Longitudinal Data Analysis
Assignment I

An investigation to study the effect of time as well as patient characteristics on the evolution of the abdominal aorta diameter in patients with abdominal aortic aneurysm (AAA).

Student:
Andrew Kamya (1849786)

Supervisors:
Prof. Dr. Geert Molenberghs
Prof. Dr. Geert Verbeke

Abstract

This is a longitudinal cohort study to investigate the effects of time as well as patient characteristics on the evolution of the diameter of the abdominal aorta for patients with abdominal aortic aneurysm (AAA). The patient characteristics as well as the diameter (in mm) of the 101 patients were collected every six months after the date of enrollment. There was a massive percentage of dropouts that was observed after Month 6 of follow up. Apart from obtaining summary statistics, individual profile plots were used to gain a general knowledge of what is expected. Moreover two-stage model, multivariate regression model and a linear mixed model were fitted to answer the research objectives. The results indicate that the age of the patient and the interaction of smoking status with time has a significant influence on the evolution of the diameter of the abdominal aorta.

Keywords : *Abdominal Aortic Aneurysm (AAA), Covariance structure, Multivariate, Two stage model, Linear mixed model (LMM)*

Contents

1 Introduction

An abdominal aortic aneurysm (AAA) is an enlargement of the abdominal aorta, which is the main artery supplying the blood to the body. AAA is mainly caused by atherosclerosis, which occurs as a result of hardening of the arteries. The inflammation or degeneration of the aortic walls caused by high blood pressure or an infection of the aorta also results in AAA. The abdominal aortic aneurysm commonly affects individuals who are 65 years and older. The high occurrence has been witnessed among smokers and mostly in males. (Voop, 2005).

The symptomatic aneurysm has a high risk of rupture and can be life-threatening as large amounts of blood spill into the abdominal cavity, which is an indication for surgery. The mortality of AAA rupture is up to 90%. 65 to 75% of patients die before they arrive at the hospital, and up to 90% die before they reach the operating room. Therefore symptomatic and large aneurysms are considered for repair by surgical methods. Often, an intervention is required if the aneurysm grows more than 1cm per year or when it is more significant than 5.5cm (Aggarwal, 2011).

Although several clinical studies have been conducted, few reliable data are available on the growth trend on the enlargement of the abdominal aorta. Therefore, the purpose why this study was done was to study the evolution of the artery diameter on patients with AAA as well as the relation with some patient characteristics.

This report is structured in 6 sections. In this section, the background information and study objective were defined. Section 2 explains more about the data. Section 3 discusses methodology used in data exploration and inference. In section 4 and 5 the interpretation of the results, discussion and recommendation were presented.

2 Data Description

The data set contained observations of 101 AAA patients aged between 52 and 86 years old. These patients were followed up and data were obtained after every six months. During these follow-up visits, the diameter of the artery and several patient characteristics were collected. In other words, the design of the study that produces the dataset is balanced but the resulting dataset is unbalanced due to missingness. The variables under the study were patient id, age of patient at baseline, coronary disease history, diameter of the abdominal aorta (in mm), smoking status, body mass index (BMI), length (in cm) and weight (in kg).

3 Methods

3.1 Exploratory analysis

Before any further investigation, an exploratory analysis was conducted on the data to discover patterns, data structure and spot any existing anomalies and to obtain clear insight on conceivable implication before model building and hypothesis testing. In this study, individual-specific profiles were used to explore the between as well as within variability. 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.

3.2 Summary Statistics

Sometimes it is necessary to perform the analysis in a situation where the correlation between two repeated measurements decreases as the period between the measurements increases. A paired t-test is appropriate in accounting for this matter by considering subject-specific differences. Alternatively, other methods have been devised to transform the number of measurements for the subjects from n_i to 1. In this study, the attention will be using Area Under the Curve (AUC) and Analysis of increments. These two methods despite having no problems of multiple testing, they also do

not explicitly assume balanced data. Nevertheless, these methods use partial information for the analysis and should be used with caution (Verbeke and Molenberghs, 2019).

3.3 Multivariate Regression Model

The multivariate regression model can be seen as an advanced multiple linear regression model where the repeated measurements of the response, as well as the repeated measurements for the predictor variables, are correlated (Izenman, 2013). This method can suitably be used in a situation where there are completely balanced/unbalanced measurements for the subjects at the fixed time points (Verbeke and Molenberghs, 2005 and 2019). The general representation of this model assumes a matrix form;

$$Y_i = X_i\beta + \epsilon_i \tag{1}$$

Where $Y_i = (Y_{i1}, Y_{i2}, ...Y_{in_i})'$ is n_i-dimensional vector of available repeated measurements of ith subject, X_i is a matrix of covariates, β is a vector of the regression parameters while ϵ_i is a vector of error components that are assumed to follow a normal distribution with expectation zero and variance Sigma. The error components capture variability that has not been explained by the systematic components. The inference for this model follows the standard classical maximum likelihood theory (Molernbeghs and Verbeke, 2005).

3.4 Two-Stage Model Formulation

Often in practice realising a balanced data is challenging since the measurements taken at fixed time point suffer from missingness. For this reason, sometimes it is necessary to use subject-specific profiles which could be approximated by linear regression function and therefore could consider using a 2-stage model which involves fitting a model for each subject separately and then find a way to explain the within and between variability that exists in the subject-specific regression coefficients (Verbeke and Molenberghs, 2000/2019).

3.4.1 Stage 1

Usually, in the first stage, a response Y_i which is the n_i-dimensional vector for all repeated measurements for the ith subject ($Y_i = (Y_{i1}, Y_{i2}, ...Y_{in_i})'$) is assumed to satisfy a linear regression model

$$\boldsymbol{Y_i = Z_i\beta_i + \epsilon_i} \tag{2}$$

Where $i = 1, ..., N$, Z_i is an (n_i x q) matrix of known covariates modelling the evolution of the response over time t_{ij} for the ith subject. β_i is a q-dimensional unknown subject-specific regression coefficient and ϵ_i a vector of residual components ϵ_{ij} $j = 1, .., n_i$. $\epsilon_i \sim N(0, \sigma^2 I_n)$ where I_n is the n_i-dimensional identity matrix (Verbeke and Molenberghs, 2000). Furthermore, this model focuses on describing only the observed variability within the subjects.

3.4.2 Stage 2

In the second stage, the multivariate regression model in 3 is used to describe the observed variability

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

between the subjects by relating the subject specific regression coefficients β_i to the known covariates K_i which is a (q x q) matrix of known covariates, β is a p-dimensional vector of unknown regression parameters while b_i is independent and $N(0, D)$.

Although a two-stage modelling techniques gives results, much caution should be taken in regarding the use of partial information. According to (Verbeke and Molenberghs, 2000), this method summarizes the vector Y_{ij} of the observed measurement for the ith subject by $\hat{\beta}_i$ leading to loss of information. Moreover, using estimates of β instead of β introduces additional random variability. Besides, the approximated covariance matrix of $\hat{\beta}$ estimate highly depends on the number of measurements available for the ith subject as well as the time points when these measurements were taken, and a two-stage model can not account for this. The Parameter b_i indicates how much the subject specific intercept deviates from the average intercept.

3.5 Random-effects Model

As discussed earlier, a two-stage model is associated with drawbacks. These problems can be curbed by combining the two stages into one model. This combination yields the linear mixed (effects) model

$$\begin{cases} Y_i = X_i\beta_i + Z_i b_i + \epsilon_i \\ bi \sim N(0, D), \\ \epsilon_i \sim N(0, \Sigma_i) \\ b_1, \dots b_n, \quad \epsilon_i, \dots \epsilon_n \quad \text{are independent} \end{cases} \tag{4}$$

From this model Y_i which 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 is the n_i dimensional vector containing residual components. Finally, D is the general (q $\times$ q) covariance matrix.

4 Results

In this section we present the exploratory data analysis of the dataset in relation to the objectives of the study. We also presented the results obtained from the two stage, multivariate regression and linear mixed model.

4.1 Explanatory results

A total of 101 patients were followed up every six months. The table 1 below shows the number of patients present at each time point. From the table, it can be observed that the number of patients in the study continues to reduce as time increases, which indicate dropout from the study. Also, only one patient had eight complete measurements. The mean and the variance of the diameter seems to increase with time steadily but reduces on the 6^{th} measurement.

Table 1: General overview

Measurement	Average	Variance	Number of Patients
1	42.97	45.79	101
2	43.94	48.61	90
3	44.95	52.90	74
4	46.62	54.85	53
5	47.39	65.31	46
6	45.05	56.39	19
7	47.00	67.00	7
8	43.00	0	1

4.1.1 Individual Profile

The interest in the longitudinal data analysis is the evolution of individual patients over time, and this does give not only an idea about the most appropriate subject-specific regression but also provides 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 diameter 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.

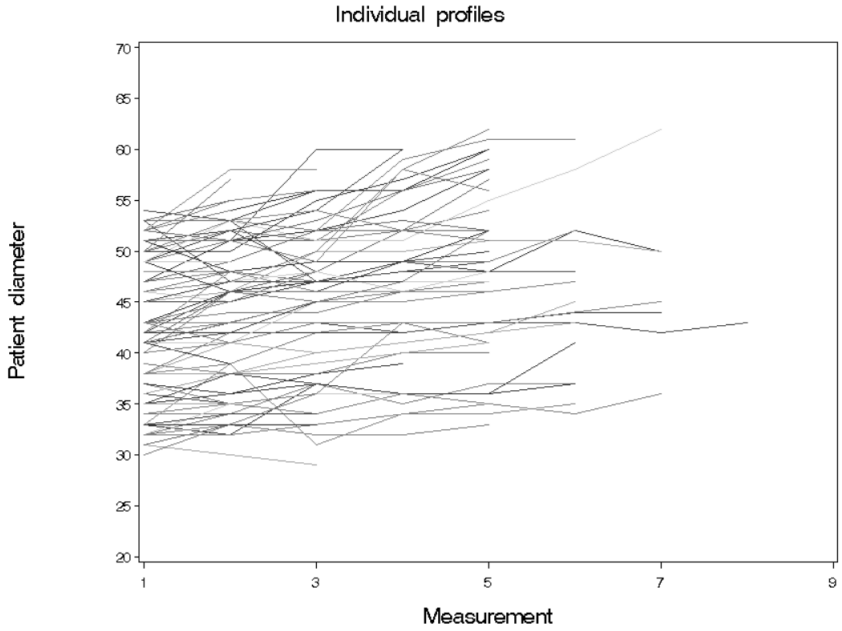

Figure 1: Individual patients profile

4.1.2 Mean Structure

This section describes the average evolution of artery diameter (in mm) of patients over time for the overall population. The mean structure gives the marginal relationship of the response with time. Figure 2 shows the mean structure of the patient. The structure indicates a linear trend with time except for the sharp decrease from the sixth measurement. 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.

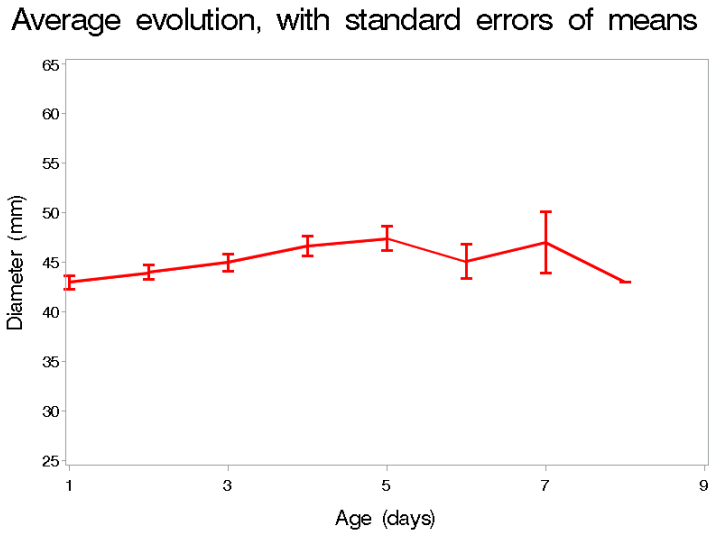

Figure 2: Mean structure

4.1.3 Variance Structure

After studying the mean structure, the evolution of variance is essential in building a longitudinal model (Verbeke and Molenberghs, 2000). Standardized squared residuals obtained from the mean structure were used to construct the variance function. The figure 3 suggests a stable variance over time with an increase in standard error at each time point; this trend is expected because of the attrition that exists in the data.

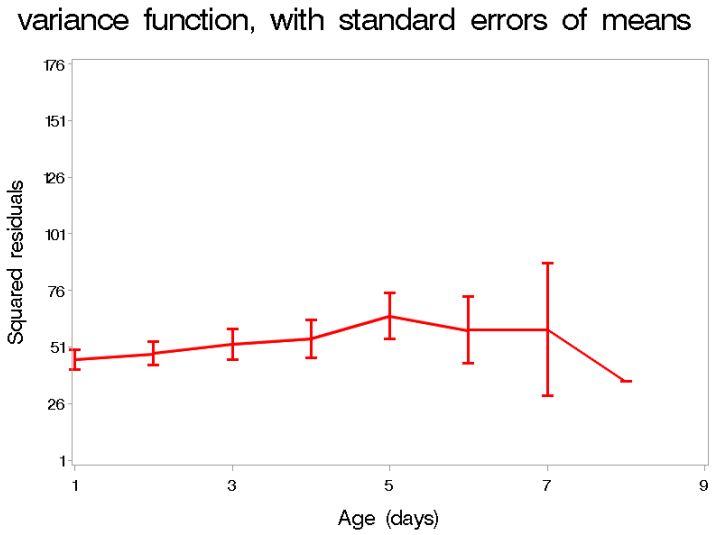

Figure 3: Variance structure

4.1.4 Correlation Structure

Usually, correlation is used to describe how the records of the subjects are correlated. In this study, correlation structure was studied using both the correlation matrix and the scatter plot matrix. But the focus here will be by use of the scatter plot matrix (Figure 4). From this plot, the off-diagonal elements obtained from the pairs of measurement occasions showed the decaying of the correlation with time. Also, the stationarity assumption suggests that the schemes remain within the diagonal bands since the measurements were collected at an equally spaced time interval, that is, measurement 1 for Month 0, measurement 2 for month 6 and the proceeding measurement obtained in that order of sequence.

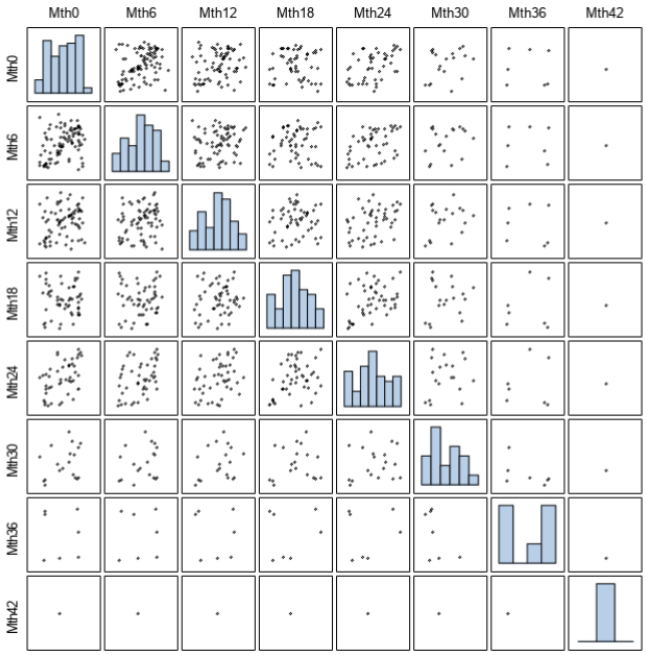

Figure 4: Correlation structure

4.2 Summary Statistics

This section present the results of the area under the curve and analysis of increment. Under the area under the curve, the trapezoidal rule was used. That is

$$AUC_i = (t_{i2} - t_{i1})\frac{(y_{i1} + y_{i2})}{2} + (t_{i3} - t_{i2})\frac{(y_{i2} + y_{i3})}{2} + ... \qquad \text{for} \quad i = 1, 2, ..., 101 \qquad (5)$$

As seen in section 2, this method compares the overall differences in the repeated measures for each subject. As regards the analysis of increment, the difference (d_i) between the last and the first observation for each subject was taken as shown in equation 6.

$$d_i = y_{in_i} - y_{i1} \qquad \text{for} \quad i = 1, 2, ..., 101 \tag{6}$$

Both methods were considered in this study because of their advantages, such as no problems with multiple testing and do not explicitly assume balance data. But their disadvantages is that they both use only partial information leading to loss of some information. The results of the analyses were shown in table 2. The findings showed that none of the variables was significant using both methods. Also, the standard errors of the Area under the curve was quite more higher than that of analysis of increment.

Table 2: Parameter Estimates for different summary statistics

	Area Under the Curve		Analysis of Increment	
	Estimate (SE)	P-value	Estimate (SE)	P-value
Intercept	-24.93 (123.58)	0.84	0.49 (5.27)	0.93
Age	1.64 (1.15)	0.15	0.04 (0.05)	0.47
BMI	1.06 (2.47)	0.66	0.02 (0.10)	0.85
Smoking	2.97 (14.03)	0.83	-0.56 (0.62)	0.37
Coronary disease	12.00(17.21)	0.49	0.91 (0.75)	0.23

4.3 Multivariate model

Different multivariate models were fitted to the data. The first model that was fitted was a saturated model adjusting for baseline characteristics and including the interaction between time and other variables, was fitted with time as a categorical variable. This model was chosen based on the exploration of the mean and variance structure which suggested that artery diameter could be modelled as a linear function of time with constant variance structure, so the first-order autoregressive variance structure (AR(1)) was used. To simplify the mean structure, a model with time as a continuous variable was fitted and the likelihood ratio test was used to compare the two models. The test was not significant and therefore, the second model was chosen. Our selected model was then fitted with different covariance structures, using heterogeneous first-order autoregressive (ARH(1)) and the heterogeneous compound symmetry (CSH) to check if our earlier assumption of constant variances was a valid one. Model 3 resulted in a non-significant p-value unlike model 4, suggesting that inference could be made based on model 3, that is, the model with heterogeneous first-order autoregressive covariance structure.

Table 3: Model Selection

Model	Mean	Covariance	Parameters	-2L	Ref	G^2	df	P-value
1	UN	AR(1)	54	1893.9				
2	$\neq$ Slopes	AR(1)	17	1930.1	1	36.2	37	0.50636
3	**$\neq$ Slopes**	**ARH(1)**	**24**	**1923.1**	**2**	**7**	**7**	**0.42888**
4	$\neq$ Slopes	CSH	24	1949.8	2	19.7	7	0.00626

Hence, the final model was model 3 where time was a continuous covariate and its interaction with age, BMI, coronary disease and smoking. The covariance structure assumed was heterogeneous first-order autoregressive covariance structure. The table 4 shows the type III test of fixed effects for the model. Results shows that the effect of age to the artery diameter was significant.

4.4 Two stage Analysis

4.4.1 Subject Specific Regression - First stage

To select the best possible regression relation between the diameter of the artery and time, a linear relationship and a quadratic relationship were assumed and examined. Each model was fitted for each subject, and the results were combined in the form of meta-analysis using R^2_{meta}. Figure

Table 4: Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr >F
Age	1	109	4.39	0.0384
BMI	1	179	2.19	0.1406
Smoking	2	138	0.2	0.8212
Coronary disease	1	103	0.42	0.516
Age*Time	1	257	0.24	0.6244
BMI*Time	1	256	0.03	0.8531
Smoking*Time	2	247	1.4	0.2479
Coronary disease*Time	1	222	0.29	0.5902

5 shows the subject specific coefficient R_i^2. The overall R_{meta}^2 of the left figure which assumes a linear relationship equals 0.79 while the overall R_{meta}^2 of the right figure which assumes quadratic relationship equals 0.86. This results implied that the linear relationship model explained about 79% of the total variation in the response while the quadratic relationships explained 86% of the total variation in the response. The dashed line shows when R^2 equals 0.3, and it was observed in both plots that 14 individual R^2 were below the line. Although there is a gain in R_{meta}^2 using the quadratic term, this can be as a result of the drawback of the R^2 which will always increase adding higher order of time. However, given the good fit already obtained in the linear model and to keep our model as parsimonious as possible, we based our analyses on the linear model.

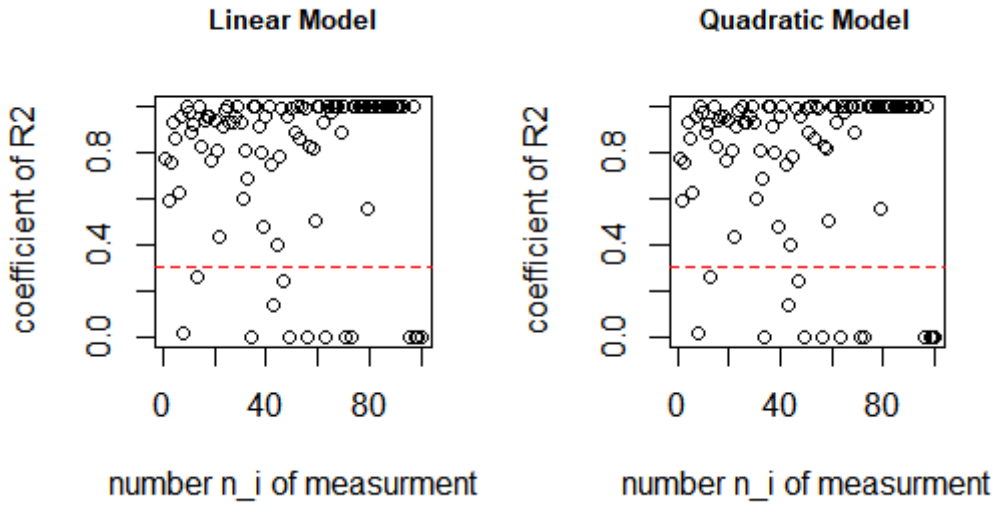

Figure 5: Subject specific coefficient R_i^2 of multiple determination (dashed lines)

4.4.2 2 stage Analysis

In this section, a two-stage model was considered, and a linear model was fitted for each patient as seen in the previous section. Table 5 shows the results of the second stage analysis.

Table 5: Results from second stage analysis

Response		Estimate	Std. Error	t value	Pr(>|t|)
β_{oi}	Intercept	20.7545	9.4876	2.19	0.0312
	Age	0.2092	0.0889	2.35	0.0206
	BMI	0.1831	0.1897	0.97	0.3368
	Smoking	0.7682	1.0771	0.71	0.4774
	Coronary disease	0.5744	1.3208	0.43	0.6646
β_{1i}	Intercept	-0.2312	2.3059	-0.10	0.9204
	Age	-0.0007	0.0213	-0.03	0.9726
	BMI	0.0643	0.0462	1.39	0.1674
	Smoking	-0.2000	0.2517	-0.79	0.4289
	Coronary disease	0.0621	0.3110	0.20	0.8422

Findings from the table 5 shows that at baseline, the evolution of the diameter of the abdominal aorta was significantly influenced by age, but at later time points, the effect of age was not significant.

4.5 Linear Mixed Model

Based on the previous discussion on the limitations of the two-stage model and the multivariate model, a linear mixed model was fitted to study whether several patient characteristics indeed affect the diameter of the abdominal aorta. A model with random intercept and random slope, the interaction of time and other covariates was fitted. Table 6 shows the result of the fitted model. Also, the effect of age and the interaction of smoke with time significantly affect the diameter of the abdominal aorta in patients with AAA.

Table 6: Type III test for Fixed Effect

Effect	Num DF	Den DF	F Value	P-value
Age	1	194	4.36	0.0381
Smoke	2	194	0.83	0.4371
Bmi	1	194	1.72	0.1908
Coronary disease	1	194	0.43	0.5143
Measurement*smoke	2	194	4.54	0.0118
Bmi*measurement	1	194	0.18	0.668
Coronary disease*Measurement	1	194	0.38	0.5392
Age*measurement	1	194	0.29	0.5913

For this analysis, the variance-covariance structure for the random effects (D) was kept unstructured majorly because of the design of the study and findings from the exploratory analysis.

The table 7 shows the comparison of the estimate of the multivariate and linear mixed model with random intercept and random slope. Noticeably from the result, the parameter estimates, standard errors, and p-values of both model show no much differences. Besides, age consistently was significant in the two models and there it can be concluded that a unit increase in age increases the diameter of the abdominal aorta by 0.1961 and 0.2013 according to Linear mixed model and Multivariate regression model respectively. As far as the evolution in the two models seems similar, linear mixed models take into account the variability from the average evolution for each subject.

Table 7: Comparison of Estimate of the Multivariate and Linear Mixed Model

	Multivariate model			Linear Mixed model		
Effect	Estimate	SE	P-value	Estimate	SE	P-value
Intercept	21.0658	9.2132	0.0238	21.9370	9.3910	0.0216
Age	0.2013	0.0960	0.0384	0.1961	0.0936	0.0375
BMI	0.2518	0.1701	0.1406	0.2268	0.1715	0.1876
Non-Smoker	-0.9919	2.6848	0.7125	-1.6329	2.1576	0.4501
Ex-Smoker	0.4747	1.1795	0.6878	0.7399	0.8351	0.3768
No Coronary Disease	-0.9316	1.4294	0.5160	-0.8343	1.3156	0.5267
Age*Time	0.0091	0.0185	0.6244	0.0084	0.0153	0.5831
BMI*Time	-0.0065	0.0351	0.8531	0.0156	0.0362	0.6680
Non-Smoker*Time	1.2846	1.8555	0.4893	0.8024	1.5312	0.6009
Ex-Smoker*Time	0.5559	1.9015	0.7702	0.0453	1.5658	0.9769
No Coronary Disease*Time	-0.1348	0.2500	0.5902	-0.1615	0.2509	0.5205

Moreover, since the random effects in a linear mixed model represent the variability in subject specific intercept and slopes that are not explained by the fixed covariates (Verbeke and Molenberghs, 2000), then there was a need to test whether the random effects were appropriate to be included in the model. Four different models were fitted as seen in table 8. The first with random intercept and random slope (model1), random intercept (model2) only, random slope (model3) only, neither random intercept nor random slope (model4). Table 8 summarizes the variance-covariance estimates and their corresponding standard errors for the four models as well as the REML for each resulting model.

Table 8: Summary of Variance components of different random effect models

Effect		Model1	Model2	Model3	Model4
Covariance of b_1		Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Var(b_{1i})	d_{11}	38.8461 (6.2590)	42.6539 (9.2430)	—-(—)	—-(—)
Var(b_{2i})	d_{22}	0.8274 (0.2297)	—-(—)	0.3124 (0.2421)	—-(—)
Cov(b_{1i}, b_{2i})	d_{12}= d_{21}	1.3698 (0.8542)	—-(—)	—-(—)	—-(—)
Residual variance					
Var (ϵ_{ij})		2.2721 (0.4249)	11.4227 (6.1581)	50.5005 (7.7418	55.43 (7.6704)
REML		1921.7	1948	1947.6	1950.0

From table 8 it can be seen that when one random effect was removed from the model, the estimated Restricted Maximum Log-likelihood (REML) value increases. Furthermore, since the models to be compared have the same mean structure and can lead to same error contrast, the estimated REML instead of maximum likelihood was used compare the models with different covariance structures as in Table 8. The likelihood ratio test statistic used in this study is a mixture of chi-squared distributions rather than the classical single chi-squared distributionss (Verbeke and Molenberghs 2000).

4.6 Marginal testing for need the random effects

The hypothesis of interest in the first case was $H_0 : d_{12} = d_{22} = 0$ where d_{12} was the covariance between random slope and random intercept, and d_{22} was the variance of random slope. The obtained p-value is statistically significant at 5%, as seen in table 9. Hence, the null hypothesis was rejected, and we concluded that the model with both random slope and intercept was appropriate. For the next model comparison, the other null hypothesis of interest was $H_0 : d_{11} = d_{12} = 0$ and was $H_0 : D = 0$ respectively. Both hypotheses lead to the same conclusion as in the first case.

Table 9: Testing for the need of random effects

Hypothesis	-2$ln(\lambda_N)$	Asymptotic H_0	P-value
Model1 vs Model 2	26.3	$\chi^2_{1:2}$	$< .0001$
Model1 vs Model 3	25.9	$\chi^2_{1:2}$	$< .0001$
Model1 vs Model 4	28.3	$\chi^2_{1:3}$	$< .0001$

4.6.1 Random effects plot comparing the Two-stage model and Random Effects model

The figure 6 shows the plot of random slope against random intercept for the two-stage (left) and linear mixed model (right). The two-stage model shows no observed trend, and the points seem scattered. However, the linear mixed model shows a positive trend between the random slope and random intercept. 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. An advantage that the Linear Mixed Model has is that it takes into account the correlation structure of the data and also uses all the information in the data.

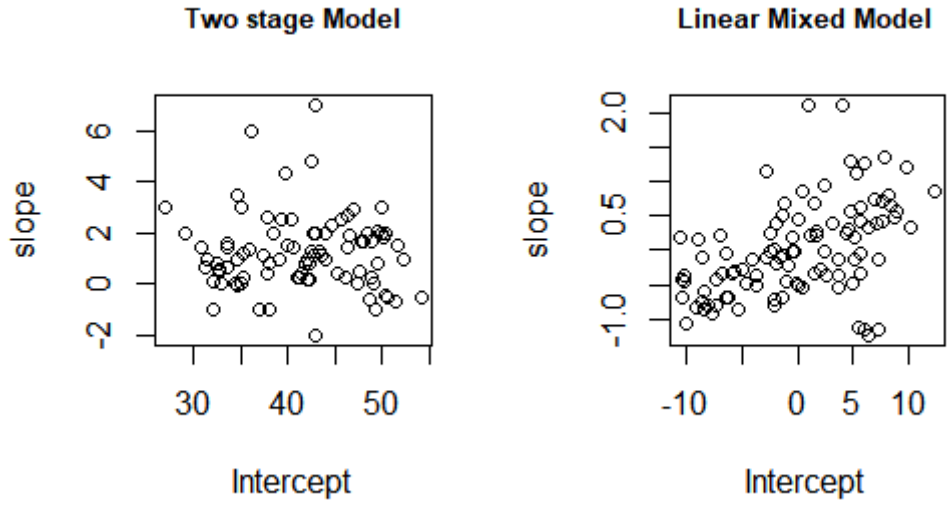

Figure 6: Two Stage Model and Linear Mixed Model plots of random effects

5 Discussion and Conclusion

From the individual profile plot, there was a clear suggestion of the presence of much variability both between and within the subjects. The summary statistics, on the other hand, both indicated that there is no effect of age, smoking, BMI and history of coronary disease on the diameter of the abdominal aorta in patients with AAA. When using the information from the summary statistics, one has to be cautious since these simple summary statistics are based on partial observations to generalize the whole conclusion and therefore the conclusion should not be made at this stage. Further, Length and Weight were removed as predictors in this analysis since all the mass of this two predictor variables were contained in body mass index.

Besides the limitations of the summary statistics, a final multivariate regression model was fitted with a heterogeneous autoregressive correlation structure. This model showed that the age of the patient has a significant effect on the evolution of the diameter of the abdominal aorta.

The two-stage model was also fitted but its use in analysis of longitudinal data has been criticized because it suffers from the fact that it does not take into account the correlation that exists on the patient observations. Finally, to overcome the limitations of the previously discussed models, a linear mixed model with random intercept and slope was fitted. The test for random effects showed that there was a need for random slope and random intercept. From the model, age and interaction of smoking status with time were significantly affecting the evolution of the diameter of the abdominal aorta. In other words, as long as the patient if older enough the effect of the

increase in the diameter of the artery is significant. Also, the smoking status which can change over time can lead to an impact on the widening of the aorta.

6 Recomendation

As a recommendation, future studies should include the gender of the subjects participating in the study since AAA may be affecting individual of the particular gender and the evolution between the different gender might be ultimately different. Furthermore, if the same study is planned, the interval between the two study visits should be reduced to mitigate the high number of dropouts.

References

[1] Vorp, D. A., & Geest, J. P. V. (2005). Biomechanical determinants of abdominal aortic aneurysm rupture. Arteriosclerosis, thrombosis, and vascular biology.

[2] Aggarwal, S., Qamar, A., Sharma, V., & Sharma, A. (2011). Abdominal aortic aneurysm: a comprehensive review. Experimental & Clinical Cardiology, 16(1), 11.y.

[3] Verbeke G. and Molenberghs G. (2009). Linear Mixed Models for Longitudinal Data.

[4] Molenberghs G, Verbeke G. (2005). Models for Discrete Longitudinal Data, 1 edn.New York: Springer-Verlag New York.

[5] Gałecki, A., & Burzykowski, T. (2013). Linear mixed-effects models using R: A step-by-step approach. Springer Science & Business Media.

[6] Robert E. Weiss. (2005). Modeling Longitudinal Data-Springer-Verlag New York.

[7] Verbeke, G. and Molenberghs, G. (2019), Introduction to Longitudinal Data Analysis: Lecture Notes, Hasselt University.

[8] Izenman A.J. (2013) Multivariate Regression. In: Modern Multivariate Statistical Techniques. Springer Texts in Statistics. Springer, New York, NY

Codes

SAS Code

```
/*****individual plot****/;
goptions reset=all ftext=swiss device=psepsf gsfname=fig1
gsfmode=replace rotate=landscape i=join;
proc gplot data=lda.aaa_1;
plot diameter*measurement=patient / haxis=axis1 vaxis=axis2 nolegend;
axis1 label=(h=2 'Measurement') value=(h=1.3) order=(1 to 10 by 2)
minor=none;
axis2 label=(h=2 A=90 'Patient diameter') value=(h=1.3) order=
(20 to 70 by 5)
minor=none;
title h=2 'Individual profiles';
run;quit;

/****mean structure plot****/
goptions reset=all;
proc gplot data=lda.aaa_2;
plot diameter*measurement=smoke / haxis=axis1 vaxis=axis2 legend=legend1;
symbol1 i=std1mjt w=2 color=red;
axis1 label=(h=1.5 'measurement') value=(h=1.5) order=(1 to 10 by 3)
minor=none;
axis2 label=(h=1.5 A=90 'Diameter (in mm)') value=(h=1.5) order=(20 to 90 by 10) minor=none;
```

```
legend1 label=none position=(top left inside) offset=(5,-3) value=(H=1.3
"Non smokers" "Ex-smokers" "Smokers") frame down=3;
title h=1.5 'Average evolution'; run;

***************Multivariate models***************
/*****Model 1 - Time Categorical - AR(1)*****/
proc mixed data = lda1 method = ml;
class patient time smoking cdisease;
model diameter =  time*age time*bmi time*cdisease time*smoking /noint ddfm=satterth s;
repeated time / type = AR(1) subject = patient r rcorr;
run;

/*****Model 2 - Time Continuous - AR(1) - Ref Model 1*****/
proc mixed data = lda1 method = ml;
class patient timeclss smoking cdisease;
model diameter =  age bmi smoking cdisease time*age time*bmi time*smoking
time*cdisease / ddfm = satterth s;
repeated timeclss / type = AR(1) subject = patient;
run;

/*****Model 3 - Time Continuous - ARH(1) - Ref Model 2*****/
proc mixed data = lda1 method = ml;
class patient timeclss smoking cdisease;
model diameter = age bmi smoking cdisease time*age time*bmi time*smoking
time*cdisease / ddfm = satterth s;
repeated timeclss / type = ARH(1) subject = patient;
run;

/*****Model 4 - Time Continuous - CSH - Ref Model 2*****/
proc mixed data = lda1 method = ml;
class patient timeclss smoking cdisease;
model diameter = age bmi smoking cdisease time*age time*bmi time*smoking
time*cdisease / ddfm = satterth s;
repeated timeclss / type = CSH subject = patient;
run;

/****LMM-1* Random Intercept and R Slope****/
proc mixed data=lda.aaa_2 method=reml empirical covtest;
class patient cdisease smoke measu_class;
model diameter = age smoke bmi cdisease smoke*measurement bmi*measurement
    age*measurement cdisease*measurement/ s chisq ;
random intercept measurement / type=un subject=patient g gcorr v vcorr solution ;
repeated measu_class / type=ar(1) subject=patient r rcorr;
ods output solutionr=lda.lmm_out; run;
run;

/****LMM-2* No random slope****/
proc mixed data=lda.aaa_2 method=reml  empirical covtest;
class patient cdisease smoke measu_class;
model diameter = age smoke bmi cdisease smoke*measurement bmi*measurement
    age*measurement cdisease*measurement/ s chisq;
random intercept / type=un subject=patient g gcorr v vcorr ;
repeated measu_class / type=AR(1) subject=patient r rcorr;
run;

/****LMM-3* No random intercept + slope****/
```

```
proc mixed data=lda.aaa_2 method=reml  empirical covtest;
class patient cdisease (ref=LAST) smoke (ref=LAST) measu_class;
model diameter = age smoke bmi cdisease smoke*measurement bmi*measurement
    age*measurement cdisease*measurement/ s chisq;
repeated measu_class / type=AR(1) subject=patient r rcorr;
run;

R Code
#############Two stage model#########
######stage 1
#####Linear
for(i in 1:100){
  Model1 <- lm(D.max.Abdominaal..mm. ~ Measurement,
            data=AAA[(AAA$Patient2==i), ], x=T)
  r2[[i]] <- summary(Model1)$r.squared
  SSE[[i]] <- sum(Model1$residuals^2)
  SSR[[i]] <- anova(Model1)$`Sum Sq`[[1]]
  ni[[i]] <- length(Model1$x[,1])
  pi <- length(Model1$coefficients)
  intercept[[i]] <- Model1$coefficients[[1]]
  slope[[i]] <- Model1$coefficients[[2]]
  SEintercept[[i]] <- summary(Model1)$coefficients[,2][1]
  SEslope[[i]] <- summary(Model1)$coefficients[,2][2]
}
R2_meta <- sum(Modelstat$SSR)/sum(Modelstat$SSR +Modelstat$SSE)

####Quadratic
for(i in 1:100){
  Model2 <- lm(D.max.Abdominaal..mm. ~ Measurement + Measurement2 ,
               data=AAA[(AAA$Patient2==i), ], x=T)
  r22[[i]] <- summary(Model2)$r.squared
  SSE2[[i]] <- sum(Model2$residuals^2)
  SSR2[[i]] <- anova(Model2)$`Sum Sq`[[1]]
  ni2[[i]] <- length(Model2$x[,1])
  pi2 <- length(Model2$coefficients)
  intercept2[[i]] <- Model2$coefficients[[1]]
  slope2[[i]] <- Model2$coefficients[[2]]
  SEintercept2[[i]] <- summary(Model2)$coefficients[,2][1]
  SEslope2[[i]] <- summary(Model2)$coefficients[,2][2]
}

######stage 2
InterceptModel <- lm(CombinedLongData$Modelstat.intercept ~
CombinedLongData$AAAresh2.Age.at.study.entry.1 +
                       CombinedLongData$AAAresh2.BMI +
                       CombinedLongData$AAAresh2.Smoking.. +
                       CombinedLongData$AAAresh2.Coronary.disease.. +
                       CombinedLongData$AAAresh2.Smoking..)
SlopeModel <- lm(CombinedLongData$Modelstat.slope ~
CombinedLongData$AAAresh2.Age.at.study.entry.1 +
                       CombinedLongData$AAAresh2.BMI +
                       CombinedLongData$AAAresh2.Smoking.. +
                       CombinedLongData$AAAresh2.Coronary.disease..)

#####Area Under the curve
AUC = trapz(AAA$Measurement, AAA$D.max.Abdominaal..mm.)
```

```
AUC = c(100)
for(i in 1:100){
  AUC[i] = trapz(AAA$Measurement[AAA$Patient2 ==i],
  AAA$D.max.Abdominaal..mm.[AAA$Patient2==i])
}

AUCModel <- lm(AUC ~ CombinedLongData$AAAresh2.Age.at.study.entry.1 +
                 CombinedLongData$AAAresh2.BMI +
                 CombinedLongData$AAAresh2.Smoking.. +
                 CombinedLongData$AAAresh2.Coronary.disease.. +
                 CombinedLongData$AAAresh2.Smoking..)

######Analysis of Increments
Increments <- lm(Increment$AOI ~ Increment$Age + Increment$BMI + Increment$Smoking +
                 Increment$CD)

####
```