used to explore the mean structure, variance structure and correlation structure. Besides, summary tables were used to give an overview of the descriptive information from the data.

Additionally, we selected the best possible regression relation between Hgb and time by testing a linear relationship and a quadratic relationship. We fitted each model for each subject who has at least 4 Hgb measurements, and then we combined the results in the form of meta-analysis using R^2_{meta}. We also performed the F test for model extension comparing the model with only the linear term and the model with additional quadratic term.

3.2 Summary Statistics

To provide general information about the evolution of Hgb level over time in hemodialysis patients, we performed some summary statistics calculations, namely, Analysis at Each Time Point, Analysis of Increments and Analysis of Are Under the Curve. Since Hgb level in male is generally higher than that of female [4], we examined how difference between female and male dialysis patient's Hgb level in the first analysis. The second summary statistics, Analysis of Increments, was a simple method to compare evolutions between subjects. Specifically, we analyzed the subject-specific changes of Hgb from baseline (y_{i1}) to the last observation (y_{in_i}): $y_{in_i} - y_{i1}$. Analysis of Area Under the Curve provided an "overall" differences across individuals without problem of multiple testing and assumption of balanced data. This was calculated by this equation: $AUC_i = (t_{i2} - t_{i1}) \times (Y_{i1} + Y_{i2})/2 + (t_{i3} - t_{i2}) \times (Y_{i2} + Y_{i3})/2 + ...$, where AUC_i indicated Area Under the Curve of individual i, t_{ij} indicated the month j ($j = 1, ..., 6$), Y_{ij} indicated the Hgb level of individual i at month j. We also predicted AUC value by baseline characteristics, such as age and sex by using linear model: $AUC_i = \beta_0 + \beta_1 age_i + \beta_2 sex_i$.

3.3 Inferential Data Analysis

To investigate the evolution of Hgb over time and the influence of EPO dose, iron deficiency status, age and sex, we employed three modeling approaches, namely multivariate model, two-stage analysis and linear mixed model, then the results from three approaches were compared. These approaches were chosen due to their appropriateness to model longitudinal data with repeated measures [5]. All models were fitted using restricted maximum likelihood (REML). REML is preferred to maximum likelihood (ML) as REML estimator is unbiased in linear mixed model, which is not the case for the ML estimator [5, 6].

Prior to modeling, transformation for the variable month and EPO dose was performed. Month was recoded to 0 to 5 so that the first month represented baseline. Because the EPO dose recorded for each month was the dose to be administered during the following month, the EPO dose of the previous month was used to model its effect on Hgb evolution. The EPO dose of the first month was assumed to be 0 IU/kg/week as the patients was not on EPO at baseline. EPO dose, age and month were considered continuous variables, and sex and iron deficiency as dummy variables in all models.

3.3.1 Multivariate model

Monthly Hgb levels were firstly predicted by multivariate model, accounting for various factors that could influence Hgb concentration. The initial model included independent variables including sex, age, iron deficiency status, time and EPO dose, and interaction terms as described in the equation 1.
Let $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{i6})^t$ the vector of monthly repeated measurement of Hgb. The multivariate model is denoted by: $Y_i = X_i\beta + \epsilon_i$ where X_i are the covariates; specifically:

$$\begin{aligned} Y_{ij} = {} & \beta_0 + \beta_1 \,\text{age}_i + \beta_2 \,\text{sex}_i + \beta_3 \,\text{iron}_{ij} + \beta_4 \,\text{dose}_{ij} + \beta_5 \,\text{month}_j + \beta_6 \,\text{month}_j^2 \\ & + \beta_7(\text{dose}_{ij} \times \text{age}_i) + \beta_8(\text{dose}_{ij} \times \text{sex}_i) + \beta_9(\text{dose}_{ij} \times \text{iron}_{ij}) + \beta_{10}(\text{dose}_{ij} \times \text{month}_j) \\ & + \beta_{11}(\text{month}_j \times \text{age}_i) + \beta_{12}(\text{month}_j \times \text{sex}_i) + \beta_{13}(\text{month}_j \times \text{iron}_{ij}) + \epsilon_{ij} \end{aligned} \tag{1}$$

We then reduced the mean structure, using F-test to compare the reduced models with Model 1. We tested the significance of quadratic term of time and removed it if this effect was obsolete. We continued reducing interaction terms between dose and sex, age, and iron. After that, we tested if we could omit the effect of month interacting with age and sex. Lastly, we further tried removing interaction of month and iron or dose of EPO. Such steps were guided by F-test in which a p-value ≥ 0.05 indicated the nested model could be reduced to be the more parsimonious one. Therefore, we could find the most parsimonious mean structure—the model with the fewest parameters that best explained the data.

As Hgb levels measured at different time points of the same patient were correlated, we compared several potential covariance structures to determine the one that best fit the data. Specifically, they were: unstructured type, simple type, compound symmetry type, banded type, first-order autoregressive type and Toeplitz type. This step was guided by comparing the -2*Log Likelihood value using the Likelihood ratio test. This helped identify the most appropriate model specification for covariance structure; estimation of effect sizes could then be more precised.

3.3.2 Two-stage analysis model

In longitudinal studies with a large number of repeated measurements and potentially irregular time intervals, traditional multivariate models may not be optimal due to their stringent assumptions and complexity, especially when dealing with unbalanced data. In such cases, a two-stage analysis has been proposed as an alternative, acknowledging its limitations [5].

Stage 1 analysis

The first stage involves fitting individual linear regression models separately for each patient to understand within-subject variability. This approach yields different slopes and intercepts for each subject, capturing their unique trajectories over time.

For our specific case, based on exploratory data analysis, the majority of the trends can be represented by a linear time trend, as subjects showed smooth evolution over time. Therefore, individuals with at least two observations were included in the analysis.

The subject-specific regression model for patient i is:

$$Y_{ij} = \beta_{1i} + \beta_{2i} t_{ij} + \epsilon_{ij}, \quad j = 0, \ldots, 5 \tag{2}$$

where:

- Y_{ij} is the Hgb measurement for subject i at time j.
- β_{1i} represents the intercept for subject i, indicating the initial level of Hgb measurement.
- β_{2i} represents the slope (or monthly trend) for subject i, showing the rate of change in Hgb level per month.
- t_{ij} is the month variable representing the month of observation, measured from 0 to 5.
- ϵ_{ij} is the error term for subject i at time j, capturing the residual variation in Hgb levels not explained by the linear trend. It is assumed to be independent and normally distributed with mean zero and covariance matrix $\sigma^2 I_{n_i}$. I_{n_i} is the n_i-dimensional identity matrix.

Stage 2 analysis

While the first stage provides individual-specific coefficients, our interest lies in understanding the average evolution and how covariates like age influence these trends across subjects. The second stage aims to explain the variability in the subject-specific regression coefficients using known covariates, capturing between-subject variability.

The subject-specific coefficients are modeled as:

$$\beta_i = K_i \beta + b_i \tag{3}$$

where:

- β_i is the $q \times 1$ vector of subject-specific coefficients (β_{1i}, β_{2i}).
- K_i is the $q \times p$ matrix of covariates for subject i
- β is the $p \times 1$ vector of unknown regression parameters (fixed effects).
- b_i is the $q \times 1$ vector of random effects for subject i, assumed to be normally distributed with mean zero and covariance matrix D.

Specifically, we modeled the between-subject variability by examining the effect of age on the intercepts and slopes:

$$\begin{aligned} \beta_{1i} &= \beta_0^{(1)} + \beta_{\text{Age}}^{(1)} \cdot \text{Age}_i + b_{1i} \\ \beta_{2i} &= \beta_0^{(2)} + \beta_{\text{Age}}^{(2)} \cdot \text{Age}_i + b_{2i} \end{aligned} \tag{4}$$

where:

- β_{1i} and β_{2i} are the intercept and slope for subject i, respectively.
- $\beta_0^{(1)}$ and $\beta_0^{(2)}$ average intercept and slope after adjusting for age

- $\beta_{\text{Age}}^{(1)}$ and $\beta_{\text{Age}}^{(2)}$ represent the effect of age on the intercept and slope, respectively.
- Age_i is the age of subject i.
- b_{1i} and b_{2i} are the random effects for the intercept and slope of subject i, capturing the deviation of each subject from the population average, assumed to be normally distributed with mean zero and the covariance matrix D.

3.3.3 Mixed-effect model

The general linear mixed model (LMM) is formulated as:

$$\begin{cases} Y_i = X_i\beta + Z_i b_i + \epsilon_i, \\ b_i \sim N(0, D), \\ \epsilon_i \sim N(0, \Sigma_i), \\ b_1, \ldots, b_n,\ \epsilon_1, \ldots, \epsilon_n \text{ are independent} \end{cases} \tag{5}$$

where Y_i is the n_i-dimensional vector for subject i, $1 \leq i \leq N$, where N is the number of subjects. X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariates. β is the p-dimensional vector containing the fixed effects, b_i is the q-dimensional vector containing the random effects, and ϵ_i is the n_i-dimensional vector containing residual components. Finally, D is the general $(q \times q)$ covariance matrix. We employed the hierarchical interpretation of the linear mixed model, which implies:

$$Y_i | b_i \sim N(X_i\beta + Z_i b_i, \Sigma_i) \quad b_i \sim N(0, D) \tag{6}$$

We employed a four-stage model building process as recommended in the literature [6] [5]. First, an elaborated LMM was fitted. The outcome variable was monthly Hgb level (g/dl). The fixed effects included all main effects of EPO dose, age, sex, iron deficiency, and month, and the interactions between dose and age, dose and sex, dose and iron, dose and month, iron and month, age and month, and sex and month. The random effects included random intercept and random slope for month, to capture the individual deviation from the average of baseline Hgb level and its evolution over time. We assumed an unstructured variance-covariance structure for the random effects (D) and simple residual covariance structure. The assumptions of the LMM (the explanatory variables are related linearly to the response, the errors have constant variance and the errors are normally distributed) were checked using diagnostic plots, from which it was concluded that these assumptions were met. Second, the residual covariance structure was refined by adding serial correlation functions. Models with Gaussian serial correlation, exponential serial correlation, and without serial correlation were compared using REML log-likelihood, the model with highest log-likelihood was chosen. Third, the random effect structure was reduced. The model with random intercept and random slope for time was compared with the model with only random intercept and model with only random slope, using Likelihood ratio test. The reduced model with comparable goodness of fit to the full model (p-value of Likelihood ratio test ≥ 0.05) was chosen. If both models gave similar goodness of fit to the full model, the model without random effect would be compared with them. Finally, the mean structure was reduced using F-test, with Satterthwaite approximation for the denominator degrees of freedom . The model from step 3 was compared with the models removing the interaction effects one by one: dose*age, dose*sex, dose*iron, age*month, sex*month, iron*month, and dose*month. The reduction was stopped when F-test was significant ($p < 0.05$). The model obtained from this step is the final model.

Additionally, to compare with the two-stage model, we fitted a LMM with only fixed effect of age, month, age*month interaction, random intercept and random slope for month, and similar residual covariance structure to the above final model.

4 Results

4.1 Exploratory Data Analyses

A total of 3823 patients were included in the study in which patients were followed up every month for the maximum of 6 months. Among these, 925 (24.2%) patients had at least 4 Hgb measurements. The number and percentage of patients having Hgb measured over time are presented in table 1. 3174 patients were measured on the first month, and this number decreased over time, to only 37 patients at the end of the 6^{th} month. 9.02% of total patients had missing month data. Additionally, 4.63% and 3.95% of total patients had EPO dose and Hgb missing, respectively. Nevertheless, we assumed all the missingness to be completely at random in this study [5]. Furthermore, there were a total of 9604 Hgb measurements, with the median number of measurements per patient of 3.

Table 1: Number of patients over time (NA = not available)

Month	Number of Patients	Percentage (%)
1	3174	83.01
2	2777	72.63
3	2183	57.09
4	1271	33.24
5	212	5.55
6	37	0.97
NA	345	9.02

4.1.1 Individual Profile

The interest in the longitudinal data analysis is the evolution of individual patients over time, and this gives an idea about the most appropriate subject-specific regression as well as providing an insight of the both between variability and within variability for the subjects. Figure 1 shows that there seems to be much within-subject variability and between-subject variability among the patients. It can also be observed that the patients have a different Hb at the start of the experiment and this also changes with time. This variation suggests that a random intercept and random slope model could be a plausible starting point.

The overall R^2_{meta} of the model assuming a linear relationship between each patient's Hgb and time equals 0.52 (figure 2). This increased by about 40% to 0.73 in the model assuming a quadratic relationship. This results implied that the linear relationship model explained about 52% of the total variation in the response while the quadratic relationships explained 73% of the total variation in the response. Such an increase in R^2_{meta} can partly be explained partly by the addition of the higher order of time. Additionally, the extension of the model to include the quadratic term was proven to be significant using the F test (p value $< .0001$). The increase in R^2_{meta} together with the significant F test suggest that the quadratic time effect can be considered in the models.

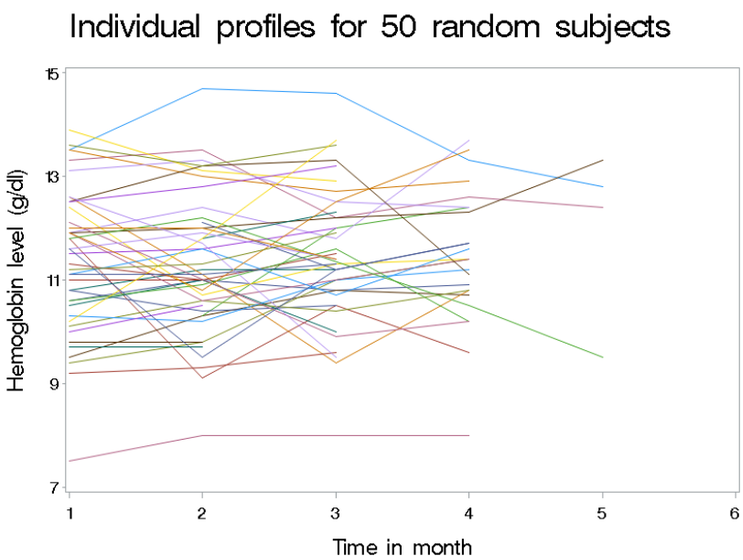

Figure 1: Individual Hgb profiles for 50 randomly selected subjects

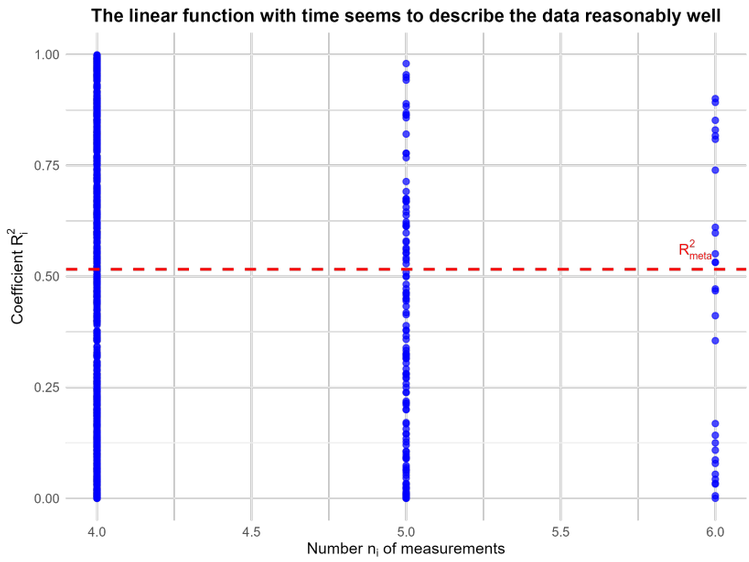

Figure 2: R squared under the linear model

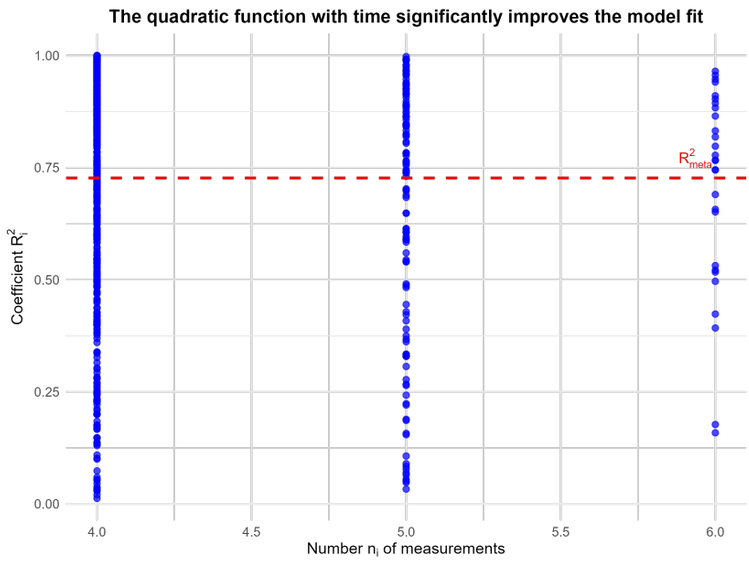

Figure 3: R squared under the quadratic model

4.1.2 Mean Structure

This section describes the average evolution of Hgb (in g/dl) of patients over time for the overall population. The mean structure gives the marginal relationship of the response with time. Figure 4 shows the mean structure of the Hgb. The structure indicates a linear trend with time. The standard error bar on the mean also becomes wider. The inflated standard errors bars may be as a result of many dropouts or by the variability of the patients. Although this variation exists, the linear mean model might be appropriate to model the mean structure, but this is subject to a formal test.

4.1.3 Variance Structure

After studying the mean structure, the evolution of variance is essential in building a longitudinal model [5]. Standardized squared residuals obtained from the mean structure were used to construct the variance function. The figure 5 suggests a stable variance over time with an increase in standard error at each time point. This increase in the standard error is expected because of the attrition that exists in the data. The constant variance suggests a random intercept model might be appropriate to describe the data.

4.1.4 Correlation Structure

In this study, we explored the correlation structure using both a correlation matrix and a scatter plot matrix. The scatter plot matrix, shown in Figure 6, illustrates that the off-diagonal elements from pairs of measurement occasions demonstrate relatively stable correlations over time, regardless of the time gap. This pattern suggests that a covariance matrix with constant correlation, such as the compound symmetry structure, might be appropriate.

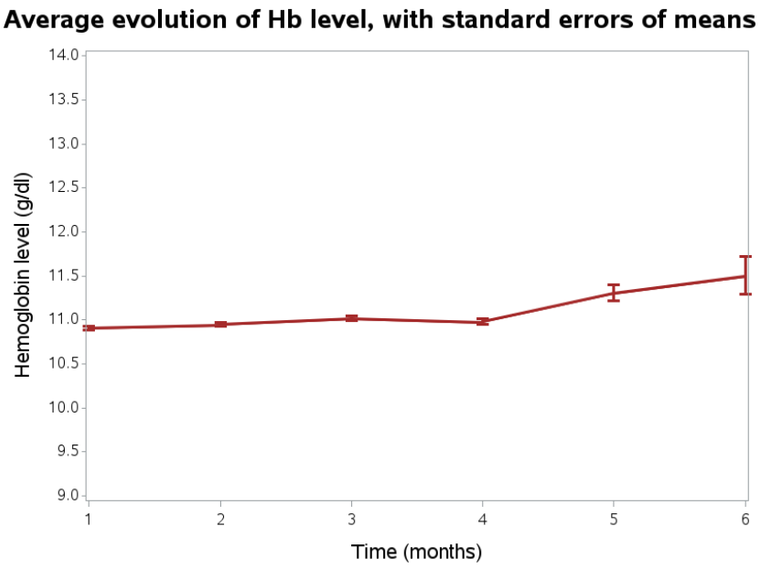

Figure 4: Average evolution of Hgb level, with standard error of means

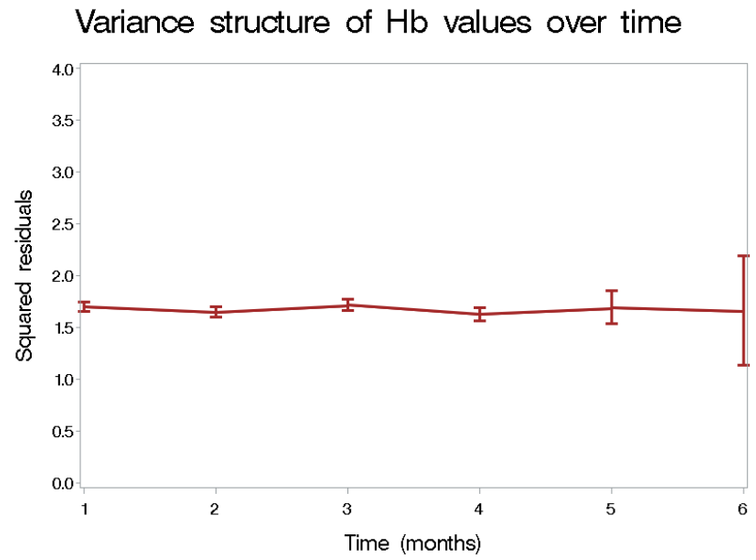

Figure 5: Variance structure of Hb values over time

Conversely, the correlation matrix presented in Table 2 indicates that the correlation between measurements decreases as time progresses. This observation implies that alternative covariance structures, like the Toeplitz or first-order autoregressive structure, may be more suitable.

Table 2: Correlation structure of monthly Hgb level

	Month 1	**Month 2**	**Month 3**	**Month 4**	**Month 5**	**Month 6**
Month 1	1.000	0.699	0.539	0.438	0.393	0.317
Month 2	0.699	1.000	0.745	0.542	0.408	0.080
Month 3	0.539	0.745	1.000	0.716	0.499	0.224
Month 4	0.438	0.542	0.716	1.000	0.640	0.615
Month 5	0.393	0.408	0.499	0.640	1.000	0.726
Month 6	0.317	0.080	0.224	0.615	0.726	1.000

4.2 Summary Statistics

4.2.1 Analysis at Each Time Point

As seen in Figure 7, Hgb levels slightly increased toward the end of the study period. There were minor differences in the median Hgb levels between female and male group over the 6-month period.

4.2.2 Analysis of Increments

As can be seen from figure 8, change of Hgb level from baseline to the last measured level varied dramatically. Majority of patients tended to have minor change in Hgb levels. However, some demonstrated a profound

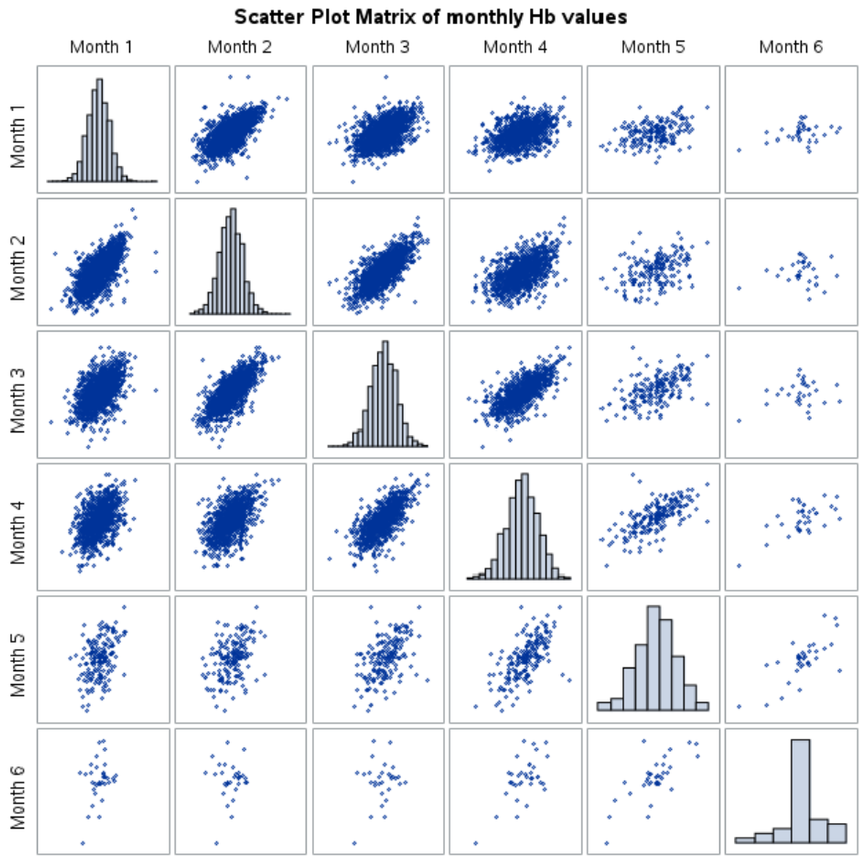

Figure 6: Correlation matrix of monthly Hb values

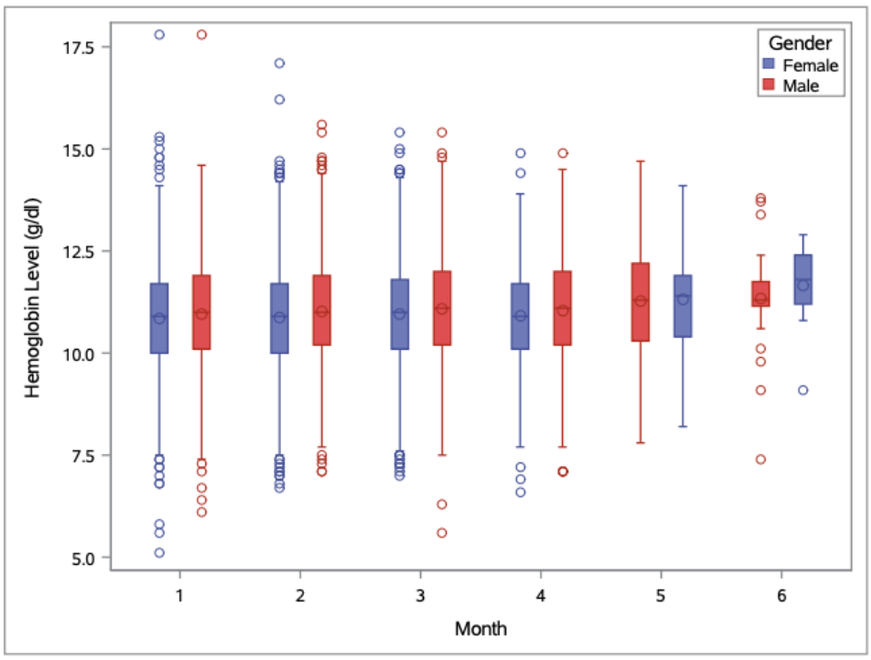

Figure 7: Distribution of Hgb levels by month for female and male patients

drop of Hgb, which was as much as 4 g/dL. Interestingly, some patients showed substantial increases of Hgb concentration.

4.2.3 Analysis of Area Under the Curve

AUC_i value ranged from -22.550 to 82.400 with a mean value of 42.703 and standard deviation of 17.792. The linear model indicated that for each age increase, AUC_i also increased 0.0731, this effect was statistically significant with p-value < 0.001. The effect of sex on AUC_i was of not statistical significant ($p = 0.1914$).

4.3 Multivariate model

As demonstrated by table 3, we first performed a reduction of time quadratic term from the Model 1. Since the type 3 $F - test$ demonstrated a p value of 0.2849 when comparing the reduced model versus the full

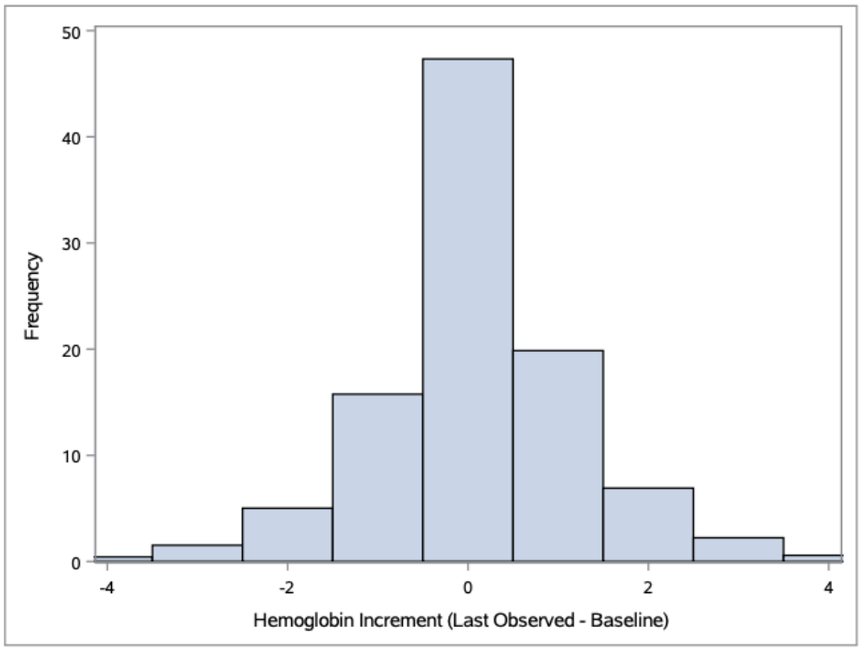

Figure 8: Distribution of Hgb increment from baseline to the last observation

model, it was possible to omit such quadratic term. Further steps to find the most parsimonious model was presented by table 3. All the F-test were conducted against Model 1.

Table 3: Comparisons of mean structures by reducing interaction terms

Reduction		**DF**	**Den. DF**	**F-value**	**Pr>F**
1	month^2	1	3471	1.14	0.2849
2	(1) plus dose*sex, dose*age and dose*iron	3	3471	1.63	0.1805
3	(2) plus month*age and month*sex	5	3471	1.95	0.0836
4	(3) plus month*iron	6	3471	2.71	0.0125
5	(3) plus month*dose	6	3471	6.73	<.0001

As can be seen from the table 3, it was reasonable to reduce five interaction terms: $dose * sex$, $dose * age$, $dose * iron$, $month * age$ and $month * sex$. Further reduction led to significantly worse goodness-of-fit compared to the nested model. Therefore, the final parsimonious model included the fixed effects of age, sex, iron status, dose of EPO, time and interaction terms of time versus time and iron deficiency.

Table 4: Comparisons of potential covariance structures

Covariance	**Parameters**	**-2 Log Likelihood**	G^2	**df**	**p**
Unstructured type	21	27693.1			
Simple type	1	32007.9	4314.80	20	< 0.001
Compound symmetry type	2	28358.7	665.594	19	< 0.001
Banded type	21	28948.1	1255.04	0	< 0.001
First-order autoregressive type	2	27755.5	62.4328	19	< 0.001
Toeplitz type	6	27735.8	42.7125	15	0.002

Table 4 indicated that the unstructured covariance fitted the data best. We then applied this covariance structure to fit the most parsimonious model. We provided estimates of this model in table 5.

4.4 Two-stage analysis

Stage-1 analysis

In the first stage analysis, which focused on within-subject variability, a subject-specific regression model was fitted to represent the monthly Hgb levels (Y_{ij}) of patients as a linear function of the month. The table 6 presents the mean (10.92 g/dl) and range for the intercept and slope parameters obtained from this analysis.

Stage-2 analysis

In the second stage analysis, modeling of the between-subject variability was performed by examining the effect of a known covariate (age) on the trend of Hgb levels across months.

Table 5: Multivariate model: fixed effect estimates

Effect	Estimate	Standard Error	DF	t value	Pr > |t|
Intercept	10.6123	0.08838	3471	120.08	< .0001
Age	0.002793	0.001309	3471	2.13	0.0329
Male	0.1037	0.03924	3471	2.64	0.0083
Iron deficiency	0.1682	0.03182	3471	5.29	< .0001
Dose	0.000485	0.000176	3471	2.76	0.0059
Month	-0.03753	0.01548	3471	-2.42	0.0154
Dose*Month	0.000585	0.000107	3471	5.49	< .0001
Month*Iron deficiency	-0.04714	0.01843	3471	-2.56	0.0106

Variable	N	Mean	Std Dev	Minimum	Maximum
Intercept	2933	10.92	1.36	4.45	18.77
month	2933	0.03	0.69	-5.00	7.10

Table 6: Summary statistics for the intercept and slope (month) in the Stage-1 Analysis.

Table 7: Effect of Age on the Intercept

Variable	Label	DF	Parameter Estimate	Standard Error	T Value	P-value
Intercept	Intercept	2931	10.6090	0.1077	98.53	< 0.0001
AGE	Slope	1	0.00499	0.00168	2.97	0.0030

Table 8: Effect of Age on the Slope

Variable	Label	DF	Parameter Estimate	Standard Error	T Value	P-value
Intercept	Intercept	2931	0.1568	0.05532	2.83	0.0046
AGE	Slope	1	-0.00187	0.000862	-2.17	0.0300

The parameter estimate for the effect of age on the intercept was 0.00499 (g/dl) (Table 7). This positive coefficient suggests that, on average, a unit increase in age is associated with a slight increase in the baseline Hgb level. For every additional year of age, the baseline Hgb level increases by approximately 0.0049 (g/dl).

The parameter estimate for the effect of age on the slope is -0.00187 (g/dl)(Table 8). This negative coefficient implies that as age increases, the rate of change in Hgb level over time decreases slightly. For every additional year of age, the monthly rate of Hgb level change decreases by approximately 0.00187 (g/dl).

4.5 Mixed-effects Model

After employing the four-stage model building procedure, we obtained the final LMM that includes the fixed effect of EPO dose, age, sex, iron deficiency, month, dose*month, and iron*month, a random intercept, and Gaussian serial correlation in the residual covariance structure. Details of the results of each model building stage were provided in appendix 1. The fixed effect estimates are provided in Table 9. The Type 3 tests of fixed effects for this model shows that the effects of EPO dose, age, sex, iron, and interaction between dose and month, iron and month were significant (Table 10).

The mean structure of this LMM was identical to that of the multivariate model. Both models yielded nearly identical results, with consistent coefficient estimates, standard errors, and significance levels. They both indicated that EPO dose had a small positive effect on baseline Hgb ($\hat{\beta} = 0.0005, p < 0.005$) and that higher doses were associated with higher baseline Hgb levels. While both models showed a slight decrease in Hgb over time, the Month effect was statistically significant only in the Multivariate Model ($\hat{\beta} = -0.0375, p = 0.0154$). Nevertheless, the Month effect was close to significant in the LMM ($\hat{\beta} = -0.0282, p = 0.0706$). The interaction between EPO dose and Month ($\hat{\beta} = 0.0006, p < 0.0001$) in both models showed a stronger effect of EPO over time, attenuating the decline in Hgb over time. Iron deficiency also positively affected baseline Hgb ($\hat{\beta} = 0.1682, p < 0.0001$), while its interaction with Month ($\hat{\beta} = -0.0471, p = 0.0106$) suggested a faster decline in Hgb for iron-deficient patients. Age had a small positive effect on baseline Hgb ($\hat{\beta} = 0.0028, p = 0.0329$), and male patients had significantly higher baseline Hgb levels ($\hat{\beta} = 0.1037, p = 0.0083$).

Table 9: Comparison of fixed effect estimates between the linear mixed model and multivariate model

	Linear Mixed Model		Multivariate Model	
Effect	**Estimate (SE)**	**P-value**	**Estimate (SE)**	**P-value**
Intercept	10.6218 (0.0885)	< 0.0001	10.6123 (0.0884)	< 0.0001
EPO dose	0.0005 (0.0002)	0.0046	0.0005 (0.0002)	0.0059
Age	0.0027 (0.0013)	0.0421	0.0028 (0.0013)	0.0329
Male	0.1008 (0.0393)	0.0102	0.1037 (0.0392)	0.0083
Iron deficiency	0.1662 (0.0317)	< 0.0001	0.1682 (0.0318)	< 0.0001
Month	-0.0282 (0.0156)	0.0706	-0.0375 (0.0155)	0.0154
EPO dose * month	0.0005 (0.0001)	< 0.0001	0.0006 (0.0001)	< 0.0001
Iron deficiency * month	-0.0433 (0.0182)	0.0174	-0.0471 (0.0184)	0.0106

Table 10: Type 3 tests of fixed effects of the linear mixed model

Effect	**Numerator DF**	**Denominator DF**	**F value**	**Pr > F**
Age	1	6067	8.02	0.0046
Male	1	6067	4.13	0.0421
EPO dose	1	6067	6.60	0.0102
Iron deficiency	1	6067	27.53	< 0.0001
Month	1	6067	3.27	0.0706
EPO dose * month	1	6067	22.23	<0.0001
Iron deficiency * month	1	6067	5.66	0.0174

The variance components of the LMM are summarized in Table 13. The random intercept had an estimated variance of 0.69, and that of for the Gaussian serial correlation component and measurement error component are 1.92 and 0.29, respectively.

Table 11: Variance components of the linear mixed model

Effect	**Parameter**	**Estimate (SE)**
Covariance of b_i		
$\text{var}(b_{1i})$	d_{11}	0.6919 (0.0556)
Measurement error variance		
$\text{var}(\epsilon_{(1)ij})$	σ^2	0.2923 (0.0164)
Gaussian serial correlation		
$\text{var}(\epsilon_{(2)ij})$	τ^2	1.9210 (0.1312)

We also fitted a LMM that has a similar structure to the two-stage model described in section 4.4 to compare the two approaches. In figure 9 we showed the plot of random slope against random intercept for the LMM (left) and two-stage model (right). The two-stage model demonstrated a slightly downward trend between the random slope and random intercept. Conversely, we observed a strong positive linear relationship between these two random effects for the LMM. This marked differences could be attributed to the fact that the two-stage model does not take into account the correlation structure in the data; hence, do not use the full information of the data. These drawbacks of two-stage modeling is overcome in LMM approach.

5 Discussion

The most parsimonious mean structure to describe the average trends in the multivariate model included the fixed effects of EPO dose, age, sex, iron deficiency, month, as well as interaction terms for dose*month and iron*month. For the covariance structure, an unstructured covariance was the most appropriate to describe variations effectively.

In the two-stage analysis, age was found to have a modest positive association with baseline Hgb levels, with each unit increase in age corresponding to a slight increase in baseline Hgb. Additionally, as age increased, the rate of change in Hgb over time decreased slightly, indicating a slower decline in Hgb among older patients.

The linear mixed model (LMM) used a mean structure identical to that of the multivariate model, with the addition of a random intercept and Gaussian serial correlation in the residual covariance. Additionally, this model showed a strong relationship between subject-specific intercepts and slopes compared to the two-stage model as it accounts for the correlation structure in the data and thus use the full information of the data.

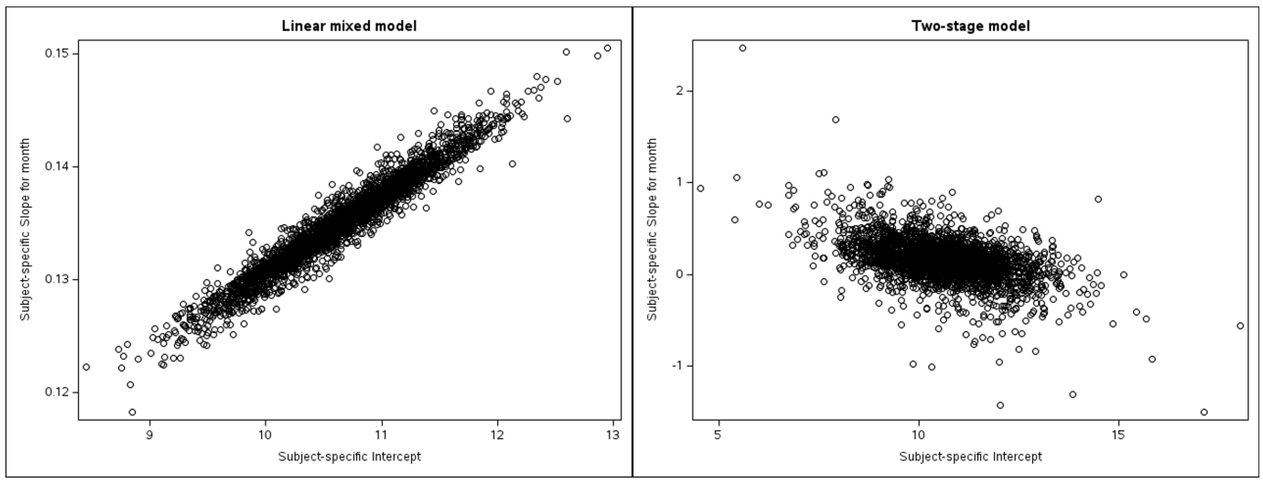

Figure 9: Correlation of subject-specific intercepts and slopes in linear-mixed model (left) and two-stage model (right)

The study investigated the evolution of Hgb level over time in hemolysis patients. In general, there was substantial within- and between-individual variability in Hgb levels during a 6-month period. This could be attributed to several factors, including baseline characteristics and time-varying factors. Our results showed that male patients had approximately 0.1 g/dL higher concentration of Hgb compared to female patient, a difference consistent with observations in healthy individuals. Although Hgb level tended to increase slightly with age, this effect was of not clinical significance. Noticeably, Hgb concentrations decreased mildly over time, indicating that dialysis patients were susceptible to anemia when disease progressed. However, such decline was mitigated by the effect of EPO dose as our results showed that higher doses of EPO were associated with higher Hgb levels. On the other hand, iron deficiency exaggerated the effect time on Hgb levels. Nevertheless, iron deficiency status somehow had a positive effect on the concentration of Hgb. This raised the question about the confounding of other factors, for example, iron supplement could explain this potentially spurious effect.

The multivariate model is well-suited for situations where measurements are taken at fixed time points and there is a small number of measurements per subject [5]. This model treats the repeated measures as multiple outcomes and models them simultaneously, capturing the correlations between measurements at different time points. In our study, Hgb levels were measured monthly over six months, providing data at fixed intervals. Additionally, there are only six measurements per subject, which aligns with the ideal conditions for applying a multivariate model.

Compared to the multivariate model, the two-stage analysis is advantageous when there is an unequal number of measurements per subject and when measurements are not taken at fixed time points. In our study, the former was the case, which provides a rationale for considering the two-stage analysis. However, this method has notable drawbacks. Firstly, subjects with only one observation cannot have a slope estimated in the first stage and are therefore excluded from the analysis. In our dataset, this led to the exclusion of a significant portion of the sample (potentially up to three-quarters), resulting in a substantial loss of information. Secondly, the standard errors of the subject-specific estimates from stage one analysis model can vary widely between subjects due to differences in the number and timing of measurements. The two-stage analysis often ignores this variability. Lastly, the method reduces rich longitudinal data to summary statistics (intercepts and slopes), potentially oversimplifying the complexity inherent in individual trajectories, subsequently impacting the power and standard errors of the estimates. Another loss of information is that this method does not fully take into account the correlation structure in the data.

The LMM addresses many of the limitations associated with the multivariate model and the two-stage analysis. In the context of our study, the LMM allows us to include all subjects, regardless of the number of measurements they have, maximizing the use of available data. Therefore, for our study investigating Hgb levels over time with varying numbers of measurements per subject, the LMM is the most appropriate choice, offering a robust and comprehensive analysis of the continuous longitudinal data.

In this project we employed the hierarchical interpretation of the LMM. This approach offers insights into how the evolution of each subject deviate from the expected evolution. However, the drawback is that it is difficult to be certain that the proposed hierarchical model is correct. Indeed, the model was estimated from the marginal model derived from the proposed hierarchical model, but one marginal model could correspond to different hierarchical models. Thus, a good fit of the marginal model cannot be interpreted as evidence for a hierarchical model [5].

Another limitation of our study is the presence of missing data. There was only 33% of patients with at least 4 measurements, and there was 8.28% of observation with at least one field missing. By performing complete-case analysis, some efficiency may be lost, and it renders biased estimates if the missingness mechanism is not completely at random [7]. Furthermore, we modeled the effect of EPO dose on Hgb level one month later, it is possible that EPO requires a longer duration to take effect [8].

Future studies should aim to collect additional Hgb measurements over extended periods to assess long-term trends. Gathering more detailed patient characteristics will allow for better control of confounding variables (such as duration of dialysis) [9], enhancing the accuracy of the findings. Additionally, recording reasons for patient dropout is crucial to address any outcome-related attrition that may bias effect estimates.

6 Conclusion

In conclusion, we observed that Hgb levels decreased slightly over time. EPO mitigated this decline as higher EPO doses led to increased Hgb levels over time. Patients with higher EPO doses also had higher baseline Hgb, as did older and male patients. While iron deficiency positively influenced baseline Hgb, it was associated with a faster decline in Hgb levels over time.

References

[1] Rajiv Saran, Bruce Robinson, Kevin C Abbott, Lawrence YC Agodoa, Nicole Bhave, Jennifer Bragg-Gresham, Rajesh Balkrishnan, Xue Dietrich, Ashley Eckard, Paul W Eggers, et al. Us renal data system 2017 annual data report: epidemiology of kidney disease in the united states. *American Journal of Kidney Diseases*, 71(3):A7, 2018.

[2] Francois Madore, Edmund G Lowrie, Carlo Brugnara, Nany L Lew, J Michael Lazarus, Kenneth Bridges, and William F Owen. Anemia in hemodialysis patients: variables affecting this outcome predictor. *Journal of the American Society of Nephrology*, 8(12):1921–1929, 1997.

[3] Angelo Karaboyas, Hal Morgenstern, Sandra Waechter, Nancy L Fleischer, Raymond Vanholder, Stefan H Jacobson, Manish M Sood, Douglas E Schaubel, Masaaki Inaba, Ronald L Pisoni, and Bruce M Robinson. Low hemoglobin at hemodialysis initiation: an international study of anemia management and mortality in the early dialysis period. *Clinical Kidney Journal*, 13(3):425–433, 07 2019.

[4] William G. Murphy. The sex difference in haemoglobin levels in adults: mechanisms, causes, and consequences. *Blood Reviews*, 28(2):41–47, Mar 2014. Epub 2014 Jan 22.

[5] G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer New York, 2000.

[6] Jing Cheng, Lloyd J. Edwards, Mildred M. Maldonado-Molina, Kelli A. Komro, and Keith E. Muller. Real longitudinal data analysis for real people: Building a good enough mixed model. *Statistics in Medicine*, 29(4):504–520, 2010.

[7] Ian R. White and John B. Carlin. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29(28):2920–2931, 2010.

[8] Carsten Lundby, Jonas Juhl Thomsen, Robert Boushel, Maria Koskolou, Jørgen Warberg, José A. L. Calbet, and Paul Robach. Erythropoietin treatment elevates haemoglobin concentration by increasing red cell volume and depressing plasma volume. *The Journal of Physiology*, 578(1):309–314, 2007.

[9] Zeynep Bal, Bahar Gurlek Demirci, Suleyman Karakose, Emre Tutal, Mehtap Erkmen Uyar, Nurhan Ozdemir Acar, and Siren Sezer. Factors influencing hemoglobin variability and its association with mortality in hemodialysis patients. *The Scientific World Journal*, 2018(1):8065691, 2018.

Appendix 1 - Model building stages for linear mixed model

Step 1: Elaborated model

An elaborated LMM was fitted. The outcome variable was monthly Hgb level (g/dl). The fixed effects included all main effects of EPO dose, age, sex, iron deficiency, and month, and the interactions between dose and age, dose and sex, dose and iron, dose and month, iron and month, age and month, and sex and month. The random effects included random intercept and random slope for month, to capture the individual deviation from the average of baseline Hgb level and its evolution over time. We assumed an unstructured variance-covariance structure for the random effects (D) and simple residual covariance structure.

Step 2: Add serial correlation

Models with Gaussian serial correlation, exponential serial correlation, and without serial correlation were compared using REML log-likelihood. The model with Gaussian or Exponential serial correlation produced higher log-likelihood than the model without serial correlation (Table 12). As Gaussian and exponential serial correlation had similar log-likelihood, the model with Gaussian serial correlation was chosen.

Table 12: Comparison of model fits between models with Gaussian serial correlation, exponential serial correlation, and without serial correlation

Residual Covariance Structure	**REML log-likelihood**
Measurement error	-14000.5
Measurement error + Gaussian	-13899
Measurement error + Exponential	-13899

Step 3: Reduce random effect structure

The model with random intercept and random slope for time (Model 1) was compared with the model with only random intercept (Model 2) and model with only random slope (Model 3), using Likelihood ratio test. Model 2 was chosen as the LR test showed that the goodness of fit did not differ from that of Model 1 (Table 13.

Table 13: Comparison of models with different random components

		Model 1	**Model 2**	**Model 3**
Effect	**Parameter**	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**
Covariance of b_i				
$\text{var}(b_{1i})$	d_{11}	0.7063 (0.1015)	0.6954 (0.0555)	
$\text{var}(b_{2i})$	d_{22}	0.0090 (0.0135)	-	0.0020 (0.0065)
$\text{cov}(b_{1i}, b_{2i})$	$d_{12} = d_{21}$	0.0072 (0.0220)	-	-
Measurement error variance				
$\text{var}(\epsilon_{(1)ij})$	σ^2	0.2881 (0.0191)	0.2930 (0.0164)	0.3601 (0.0114)
Gaussian serial correlation				
$\text{var}(\epsilon_{(2)ij})$	τ^2	1.8476 (0.1732)	1.9198 (0.1317)	3.3898 (0.1275)
REML log-likelihood	-13899	-13901.1	-13929.55	
LR test p-value comparing with model 1		-	0.0833	< 0.0001

Step 4: Reduce mean structure

The mean structure was reduced using F-test, with Satterthwaite approximation for the denominator degrees of freedom. The model from step 3 was compared with the models removing the interaction effects one by one: dose*age, dose*sex, dose*iron, age*month, sex*month, iron*month, and dose*month. Removing dose*age, dose*sex, dose*iron, age*month, and sex*month did not significantly change the model fit compared to the full model (Table 14). Continuing to remove month*iron or month*dose significantly worsen the model fit, thus the procedure stopped here.

Table 14: F-test results to reduce mean structure of LMM

Reduction	DF	Den. DF	Chi-square	F-value	Pr > ChiSq	Pr>F
$dose * sex$, $dose * age$ and $dose * iron$	3	6062	3.61	1.20	0.3066	0.3067
$month * age$ and $month * sex$	5	6062	9.04	1.81	0.1075	0.1077
$month * iron$	6	6062	14.73	2.45	0.0225	0.0226
$month * dose$	6	6062	31.35	5.23	<.0001	<.0001

Appendix 2 - R/SAS code

Exploratory Data Analysis

```
1
2  /* ----------- Individual profile ------------ */
3
4  /* Select 50 random patients */
5
6  proc surveyselect data=m.hemodialysis out=sampled_patients
7     method=srs n=50 seed=1003; /*simple random sampling */
8     id ID;
9  run;
10
11 proc sql;
12    create table subset_50_hemo as
13    select a.*
14    from m.hemodialysis as a
15    inner join sampled_patients as b
16    on a.ID = b.ID;
17 quit;
18
19 /*proc print data=subset_50_hemo;
20    title "Data from Random Sample of 50 Patients";
21 run; */
22
23 /* Plot Individual profiles of 50 random patients */
24 proc sort data=subset_50_hemo;
25 by ID month;
26 run;
27
28 proc gplot data=subset_50_hemo;
29 plot hb*month=id / haxis=axis1 vaxis=axis2 nolegend;
30 axis1 label=(h=2 'Time in month') value=(h=1.5)
31 order=(0 to 6 by 1) minor=none;
32 axis2 label=(h=2 A=90 'Hemoglobin level (g/dl)') value=(h=1.5)
33 order=(4 to 18 by 2) minor=none;
34 title h=3 'Individual profiles for 50 random subjects';
35 run;quit;
36
37
38 /* ----------Mean structure ----------*/
39
40 proc gplot data=m.hemodialysis;
41 plot hb*month / haxis=axis1 vaxis=axis2;
42 symbol c=brown i=std1mjt w=2 mode=include;
43 axis1 label=(h=2 'Time (months)') value=(h=1.5) order=(1 to 6 by 1) minor=none;
44 axis2 label=(h=2 A=90 'Hemoglobin level (g/dl)') value=(h=1.5) order=(9 to 14 by 0.5)
45 minor=none;
46 title h=3 'Average evolution of Hb level, with standard errors of means';
47 run;quit;
48
49
50 /* -------- Variance structure --------- */
51
52 /* Calculate monthly mean Hb */
53 proc means data=m.hemodialysis noprint;
54    class month;
55    var hb;
56    output out=mean_hb mean=mean_hb;
57 run;
58
59 /* Calculate variance by month */
60
61 proc sort data=m.hemodialysis;
62    by month;
63 run;
64
65 proc sort data=mean_hb;
```

```
   by month;
run;

data variance_calc;
   merge m.hemodialysis (in=a) mean_hb (in=b);
   by month;
   if a and b; /*Keeps only records where month is present in both datasets */
   /* Calculate squared deviation for each observation */
   sq_residual = (hb - mean_hb)**2;
run;


/* Plot var(t) and its standard error by month */

ods graphics / reset=all outputfmt=png imagename="var_structure" imagefmt=png;
ods listing gpath="/home/u62794741/LDA/";
goptions reset=all ftext=swiss device=psepsf gsfname=fig4 gsfmode=replace
rotate=landscape;

proc gplot data=variance_calc;
plot sq_residual*month / haxis=axis1 vaxis=axis2;
symbol c=brown i=std1mjt w=2 mode=include;
axis1 label=(h=2 'Time (months)') value=(h=1.5) order=(1 to 6 by 1) minor=none;
axis2 label=(h=2 A=90 'Squared residuals') value=(h=1.5) order=(0 to 4 by 0.5)
minor=none;
title h=3 'Variance structure of Hb values over time';
run;quit;
ods graphics off;

/* ---- Correlation structure --- */

proc sort data=m.hemodialysis;
   by ID month;
run;

proc transpose data=m.hemodialysis out=transposed_hemo prefix=hb_;
   by ID;
   id month;
   var hb;
run;

data transposed_hemo;
   set transposed_hemo;
   label hb_1 = "Month 1"
         hb_2 = "Month 2"
         hb_3 = "Month 3"
         hb_4 = "Month 4"
         hb_5 = "Month 5"
         hb_6 = "Month 6";
run;

proc sgscatter data=transposed_hemo;
   matrix hb_1 hb_2 hb_3 hb_4 hb_5 hb_6 / diagonal=(histogram);
   title "Scatter Plot Matrix of monthly Hb values";
run;


# test linear trend ------------------------------------------------------

## fit linear model for each patient ----------------------------------------
# Initialize lists to store statistics for each patient
r2 <- list()
SSE <- list()
SSR <- list()
ni <- list()
intercept <- list()
slope <- list()
SEintercept <- list()
```

```r
SEslope <- list()

# Loop through each unique patient ID
for (i in unique(hemo$ID)) {
  # Subset data for the current patient
  patient_data <- hemo[hemo$ID == i, ]

  # Fit the linear model for the current patient
  Model1 <- lm(hb ~ month, data = patient_data, x = TRUE)

  # Calculate and store the statistics
  r2[[i]] <- summary(Model1)$r.squared
  SSE[[i]] <- sum(Model1$residuals^2)
  SSR[[i]] <- anova(Model1)$`Sum Sq`[1]
  ni[[i]] <- length(Model1$x[, 1])
  intercept[[i]] <- Model1$coefficients[1]
  slope[[i]] <- Model1$coefficients[2]
  SEintercept[[i]] <- summary(Model1)$coefficients[1, 2]
  SEslope[[i]] <- summary(Model1)$coefficients[2, 2]
}

# Calculate the meta R-squared statistic
R2_meta <- sum(unlist(SSR)) / sum(unlist(SSR) + unlist(SSE))

# Display the meta R-squared result
R2_meta
# 0.5160457

# test quadratic trend ----------------------------------------------------
# Initialize lists to store statistics for each patient
r2 <- list()
SSE <- list()
SSR <- list()
ni <- list()
intercept <- list()
slope <- list()
SEintercept <- list()
SEslope <- list()

# Loop through each unique patient ID
for (i in unique(hemo$ID)) {
  # Subset data for the current patient
  patient_data <- hemo[hemo$ID == i, ]

  # Fit the quadratic model (including month squared term) for the current patient
  Model1 <- lm(hb ~ month + I(month^2), data = patient_data, x = TRUE)

  # Calculate and store the statistics
  r2[[i]] <- summary(Model1)$r.squared
  SSE[[i]] <- sum(Model1$residuals^2)
  SSR[[i]] <- anova(Model1)$`Sum Sq`[1]
  ni[[i]] <- length(Model1$x[, 1])
  intercept[[i]] <- Model1$coefficients[1]
  slope[[i]] <- Model1$coefficients[2]
  SEintercept[[i]] <- summary(Model1)$coefficients[1, 2]
  SEslope[[i]] <- summary(Model1)$coefficients[2, 2]
}

# Calculate the meta R-squared statistic
R2_meta <- sum(unlist(SSR)) / sum(unlist(SSR) + unlist(SSE))

# Display the meta R-squared result
R2_meta
# 0.7270484

# test for model extension ------------------------------------------------

# Initialize lists to store statistics for each patient
```

```r
# Initialize lists to store statistics for each patient
SSE_R <- list() # SSE for the linear model
SSE_F <- list() # SSE for the extended (quadratic) model
p <- 2 # Number of parameters in the original (linear) model
p_ext <- 1 # Number of parameters added in the extended (quadratic) model (for month^2 term)
ni <- list() # Number of observations for each patient

# Loop through each unique patient ID
for (i in unique(hemo$ID)) {
  # Subset data for the current patient
  patient_data <- hemo[hemo$ID == i, ]

  # Fit the linear model (original model)
  Model_R <- lm(hb ~ month, data = patient_data, x = TRUE)

  # Fit the quadratic model (extended model)
  Model_F <- lm(hb ~ month + I(month^2), data = patient_data, x = TRUE)

  # Calculate SSE for both models
  SSE_R[[i]] <- sum(Model_R$residuals^2)
  SSE_F[[i]] <- sum(Model_F$residuals^2)

  # Number of observations for the current patient
  ni[[i]] <- length(Model_R$fitted.values)
}

# Calculate the individual F-statistics for each patient
F_individual <- list()
for (i in unique(hemo$ID)) {
  SSE_R_i <- SSE_R[[i]]
  SSE_F_i <- SSE_F[[i]]
  n_i <- ni[[i]]

  # Calculate the F-statistic for this patient
  F_individual[[i]] <- ((SSE_R_i - SSE_F_i) / p_ext) / (SSE_F_i / (n_i - p - p_ext))
}

# Calculate the meta F-statistic
F_meta_numerator <- sum(unlist(SSE_R) - unlist(SSE_F)) / p_ext
F_meta_denominator <- sum(unlist(SSE_F)) / (sum(unlist(ni)) - p - p_ext)

F_meta <- F_meta_numerator / F_meta_denominator

# You may also want to calculate the degrees of freedom for the null distribution
df1 <- p_ext # The number of parameters added in the extended model
df2 <- sum(unlist(ni)) - p - p_ext # The remaining degrees of freedom

# Calculate the p-value from the F-distribution
p_value <- 1 - pf(F_meta, df1 = df1, df2 = df2)
```

Summary statistics

```
/*Male and female Hgb*/
proc format;
    value sexfmt
        1 = 'Male'
        2 = 'Female';
run;
proc sgplot data=mylib.hemodialysis;
    where SEX is not missing;
    vbox Hb / category=MONTH group=SEX boxwidth=0.4;
    xaxis label="Month";
    yaxis label="Hgb Level (g/dl)";
    keylegend / title="Gender" location=inside position=topright across=1;
    format SEX sexfmt.;
run;
/*Increment from baseline to last observed*/
proc sort data=mylib.hemodialysis;
```

```
    by ID MONTH;
run;
data increments;
    set mylib.hemodialysis;
    by ID;

    retain Hb_baseline;
    if first.ID then Hb_baseline = Hb; /*baseline as the first Hb value*/

    if last.ID then do;
        Hb_increment = Hb - Hb_baseline; /*calculate increment from baseline to last observed*/
        output;
    end;
run;
proc sgplot data=increments;
    histogram Hb_increment / nbins=20;
    xaxis label="Hgb Increment (Last Observed - Baseline)" min=-4 max=4;
    yaxis label="Frequency";
run;
/*Calculate AUCi */
proc sort data=mylib.hemodialysis;
    by ID MONTH;
run;
data AUCi_summary;
    set mylib.hemodialysis;
    by ID;

    retain AUCi 0;

    if not first.ID then do;
        AUCi + (Hb + lag(Hb)) / 2 * (MONTH - lag(MONTH));
    end;
    if last.ID then output;
    keep ID AUCi;
run;
proc sort data=mylib.hemodialysis;
    by ID;
run;
data AUCi_with_covariates;
    merge AUCi_summary(in=a) mylib.hemodialysis(keep=ID age SEX);
    by ID;
    if a;
run;
proc glm data=AUCi_with_covariates;
    class SEX;
    model AUCi = age SEX / solution;
    title "Modeling AUC with Age and Sex";
run;
```

Variable transformation

```
    /* Create new DOSE2 variable with 1 month lag */
proc sort data=m.hemodialysis;
by ID month;
run;

data m.hemodialysis2;
   set m.hemodialysis;
   by ID month;
   retain prev_DOSE;
   if first.ID then DOSE2 = 0;
   else DOSE2 = prev_DOSE;
   prev_DOSE = DOSE;

run;

/* reparameterize month */
/* Dummy variables for sex */
data m.hemodialysis2;
```

```
   set m.hemodialysis2;
   month_class = month;
   month0 = month - 1;
   month0_class = month0;
   SEX_male = (SEX = 1);
run;
```

Multivariate model

```
/*Multivariate model full, quadratic term check*/
proc mixed data=mylib.hemodialysis2 method=reml;
   class ID month0_class;
   model hb = DOSE2 AGE SEX_male iron month0 month0*month0
              DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0
              iron*month0 AGE*month0 SEX_male*month0 / solution;
   repeated month0_class / type=un subject=ID r rcorr;
   title "Fixed-effects model with repeated measures for Hb";
   contrast 'Reduce interaction of quadratic term' month0*month0 1, /chisq;
run;

/*Reduce dose interact with iron, age, sex*/
proc mixed data=mylib.hemodialysis2 method=reml;
       class ID month0_class;
       model hb = DOSE2 AGE SEX_male iron month0
       DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0
       iron*month0 AGE*month0 SEX_male*month0 / solution;
       repeated month0_class / type=un subject=ID r rcorr;
       title "Fixed-effects model with repeated measures for Hb";
       contrast 'Reduce interaction of DOSE' DOSE2*iron 1,
                                                          DOSE2*AGE 1,
                                                          DOSE2*SEX_male 1, /chisq;
run;
/*Reduce dose interact with iron, age, sex, and age*month, sex*month*/
proc mixed data=mylib.hemodialysis2 method=reml;
       class ID month0_class;
       model hb = DOSE2 AGE SEX_male iron month0
       DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0
       iron*month0 AGE*month0 SEX_male*month0 / solution;
       repeated month0_class / type=un subject=ID r rcorr;
       title "Fixed-effects model with repeated measures for Hb";
       contrast 'Further reduce interaction of Dose & month' DOSE2*iron 1,
                                                          DOSE2*AGE 1,
                                                          DOSE2*SEX_male 1,
                                                          AGE*month0 1,
                                                          SEX_male*month0 1,/chisq;
run;
/*Further reduce iron*month*/
proc mixed data=mylib.hemodialysis2 method=reml;
       class ID month0_class;
       model hb = DOSE2 AGE SEX_male iron month0
       DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0
       iron*month0 AGE*month0 SEX_male*month0 / solution;
       repeated month0_class / type=un subject=ID r rcorr;
       title "Fixed-effects model with repeated measures for Hb";
       contrast 'Further reduce interaction of iron & month' DOSE2*iron 1,
                                                          DOSE2*AGE 1,
                                                          DOSE2*SEX_male 1,
                                                          AGE*month0 1,
                                                          SEX_male*month0 1,
                                                          iron*month0 1,/chisq;
run;
/*Or further reduce DOSE2*month*/
proc mixed data=mylib.hemodialysis2 method=reml;
       class ID month0_class;
       model hb = DOSE2 AGE SEX_male iron month0
       DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0
       iron*month0 AGE*month0 SEX_male*month0 / solution;
       repeated month0_class / type=un subject=ID r rcorr;
       title "Fixed-effects model with repeated measures for Hb";
```

```
        contrast 'Further reduce interaction of dose2 & month' DOSE2*iron 1,
                                                               DOSE2*AGE 1,
                                                               DOSE2*SEX_male 1,
                                                               AGE*month0 1,
                                                               SEX_male*month0 1,
                                                               DOSE2*month0 1,/chisq;
run;
/*The parsimonious model, type un*/
proc mixed data=mylib.hemodialysis2 method=reml;
        class ID month0_class;
        model hb = DOSE2 AGE SEX_male iron month0
    DOSE2*month0 iron*month0 / solution;
        repeated month0_class / type=un subject=ID r rcorr;
        title "The most parsinomious model type un";
run;
/*Find covariance structure*/
/* type=un */
proc mixed data=mylib.hemodialysis2 method=reml;
   class ID month0_class;
   model hb = DOSE2 AGE SEX_male iron month0 DOSE2*month0 iron*month0 / solution;
   repeated month0_class / type=un subject=ID r rcorr;
   ods output FitStatistics=fit_un;
run;
/* type=simple */
proc mixed data=mylib.hemodialysis2 method=reml;
   class ID month0_class;
   model hb = DOSE2 AGE SEX_male iron month0 DOSE2*month0 iron*month0 / solution;
   repeated month0_class / type=simple subject=ID r rcorr;
   ods output FitStatistics=fit_simple;
run;
data LRT_1;
   merge fit_simple(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_simple))
         fit_un(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_un));
   LRT_1 = LL_simple - LL_un; /* Test statistic */
   df = '1:21'; /* Degrees of freedom: difference in the number of parameters */
   p_value = 0.5*(1 - probchi(LRT_1, 1) + 1 - probchi(LRT_1, 21)); /* p-value for chi-square test
       */
run;
proc print data=LRT_1;
   var LRT_1 df p_value;
   title "LR test: un vs simple";
run;
/* type=cs */
proc mixed data=mylib.hemodialysis2 method=reml;
   class ID month0_class;
   model hb = DOSE2 AGE SEX_male iron month0 DOSE2*month0 iron*month0 / solution;
   repeated month0_class / type=cs subject=ID r rcorr;
   ods output FitStatistics=fit_cs;
run;
data LRT_2;
   merge fit_cs(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_cs))
         fit_un(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_un));
   LRT_2 = LL_cs - LL_un; /* Test statistic */
   df = '2:21'; /* Degrees of freedom: difference in the number of parameters */
   p_value = 0.5*(1 - probchi(LRT_2, 2) + 1 - probchi(LRT_2, 21)); /* p-value for chi-square test
       */
run;
proc print data=LRT_2;
   var LRT_2 df p_value;
   title "LR test: un vs cs";
run;
/* type=banded */
proc mixed data=mylib.hemodialysis2 method=reml;
   class ID month0_class;
   model hb = DOSE2 AGE SEX_male iron month0 DOSE2*month0 iron*month0 / solution;
   repeated month0_class / type=un(2) subject=ID r rcorr;
   ods output FitStatistics=fit_banded;
run;
```

```
data LRT_3;
   merge fit_banded(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_banded))
         fit_un(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_un));
   LRT_3 = LL_banded - LL_un; /* Test statistic */
   df = '21:21'; /* Degrees of freedom: difference in the number of parameters */
   p_value = 0.5*(1 - probchi(LRT_3, 21) + 1 - probchi(LRT_3, 21)); /* p-value for chi-square
       test */
run;
proc print data=LRT_3;
   var LRT_3 df p_value;
   title "LR test: un vs banded";
run;
/* type=autoregressive */
proc mixed data=mylib.hemodialysis2 method=reml;
   class ID month0_class;
   model hb = DOSE2 AGE SEX_male iron month0 DOSE2*month0 iron*month0 / solution;
   repeated month0_class / type=ar(1) subject=ID r rcorr;
   ods output FitStatistics=fit_ar;
run;
data LRT_4;
   merge fit_ar(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_ar))
         fit_un(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_un));
   LRT_4 = LL_ar - LL_un; /* Test statistic */
   df = '2:21'; /* Degrees of freedom: difference in the number of parameters */
   p_value = 0.5*(1 - probchi(LRT_4, 2) + 1 - probchi(LRT_4, 21)); /* p-value for chi-square test
       */
run;
proc print data=LRT_4;
   var LRT_4 df p_value;
   title "LR test: un vs autoregressive";
run;
/* type=toep*/
proc mixed data=mylib.hemodialysis2 method=reml;
   class ID month0_class;
   model hb = DOSE2 AGE SEX_male iron month0 DOSE2*month0 iron*month0 / solution;
   repeated month0_class / type=toep subject=ID r rcorr;
   ods output FitStatistics=fit_toep;
run;

data LRT_5;
   merge fit_toep(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_toep))
         fit_un(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_un));
   LRT_5 = LL_toep - LL_un; /* Test statistic */
   df = '6:21'; /* Degrees of freedom: difference in the number of parameters */
   p_value = 0.5*(1 - probchi(LRT_5, 6) + 1 - probchi(LRT_5, 21)); /* p-value for chi-square test
       */
run;
proc print data=LRT_5;
   var LRT_5 df p_value;
   title "LR test: un vs toeplitz";
run;
```

Two-stage analysis

```
/*The 2-stage model formulation- MAIN CODE*/
/* ---------- Data manipulation ---------------- */
/* Create new DOSE2 variable with 1 month lag */

proc sort data=mylib.hemodialysis;
by ID month;
run;

data mylib.hemodialysis2;
   set mylib.hemodialysis;
   by ID month;

   /* Retain the previous month's DOSE for each ID */
   retain prev_DOSE;

```

```
   /* Initialize DOSE2 */
   if first.ID then DOSE2 = 0;
   else DOSE2 = prev_DOSE;

   /* Update prev_DOSE with the current DOSE for the next observation */
   prev_DOSE = DOSE;

run;

/* reparameterize month */
/* Dummy variables for sex */
data mylib.hemodialysis2;
   set mylib.hemodialysis2;
   month_class = month;
   month0 = month - 1;
   month0_class = month0;
run;


/* Step 1: Count the number of valid observations per ID */
proc sql;
   create table valid_ids as
   select ID
   from mylib.hemodialysis2
   where Hb is not missing and MONTH0 is not missing
   group by ID
   having count(*) >= 2; /*Maybe it is best to remove the equal one because you can not see trend
       on 2 points*/
quit;

/* Step 2: Keep only records for valid IDs -2023 subjects*/
proc sort data=valid_ids;
   by ID;
run;

proc sort data=mylib.hemodialysis2 out=hemo_valid;
   by ID;
run;

data hemo_valid;
   merge hemo_valid(in=a) valid_ids(in=b);
   by ID;
   if a and b;
run;

/*Step 3: Sort Data by ID*/
proc sort data=hemo_valid;
   by ID;
run;

/* Step 4: Fit individual linear models */
proc datasets library=work nolist;
   delete individual_params;
quit;

ods output ParameterEstimates=individual_params;

proc reg data=hemo_valid plots=none;
   by ID;
   model Hb = MONTH0;
run;
quit;

/*Step 5: Process the Parameter Estimates*/
/* Step 5: Reshape the dataset to have one record per patient */
proc transpose data=individual_params out=patient_params(drop=_name_);
   by ID;
   id Variable;
```

```
   var Estimate;
run;
/* Step 1: Sort the dataset by ID */
proc sort data=patient_params out=sorted_data;
    by ID;
run;
/* Step 6: Extract unique ID and AGE from the original dataset */
proc sort data=hemo_valid(keep=ID AGE) nodupkey out=patient_age;
   by ID;
run;

/* Merge with patient_params */
data stage2_data;
   merge patient_params(in=a) patient_age(in=b);
   by ID;
   if a and b;
run;
/* Calculating the mean of Intercept and month */
proc means data=work.stage2_data;
   var Intercept month0;
run;
/*Step 7: Fit Regression Models in Stage 2*/
/*Modelling intercept*/
proc reg data=stage2_data;
   model Intercept = AGE;
   title 'Effect of Age on Intercepts';
run;
quit;

/*Modelling slopes*/
proc reg data=stage2_data;
   model MONTH0 = AGE;
   title 'Effect of Age on Slopes';
run;
quit;
```

Linear mixed model

```
/* --- Fit elaborated mixed model with random intercept and random slope ---*/
proc mixed data=m.hemodialysis2 method=reml PLOTS(MAXPOINTS= 10000)=all;
       class ID month0_class;
       model hb = DOSE2 AGE SEX_male iron month0
       DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0 iron*month0 AGE*month0 SEX_male*month0 /
           solution;
       random intercept month0 / type=un subject=ID g gcorr v vcorr;
       repeated month0_class / type=simple subject=ID r rcorr;
       ods graphics on/discretemax=3500;
       title "Elaborated mixed model with month0 & dummy SEX";
run;
quit;

/* -------- Model with serial correlation ---------------*/

/* Exponential */

proc mixed data=m.hemodialysis2 method=reml;
       class ID month0_class;
       model hb = DOSE2 AGE SEX_male iron month0
       DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0 iron*month0 AGE*month0 SEX_male*month0 /
           solution;
       random intercept month0 / type=un subject=ID g gcorr v vcorr;
       repeated month0_class / type=sp(gau)(month0) local subject=ID r rcorr;
       title "Elaborated mixed model with Exponential serial correlation";
run;
quit;

/* Gaussian */
ods select FitStatistics CovParms SolutionF SolutionR;
proc mixed data=m.hemodialysis2 method=reml covtest;
```

```
        class ID month0_class;
        model hb = DOSE2 AGE SEX_male iron month0
        DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0 iron*month0 AGE*month0 SEX_male*month0 /
            solution;
        random intercept month0 / type=un subject=ID g gcorr v vcorr;
        repeated month0_class / type=sp(gau)(month0) local subject=ID r rcorr;
        title "Elaborated mixed model with Gaussian serial correlation";
        ods output FitStatistics=FitFull;
run;
quit;

/* -------- Reduce random effect ---------------*/

/* Only intercept */
ods exclude SolutionR ClassLevels;
proc mixed data=m.hemodialysis2 method=reml covtest;
        class ID month0_class;
        model hb = DOSE2 AGE SEX_male iron month0
        DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0 iron*month0 AGE*month0 SEX_male*month0 /
            solution;
        random intercept / type=un subject=ID g gcorr v vcorr;
        repeated month0_class / type=sp(gau)(month0) local subject=ID r rcorr;
        title "Random Intercept-only mixed model";
        ods output FitStatistics=FitReduced;
run;
quit;

/* LR test random slope + intercept vs random intercept */
data LRT_Results;
   merge FitFull(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_full))
         FitReduced(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_reduced));
   LRT_stat = LL_reduced - LL_full; /* Test statistic */
   df = '1:2'; /* Degrees of freedom: difference in the number of parameters */
   p_value = 0.5*(1 - probchi(LRT_stat, 1) + 1 - probchi(LRT_stat, 2)); /* p-value for chi-square
       test */
run;

proc print data=LRT_Results;
   var LRT_stat df p_value;
   title "LR test: full model vs intercept only";
run;

/* Only slope */
ods exclude SolutionR ClassLevels;
proc mixed data=m.hemodialysis2 method=reml covtest;
        class ID month0_class;
        model hb = DOSE2 AGE SEX_male iron month0
        DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0 iron*month0 AGE*month0 SEX_male*month0 /
            solution;
        random month0 / type=un subject=ID g gcorr v vcorr;
        repeated month0_class / type=sp(gau)(month0) local subject=ID r rcorr;
        title "Random slope-only mixed model";
        ods output FitStatistics=FitReduced2;
run;
quit;

/* LR test random slope + intercept vs random slope */
data LRT_Results2;
   merge FitFull(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_full))
         FitReduced2(where=(Descr="-2 Res Log Likelihood") rename=(Value=LL_reduced));
   LRT_stat = LL_reduced - LL_full; /* Test statistic */
   df = '1:2'; /* Degrees of freedom: difference in the number of parameters */
   p_value = 0.5*(1 - probchi(LRT_stat, 1) + 1 - probchi(LRT_stat, 2)); /* p-value for chi-square
       test */
run;

proc print data=LRT_Results2;
   var LRT_stat df p_value;
```

```
93    title "LR test: full model vs slope only";
94 run;
95
96 /* Model with random intercept only was not significantly different from full model
97 -> Keep random intercept only */
98
99
100 /* -------- Reduce fixed effect ---------------*/
101 ods exclude SolutionR ClassLevels;
102 proc mixed data=m.hemodialysis2 method=reml;
103        class ID month0_class;
104        model hb = DOSE2 AGE SEX_male iron month0
105        DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0 iron*month0 AGE*month0 SEX_male*month0 /
               solution;
106        random intercept / type=un subject=ID g gcorr v vcorr;
107        repeated month0_class / type=sp(gau)(month0) local subject=ID r rcorr;
108        title "Random intercept model";
109        contrast 'Reduce interaction of DOSE' DOSE2*iron 1,
110                                                                    DOSE2*AGE 1,
111                                                                    DOSE2*SEX_male 1, /chisq;
112 run;
113 quit;
114
115 ods exclude SolutionR ClassLevels;
116
117 proc mixed data=m.hemodialysis2 method=reml;
118        class ID month0_class;
119        model hb = DOSE2 AGE SEX_male iron month0
120        DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0 iron*month0 AGE*month0 SEX_male*month0 /
               solution;
121        random intercept / type=un subject=ID g gcorr v vcorr;
122        repeated month0_class / type=sp(gau)(month0) local subject=ID r rcorr;
123        title "Random intercept model";
124        contrast 'Reduce intera Dose & month' DOSE2*iron 1,
125                                                                    DOSE2*AGE 1,
126                                                                    DOSE2*SEX_male 1,
127                                                                    AGE*month0 1,
128                                                                    SEX_male*month0 1,/chisq;
129 run;
130 quit;
131 ods exclude SolutionR ClassLevels;
132 proc mixed data=m.hemodialysis2 method=reml;
133        class ID month0_class;
134        model hb = DOSE2 AGE SEX_male iron month0
135        DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0 iron*month0 AGE*month0 SEX_male*month0 /
               solution;
136        random intercept / type=un subject=ID g gcorr v vcorr;
137        repeated month0_class / type=sp(gau)(month0) local subject=ID r rcorr;
138        title "Random intercept model";
139        contrast 'Reduce intera Dose & month' DOSE2*iron 1,
140                                                                    DOSE2*AGE 1,
141                                                                    DOSE2*SEX_male 1,
142                                                                    AGE*month0 1,
143                                                                    SEX_male*month0 1,
144                                                                    iron*month0 1,/chisq;
145 run;
146 quit;
147 ods exclude SolutionR ClassLevels;
148 proc mixed data=m.hemodialysis2 method=reml;
149        class ID month0_class;
150        model hb = DOSE2 AGE SEX_male iron month0
151        DOSE2*AGE DOSE2*SEX_male DOSE2*iron DOSE2*month0 iron*month0 AGE*month0 SEX_male*month0 /
               solution;
152        random intercept / type=un subject=ID g gcorr v vcorr;
153        repeated month0_class / type=sp(gau)(month0) local subject=ID r rcorr;
154        title "Random intercept model";
155        contrast 'Reduce intera Dose & month' DOSE2*iron 1,
156                                                                    DOSE2*AGE 1,
```

```
                                                                    DOSE2*SEX_male 1,
                                                                    AGE*month0 1,
                                                                    SEX_male*month0 1,
                                                                    DOSE2*month0 1,/chisq;
run;
quit;


/* -------- Final model ----------------*/
ods exclude SolutionR ClassLevels;
proc mixed data=m.hemodialysis2 method=reml PLOTS(MAXPOINTS= 10000)=all covtest;
        class ID month0_class;
        model hb = DOSE2 AGE SEX_male iron month0 DOSE2*month0 iron*month0 / solution;
        random intercept / type=un subject=ID solution g gcorr v vcorr;
        repeated month0_class / type=sp(gau)(month0) local subject=ID r rcorr;
        ods graphics on/discretemax=3500;
        ods output SolutionR=RandomIntercepts; /* Save random intercepts to a dataset */
        title "Final Random intercept model";
run;
ods select all;
quit;

/* -------- Model with only age & time effect, to compare with 2-stage --------*/
ods exclude SolutionR;
proc mixed data=m.hemodialysis2 method=reml;
        class ID month0_class;
        model hb = AGE month0 AGE*month0 / solution;
        random intercept month0/ type=un subject=ID solution g gcorr v vcorr;
        repeated month0_class / type=sp(gau)(month0) local subject=ID r rcorr;
        title "Model with only age and month to compare with 2-stage";
        ods output SolutionR=RandomEffects; /* Save random intercepts to a dataset */

run;
ods select all;
quit;

/* Filter random intercept and slope values */
data RandomInterceptSlope;
    set RandomEffects;
    where Effect in ('Intercept', 'month0'); /* Select intercept and slope */
    if Effect = 'Intercept' then RandomIntercept = Estimate;
    else if Effect = 'month0' then RandomSlope = Estimate;
    retain RandomIntercept RandomSlope; /* Keep values */
    by ID;
    if last.ID; /* Output only the last record per ID containing both values */
run;

/* Create scatter plot */

%let FixedIntercept = 10.6306;
%let FixedSlope = 0.1349;

data SubjectInterceptSlope;
    set RandomInterceptSlope;
    SubjectIntercept = &FixedIntercept + RandomIntercept;
    SubjectSlope = &FixedSlope + RandomSlope;
run;
ods graphics / reset=all outputfmt=png imagename="Age_subject_intercept_slope" imagefmt=png;
ods listing gpath="/home/u62794741/LDA/";
proc sgplot data=SubjectInterceptSlope;
    scatter x=SubjectIntercept y=SubjectSlope;
    xaxis label="Subject-specific Intercept";
    yaxis label="Subject-specific Slope for month";
    title "Linear mixed model";
run;
ods graphics off;
```