Linear Mixed model group assignment

Group 10[1,✉]

[1] Statistics and Data Science, Leiden University

This study investigates how joint destruction evolves over time in rheumatoid arthritis patients and how it is related to gender, age and genetic background of the patient.

1. Introduction

Joint destruction in rheumatoid arthritis has been argued to be impaired over time. This impairment might also be related to other patients' profiles, e.g., age, gender, and genetic background. Meanwhile, health consciousness and medical strategies differ among periods might also influence this progress. In this study, we investigate how joint destruction evolves over time in rheumatoid arthritis patients and how it is related to gender, age and genetic background of the patient whilst considering society's health consciousness and medical strategies by including the period of collecting data. The dataset we used is from group assignment 10 in the linear mixed model.

1.1. Research Question.

How joint destruction evolves over time in rheumatoid arthritis patients and how it is related to the gender, age and genetic background of the patient. In greater detail, we intended to test the main effect of time and its interaction with gender, age, and genetic background.

2. Methods

2.1 Information about the dataset.

This analysis used the dataset distributed to group 10 in the linear mixed model course. The original study included 500 patients were included between 1990 and 2006. Each patient was intended to be followed up 8 times. The joint destruction was quantified by single nucletide polymorphism (SNP) and evaluated by the clinicians as low or high based on a combination of clinical parameters, i.e., Severity. In terms of the period of data collection, a patient who joined before "1996-01-01" belongs to period A while a patient who joined after "1996-01-01" belongs to period B.

2.2. Descriptive Information of Variables in the Dataset.

The long-format dataset includes 8 variables (i.e., id, Period, Age, Sex, SNP, visit, SHS and Severity). The SHS and Severity are response variables representing the disease progression. The detailed information about variables in the dataset is as follows:

- **SHS**: ranges from 0 to 448, and the higher the score the higher the disease activity is. The mean SHS is 25.85 and the standard derivative is 16.58.
- **Severity**: Disease severity (0: low, 1: high) at each of the 8 visits, respectively.
- **Visit**: Visit number, taking values of 0, 1, 2, 3, 4, 5, 6, and 7.
- **Visit.f**: Visit number, factorized Visit, taking values of Visit0, Visit1, Visit2, Visit3, Visit4, Visit5, Visit6, and Visit7.
- **Period**: before vs. after '1996-01-01', with 1368 and 1398 observations respectively.
- **SNP**: single nucleotide polymorphism (no variant alleles vs. variant allele vs. variant alleles), with 1394, 818, and 554 observations respectively.
- **Age**: age of the patient at inclusion, ranging from 41 to 88 with 63.95 and 13.97 as mean and standard derivatives, respectively.
- **sex**: sex indicator (male vs. female). There are 1309 men and 1457 women.
- **id**: patient indicator, ranging from 1 to 500.

2.3. Data Analysis on SHS.

2.3.1. Missing Data Issues Checking.

Before conducting the analysis, we check whether there are any missing data issues in terms of SHS in our dataset. As shown in *Table 1*, there are missing follow-ups, but the predictor variables are complete.
According to the description of the data, some of the planned follow-ups were not recorded due to e.g., loss to follow-up, relocation and remission. Such reasons are not relevant to the observed data, neither are unobserved data. Thus, we argue that the data is missing completely at random (MCAR).

2.3.2. Exploratory Data Analysis.

Mean Structure.

We first conducted an Exploratory Data Analysis (EDA) to inform us of a starting point of the mean structure, correlation and variance structure of the further modelling. For the mean structure, it can shown that there is an almost linear but probably also some quadratic trend between SNP and visiting time (see in *Fig.1*). This relationship between SNP and visiting time might also vary across sexes (male vs. female, see in *Fig.2*), inclusion period (before vs. after '1996-01-01', *Fig.3*) and SNP (no variant alleles vs. variant allele vs. variant alleles, *Fig.4*). Alongside, the SNP is also linearly affected by age (see in *Fig.5*). This linear relationship might also be affected by the visiting time (see in *Fig.6*).

Table 1. Count of missing data when measuring SHS

id	Period	Age	sex	SNP	Visit0	Visit1	Visit2	Visit3	Visit4	Visit5	Visit6	Visit7
0	0	0	0	0	0	173	168	168	176	178	177	194

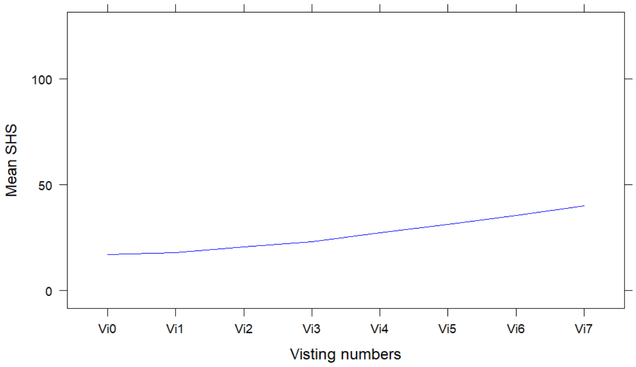

Fig. 1. EDA between SHS and visit.f

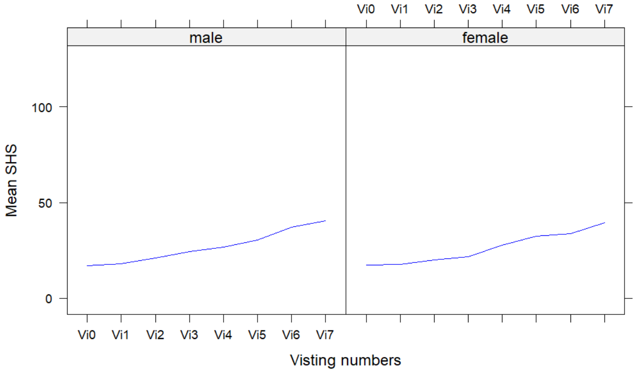

Fig. 2. EDA between SHS and visit across sex

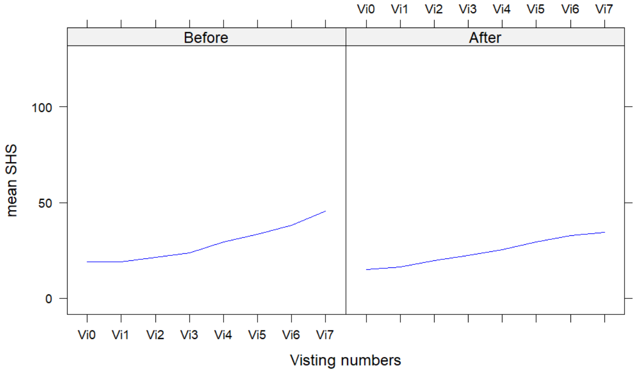

Fig. 3. EDA between SHS and visit.f across Period

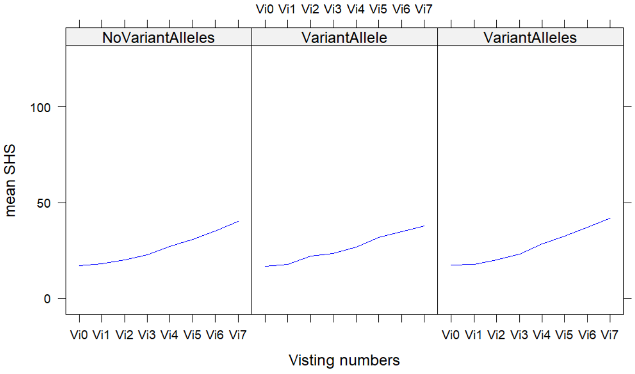

Fig. 4. EDA between SHS and visit.f across SNP

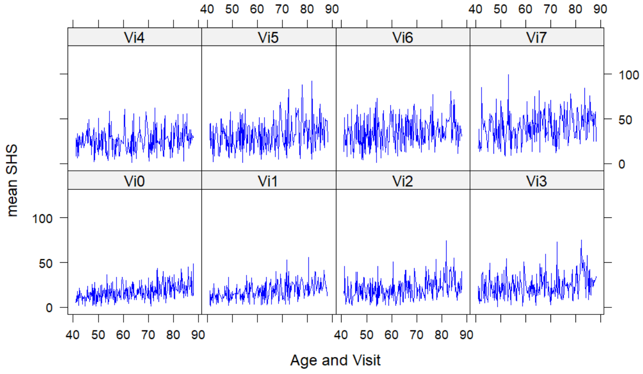

Fig. 6. EDA between SHS and Age across visit.f

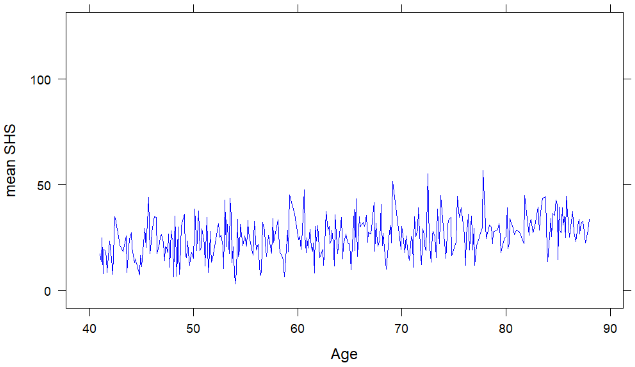

Fig. 5. EDA between SHS and Age

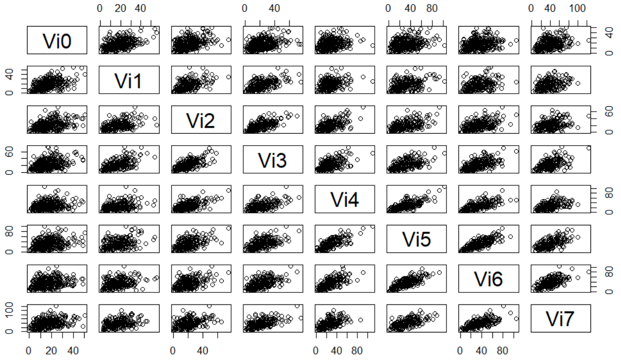

Fig. 7. EDA between SHS and Age across visit.f

Correlation Structure.

For the correlation structure, it can be seen in *Fig.7* that the correlation between different visits varies.

Variance Structure.

For the variance structure, it can be seen in *Fig.8* that it is different among visiting times.

2.3.3. Model building.

To obtain the most descriptive model on the given data, we used the *lme()* function in *nlme* package to build up the relationship between SHS and the predictors (i.e., Period, Age, Sex, SNP, visit, SHS and Severity) by taking account of the correlation between the different observations. In greater detail, we started the model with the most elaborate mean structure we observed in the section of EDA, i.e., SHS ˜ (Age + Sex + SNP + Period) * Visit.f.. With this mean structure, we intended to obtain the corresponding appropriate variance-covariance structure based on (a) for the nested model, we conducted an approximate LRT test which follows $0.5\chi^2_0 + 0.5\chi^2_1$ distribution under the null hypothesis; (b) for the non-nested model, the model with lowest Akaike's Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC) were selected. With the appropriate variance-covariance structure, we simplified the mean structure based on the item's significance and our research interest. The assumptions of the linear mixed model were checked by visualizing the conditional and marginal residuals. To obtain the information of our interested test, we used the LRT, f and t-test.

2.4. Data Analysis on Disease Severity.

2.4.1. Missing Data Issues Checking.

Before conducting the analysis, we check whether there are any missing data in terms of the Severity issues in the variables in our dataset. As shown in *Table 2*, there are missing follow-ups, but the predictor variables are complete. According to the description of the data, some of the planned follow-ups were not recorded due to e.g., loss to follow-up, relocation and remission. Such reasons are not relevant to the observed data, neither are unobserved data. Thus, we argue that the data is missing completely at random (MCAR).

2.4.2. Exploratory Data Analysis.

Mean Structure.

We first conducted an Exploratory Data Analysis (EDA) to inform us of a starting point of the mean structure, correlation and variance structure of the further modelling.

Table 2. Count of missing data when measuring Severity

id	Period	Age	sex	SNP	Visit0	Visit1	Visit2	Visit3	Visit4	Visit5	Visit6	Visit7
0	0	0	0	0	0	173	168	168	176	178	177	194

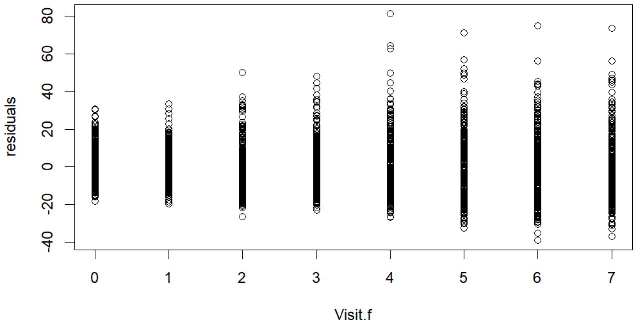

Fig. 8. EDA between SHS and Age across visit.f

For the mean structure, it shows that the percentage of patients with high severity is increasing with the visiting time (see in *Fig.9*). This relationship varies across sex (male vs. female, see in *Fig.10*), period (before vs after "1996-01-01", see in *Fig.11*) and age (see in

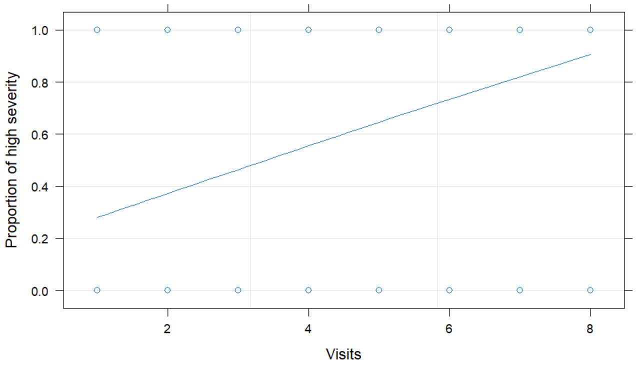

Fig. 9. EDA between Severity and visit.f

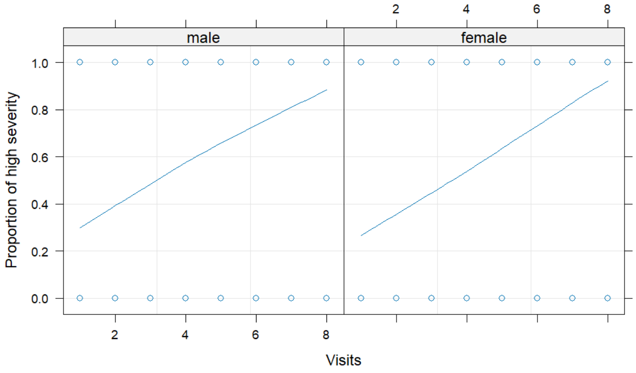

Fig. 10. EDA between Severity and visit.f across sex

Correlation Structure and variance structure . Checking the variance-covariance structure in generalized linear mixed model is difficult in general. Thus, we left it to the model

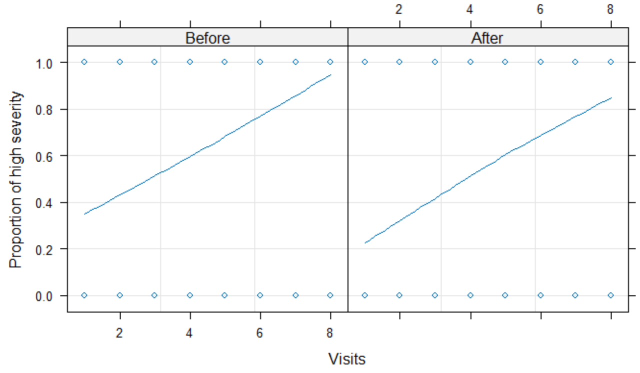

Fig. 11. EDA between Severity and visit.f across Period

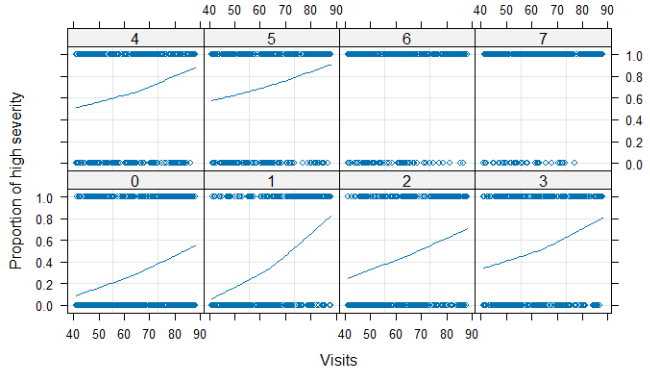

Fig. 12. EDA between Severity and Age across visit.f

selection part. We believe that the model with the best fit has an appropriate variance-covariance structure.

2.4.3. Model building.

To obtain the most descriptive model on the given data, we used the *mixed_ model()* function in *GLMMadaptive* package to build up the relationship between Severity and the predictors (i.e., Period, Age, Sex, SNP, visit, SHS and Severity) by taking account of the correlation between the different observations. In greater detail, we started the model with the most elaborate mean structure we observed in the section of EDA, i.e., Severity ~ (Age + Sex + SNP + Period) * Visit.f. With this mean structure, we intended to obtain the corresponding appropriate variance-covariance structure based on (a) for the nested model, we conducted an approximate LRT test which follows $0.5\chi_0^2 + 0.5\chi_1^2$ distribution under the null hypothesis; (b) for the non-nested model, the model with lowest Akaike's Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC) were selected. With the appropriate variance-covariance structure, we simplified the mean structure based on the item's significance and our research interest. To obtain the information of our interested test, we used the LRT, f and t-test.

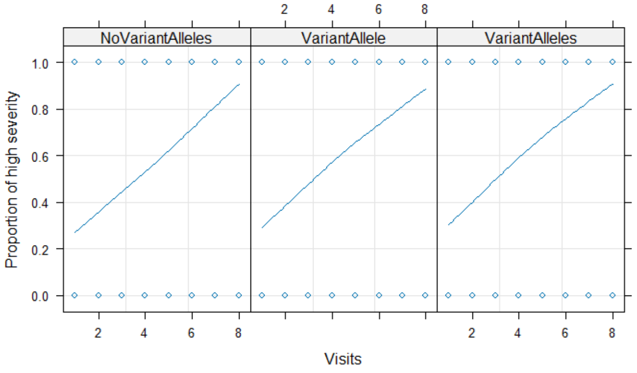

Fig. 13. EDA between Severity and Age across SNP

3. Result

Result of linear mixed effect model for SHS.

According to the mixed effect model for SHS, The mean sharp index increases over visits when the ages are the same during visits in both period Before and period After. The rate of change would increase over time in both period A and period B. The rate of change of the mean sharp index at $visit_j$ for period A is 0.4306+0.962$Visit_{ij}$, and the rate of change of the mean sharp index at visit j for period B is 2.19+0.1958$Visit_{ij}$. The mean sharp index and it's evolution over time all are not related to sex and SNP of the patients for period A and Period B. The mean sharp index would be higher for the older people during same visit time for period Before and Period After. The sharp index would increase by 26.85% with one year increase in age during same visit time for period A and period B

Result of generalized linear mixed effect model for Severity.

The results of our generalized linear mixed effect model for Severity show (seen in Table 4) that: 1). the relationship between time (quantified by visit) and Severity is not linearly significant ($\beta = 0.1303, se = 0.1050, z = 1.2410, p = 0.2146$) but quadratically yes ($\beta = 0.0363, se = 0.0159, z = 2.2797, p = 0.0226$). Note that the linear terms and quadratic terms in this study are not linearly independent. The rate of Severity in terms of the time is 0.2029. For the interaction between visit and other variables, we can see a significant interaction between age and visit (either linear or quadratic term), also that between Period and visit (either linear or quadratic term).

4. Discussion

4.1 SHS.

In general, our findings corroborate the progressive nature of rheumatoid arthritis, where the absence of effective management strategies invariably leads to the exacerbation of joint damage over time. Notably, the Sharp index exhibited a significant upward trend, highlighting the relentless progression of the disease. The result indicates that the mean Sharp index increases with each visit and more sharply with age, showing no significant correlation with sex or SNP of patients. Specifically, the rate of change per visit differs between period A (0.4306 + 0.962Visit) and period B (2.19 + 0.1958Visit), with a steeper increase observed in the period B. Importantly, the index increases by 26.85 % for every additional year of age, highlighting age as a critical factor in its progression. This suggests that monitoring strategies should prioritize age and visit frequency to effectively manage conditions assessed by the Sharp index. A detailed examination of the factors influencing the Sharp index revealed that both the chronological progression of the study (denoted by 'Visit') and the age of the patients (grouped into 'Age') exert a substantial positive influence on the index. This could reflect the cumulative effects of the disease over time and the impact of ageing on disease progression. Conversely, The inclusion period ('Period') presents a negative effect on the Sharp index, indicating that patients included in the study at later periods tend to have lower Sharp index scores compared to those included earlier.

4.2 Severity.

With respect to the Severity, we found that it is affected by time. This relationship is also affected by the period and age. Note that the p-values were obtained in the marginal coefficients directly, suggesting that the inferential interpretation might be to some extent invalid if we change the order of the terms. Idealistically, we could be able to use the multivariate Wald t-test, F-test or LRT. However, we were not able to obtain the marginal variance-covariance matrix. Thus, we were kind of limited b

5. Acknowledgement

The code of the analysis can be seen in the file *Group10LME.Rmd* file.

Yufang Wang contributes to the EDA of both models, the writing for the most part except the result of linear mixed model and section 4.1. She also helped with the modelling of the generalized linear mixed model.

Qianqian Bai and Qingyu Meng contribute to the modelling of linear mixed model, the writing of results and discussion for the linear mixed model.

Weimin Wang and Yuqiang zhan contribute to the modelling of the generalized linear mixed model.

Table 3. Output of the mixed effect model for SHS

	Estimate	**L 0.95CI**	**U 0.95CI**	**p-value**
Visit	0.000	0.000	0.000	0.000
PeriodB	-0.054	0.015	-0.083	-0.025
Age	3.7505	3.0874	4.4136	0.0000
I($Visit^2$)	0.48101	0.3728	0.5892	0.0000
$Visit : PeriodAfter$	1.7595	0.7304	2.7885	0.0008 PeriodB:I($Visit^2$)
-0.3831	-0.5347	-0.2316	0.000	

Table 4. Output of the mixed effect model for Severity

Intercept	Estimate	se	z-value	p-value
Intercept	-0.4094	0.1646	-2.4880	0.0128
Age	0.7076	0.0968	7.3067	0.0001
sexfemale	-0.1294	0.1832	-0.7062	0.4801
SNPVariantAllele	0.0048	0.2162	0.0222	0.9823
SNPVariantAlleles	-0.0276	0.2312	-0.1193	0.9050
PeriodAfter	-0.6405	0.2037	-3.1445	0.0017
Visit	0.1303	0.1050	1.2410	0.2146
I($Visit^2$)	0.0363	0.0159	2.2797	0.0226
Age:Visit	-0.1202	0.0621	-1.9361	0.0529
sexfemale:Visit	-0.0911	0.1111	-0.820	0.412
SNPVariantAllele:Visit	0.1808	0.1354	1.3349	0.1819
SNPVariantAlleles:Visit	0.1863	0.1367	1.3632	0.1728
PeriodAfter:Visit	0.2551	0.1172	2.1768	0.0295
Age:I($Visit^2$)	0.0171	0.0088	1.9429	0.0521
sexfemale:I($Visit^2$)	0.0209	0.0153	1.3669	0.1716
SNPVariantAllele:I($Visit^2$)	-0.0315	0.0194	-1.6253	0.1041
SNPVariantAlleles:I($Visit^2$)	-0.0289	0.0194	-1.4853	0.1375
PeriodAfter:I($Visit^2$)	-0.0421	0.0157	-2.6835	0.0073