Longitudinal Data Analysis: Assignment 2

Authors:
Campbell McDuling

Student Numbers:
MCDCAM001

May, 2023

Contents

1 Introduction

Mixed effects models are a class of regression models that are commonly used to analyse grouped data. These express an outcome as a linear combination of fixed effects (population-level parameters) and random effects (grouping-level parameters). Mixed effects models are thus often used to analyse longitudinal data, where the random effects effectively capture the inherent correlation between observations that come from the same group. The nature of the data-generating mechanism can be made further complex when the outcome of interest varies with covariates in a non-linear manner. One way of capturing this is to introduce polynomials or spline functions which enter the model through a linear parameterisation. This approach, however, is limited to approximating the underlying mechanism within the range of the observed data. Furthermore, these models can quickly become overly complex - and typically as higher order polynomials are required to capture more complicated non-linear trends, the models become less interpretable and less parsimonious [7]. For example, the fifth term in a polynomial function may not have any directly comprehensible meaning in relation to an organism's growth over time. Non-linear mixed effects models (NLMEs) are an alternative to this approach, where the model is often mechanistic and derived from some theoretical knowledge of the underlying data generating mechanism. For example, growth curves will often be suitably modeled by a logistic function of time - where the natural plateau of an organism's growth would be captured by a horizontal asymptote. In this report, the first-order compartment model is used to model the pharmacokinetic trajectory of anti-malarial drugs in patients' blood after a single oral dose - where the model's parameters directly measure the absorption, clearance and elimination rates as the drugs are processed by the body.

The general model formulation for NLMEs is not very different from that of LME's. In the first stage, the within-group variation is captured by modeling each group's outcome as a function a primary covariate (eg: drug concentration is modeled as a non-linear function of time). This results in separate parameter estimates for each group. In the second stage, the between-group variation is modeled by regressing these parameter estimates on some relevant group-varying covariates (fixed effects) and random effects on parameters which differ significantly between groups. These models still make the assumptions that both the residual errors and random effects are normally distributed with mean zero. While usually these models also make the assumption that the residual errors are uncorrelated with a constant variance, these assumptions can be relaxed by explicitly modeling the error covariance structure.

This report seeks to demonstrate how non-linear mixed effects models can be used to model complex relationships often seen in pharmacokinetic data - highlighting the main statistical considerations therein. Specifically, the analysis aims to evaluate whether there are differences between the study-arms with respect to the participants' pharmacokinetic mechanisms. The report proceeds as follows: a brief description of the data is provided; a thorough exploratory data analysis is presented; the model building procedure, and the statistical methods underlying it, are carefully laid out whilst providing results to illustrate the impact of different components of the model; some model diagnostics are given to assess the fit of the final model; methods to relax some of the assumptions of the model - by explicitly modelling the residual covariance structure - are demonstrated; a brief discussion and conclusion are provided.

2 Data Description

The data used in this report comes from a randomized clinical trial to compare the efficacy between two drug regimens in treating malaria. Participants were recruited at one of four sites in Mozambique and randomly assigned to one of two treatment arms: sulfadoxine & pyrimethamine (SP); and SP with artemisinin (SP/ACT). These subjects were then followed up at 9 time points over the next 42 days, and outcomes were recorded at each instance. These outcomes related to the prevalence of parasites in blood, drug concentration levels and haemoglobin counts. Of primary interest in this report are the concentrations of Pyrimethamine and Sulfadoxine in the blood plasma over the course of the study. The data also included selected covariates for each subject (sex; age at baseline; weight at baseline) as well as the site where the participant was enrolled. 408 unique participants were included in the original data.

3 Exploratory Analysis

3.1 Data Validation

Before exploring distributions and summary statistics of the variables of interest, the validity and consistency of the data needs to be investigated. This process, and the impact on the sample size, is illustrated in Figure 1. Firstly, all participants with non-zero drug concentrations at baseline were removed from the dataset - reducing the sample size to 296 participants. This was done as it is considered an assumption of the first-order compartment model that the drug concentration at baseline is zero and thus retaining these values would likely distort the between-subject variation with respect to the model parameters - if the model could fit to the data at all. This also removed all participants who had missing baseline values. Since the first order compartment has three parameters, it requires at least four observations to be estimated. Therefore, in the next step only participants with more than three non-missing measurements for both drugs, were retained - reducing the sample size to 271 subjects. Lastly, one subject was missing an important covariate - their weight measurement - and was removed to avoid complicating the modeling procedure - where weight may be a significant factor. This process thus reduced the sample size to 270 participants, where each had complete covariates, and at least four observations for each drug concentration - including a zero value at baseline.

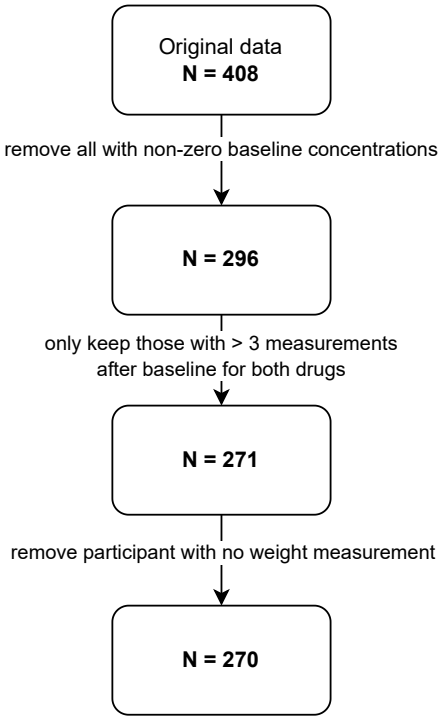


Figure 1: Sample size flow diagram

3.2 Descriptive Statistics

Table 1 provides some summary statistics of baseline continuous variables, stratified by study arm. There were 135 participants in the SP and SP/ART arms, respectively. The study arms were comparable with respect to their baseline weight measurements, with both distributions being skewed to the right (mean median). The weights of the SP/ART arm were slightly larger than those of the other group on average. In terms of age, the distributions are centered around about 15 and 16 years respectively, and right-skewed, 50% of participants under the age of 10 and 11, respectively. The SP/ART group was thus slightly older on average.

Table 1: Continuous Variables by Study Arm

Variable	SP(n=135)	SP/ART (n=135)
Weight		
Min.	10.00	10.00
1st Qu.	15.00	15.00
Median	27.00	26.75
Mean	33.68	33.34
3rd Qu.	52.50	55.00
Max.	99.00	84.00
Age		
Min.	2.00	2.00
1st Qu.	5.00	5.00
Median	10.00	11.00
Mean	15.37	16.29
3rd Qu.	23.00	23.50
Max.	63.00	65.00

Table 2 below shows the numbers and proportions of each level of sex and study site for the sample, stratified by study arm. The proportions of each sex were similar across the study arms, although the SP arm had a greater proportion of females (60.0% vs 55.6%). The participants from each study arm were not identically distributed across the study sites. A disproportionately large amount of participants in both arms were treated at Magude, while a very small number were treated at Catuane. A much larger proportion of SP/ART participants were treated at the Boane site when compared to the sites associated with the SP group. Consequently, the proportion of SP participants treated at each of the other sites was slightly greater. It may thus be important to account for the multi-center design of the study in the analysis, especially if between-site differences are expected or hypothesized. It is not shown here, but the study year and country for all participants was 2003 and Mozambique, respectively.

Table 2: Discrete variables by Study Arm

Variable	Level	SP (n [%])	SP/ART (n [%])
Sex	Male	54 [40.0%]	60 [44.4%]
	Female	81 [60.0%]	75 [55.6%]
Study site	Boane	29 [21.5%]	42 [31.1%]
	Catuane	12 [8.9%]	10 [7.4%]
	Magude	73 [54.1%]	66 [48.9%]
	Namaacha	21 [15.6%]	17 [12.6%]

3.3 Cross-sectional Profiles

In this section, the empirical distributions of the concentration of each drug across the participants at each visit date are explored. Figure 2 shows the distributions of Pyrimethamine concentrations at each of the study days, for the sample stratified by study arm. The baseline visit is not shown since all values are zero at this time point. The distribution is relatively symmetrical across the first three study days, but as the study progresses it becomes dominated by very small values. This is an intuitive result, as one would expect to see sequentially lower drug concentrations as time passes from the dosage and the drug is cleared from the body. This can be seen across the entire study window, where even over the first three days the center of the distribution seems to shift leftwards - from around 250 units to less than 200 units. The final three study visits are clearly categorized by almost all near-zero Pyrimethamine concentrations, with some outlying higher values. There do not seem to be clear differences between the groups with respect to the shape of the distributions at each visit. However, there could be differences in the rates of change that would be very difficult to see with these plots.

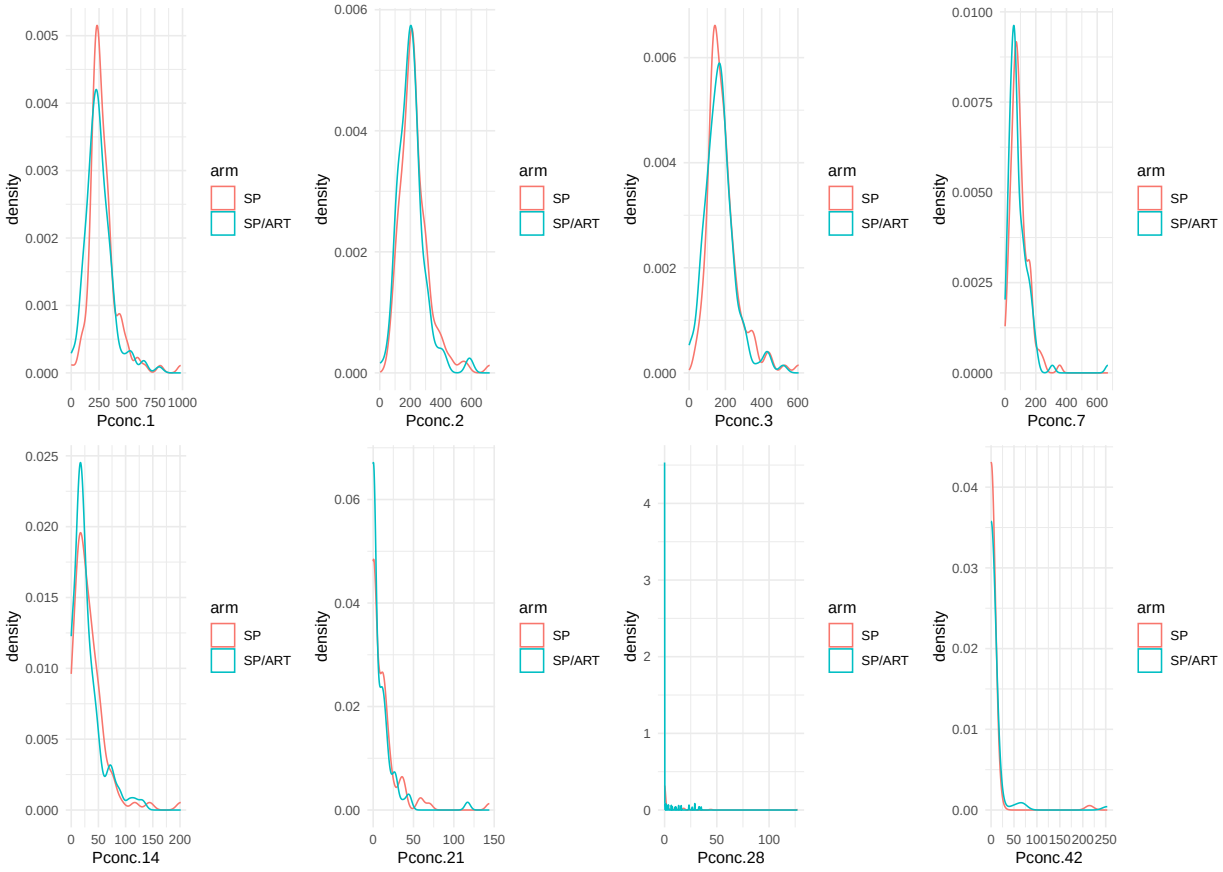


Figure 2: Distribution of Pyrimethamine concentrations at each visit - by study arm

Considering the distribution of Sulfadoxine in Figure 3, one can make some similar observations as above. The distributions are also more symmetrical to begin with, shifting leftwards at each subsequent visit, and becoming dominated by very small values in the final few visits. Interestingly, the Sulfadoxine concentrations do not become so extremely right-skewed as early as with the Pyrimethamine concentrations. Where almost all of the subjects have near-zero Pyrimethamine concentrations by visit 21, there remain a much larger and more dispersed proportion of subjects with non-zero Sulfadoxine concentrations at these latter visit dates. This may have something to do with Sulfadoxine having a longer half-life than Pyrimethamine, however this is merely evidence-based conjecture and would need to be confirmed with clinical input. As

with Pyrimethamine, there are no substantial between-arm differences in the shapes of the distributions up to day 7. Between days 14 and 28, however, there do seem to be differences between the study arms. The SP/ART arm's concentrations seem to approach zero at a faster rate than the SP arm, perhaps signalling a difference in the average rate of elimination and/or clearance of the drug between the two groups.

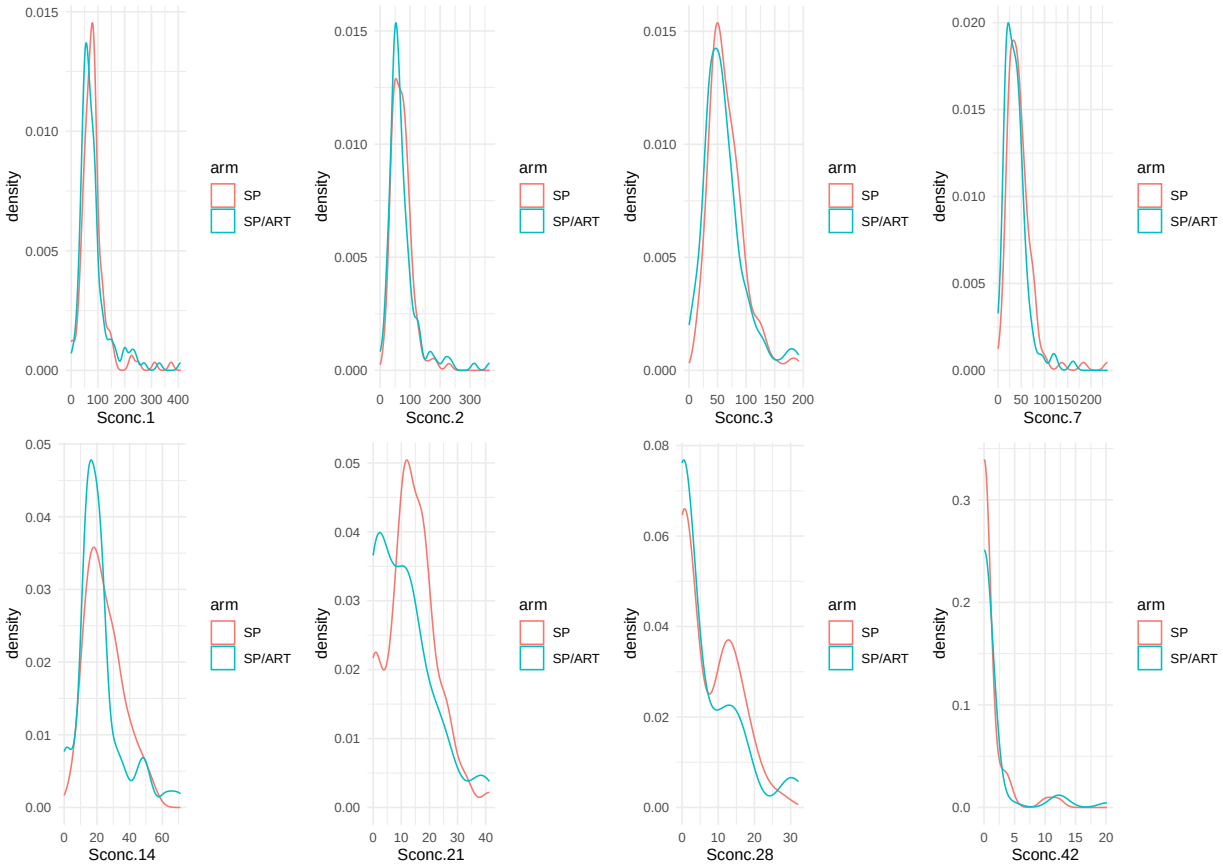
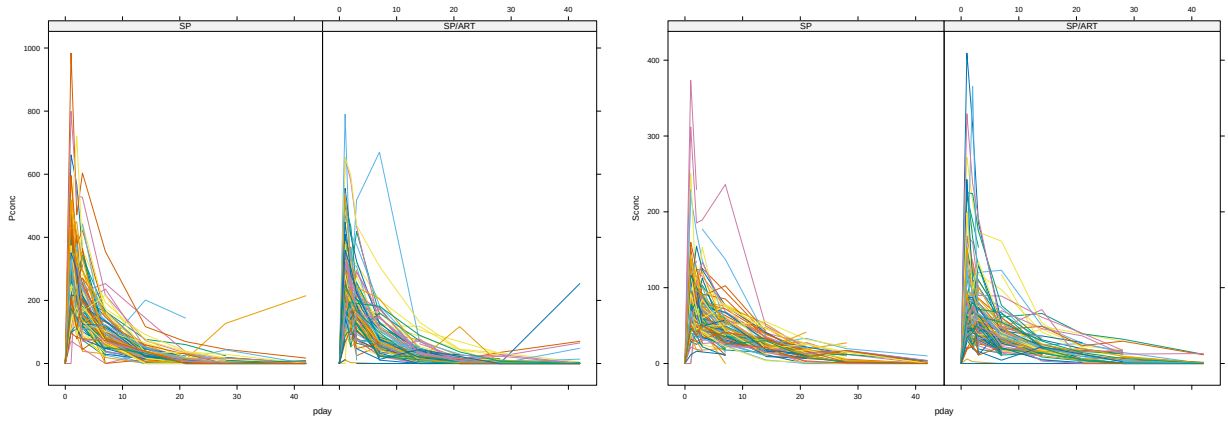


Figure 3: Distribution of Sulfadoxine concentrations at each visit - by study arm

3.4 Longitudinal Profiles

In this section of the report, the longitudinal profiles of the drug concentrations are explored. First individual profiles are illustrated, followed by mean profiles and finally the profiles of the variance of the concentrations are explored.

Figures 4a and 4b illustrate the drug concentrations for each of the 270 participants at each study day for Pyrimethamine and Sulfadoxine, respectively. Each plot is stratified by study arm, such that the left panel shows the participants in the SP arm. The first observation is that there is clearly an underlying trend with respect to the change in drug concentrations over the study window. Almost all of the subjects seem to experience an initial rapid rise in drug levels, followed by steady reduction - which seems to tend towards zero. The second observation here is that there is significant between-subject variation, which seems to be greatest around the first 2 or 3 visits, and is lessened over time as all subject's drug concentrations appear to near zero. There also appear to be a small number of participants whose profiles do not follow the general trend. There is one participant in each of the study arms whose Pyrimethamine levels rise towards the end of the study. There are also a few participants in the SP/ART arm whose drug levels appear remain near zero over the first few study days. Lastly, there also appear to be a small number of participants whose profiles reflect a 'double-peak' over the first seven study days, rather than the single-peak that most of the trajectories seem to follow.



(a) Pyrimethamine

(b) Sulfadoxine

Figure 4: Individual longitudinal profiles - stratified by study arm

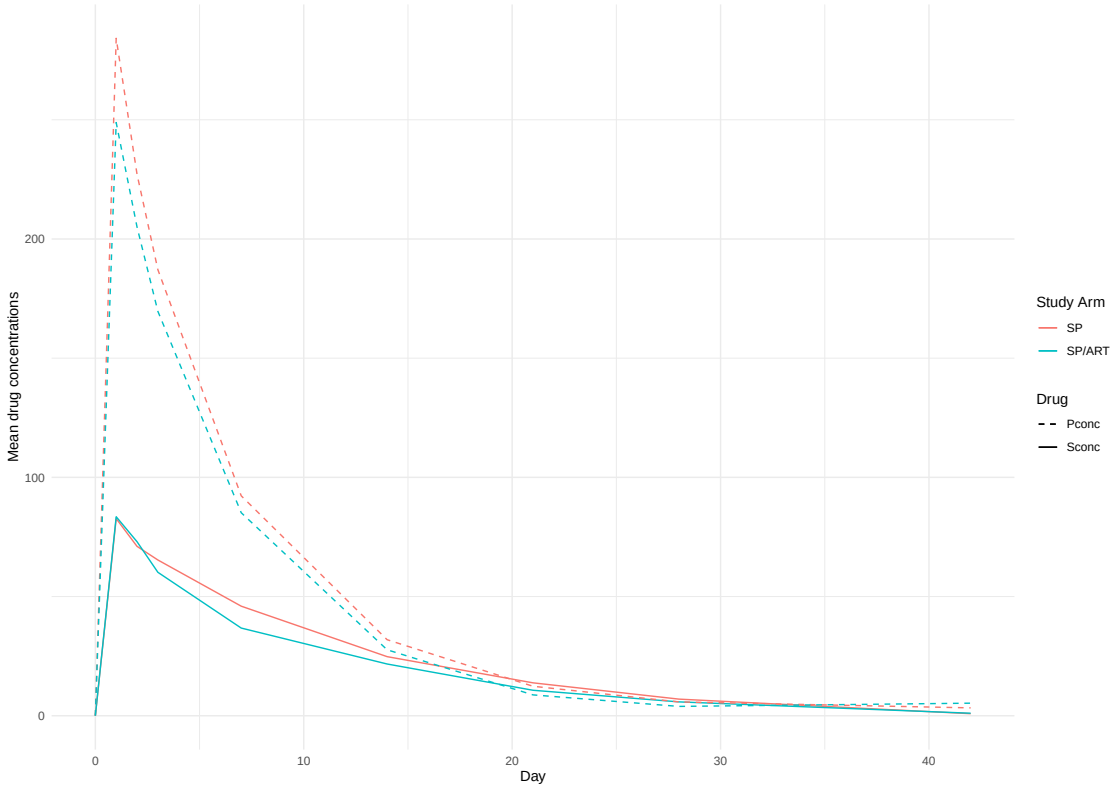


Figure 5: Mean longitudinal profiles of drug levels - by study arm

While the individual profiles provide some important insight into between-subject variation, they are rather messy and it is thus difficult to identify the average trend and compare this between sample groups (such as the study arms). Figure 5 thus plots the mean drug concentrations at each study day, for the sample stratified by study arm. The general trend is now much clearer, where one can see a rapid rise in drug concentration from zero at baseline which peaks at day one. The maximum average concentration of Pyrimethamine is much greater than that of Sulfadoxine, and it also is cleared much more rapidly over the following few days until the average concentrations of both drugs are approximately equal by day 21. Considering between-group differences in Pyrimethamine levels first, the SP arm appears to peak at about 30 units higher than the SP/ART arm and this difference in level is sustained until around day 28. This may indicate that the groups differ with respect to absorption of the drug but there is less visual evidence to suggest a difference between the elimination and clearance of the drug. In contrast, the peak Sulfadoxine levels do not seem to differ between the study arms, however the SP/ART arm seems to clear the drug at a faster rate than the SP arm - before the levels become similar again around day 28.

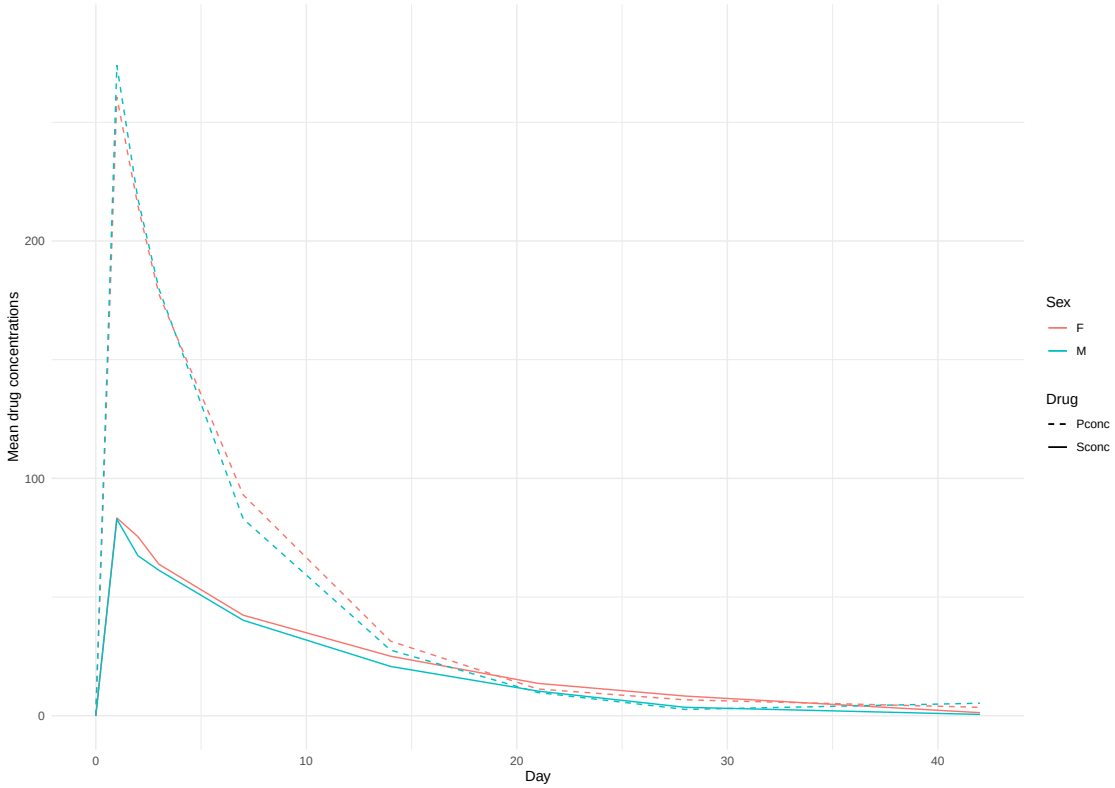


Figure 6: Mean longitudinal profiles of drug levels - by sex

To investigate whether a subject's sex has an impact on the pharmacokinetic process of these drugs, Figure 6 illustrates the mean profiles for the sample stratified by sex. Considering Pyrimethamine first, the average male's concentration peaks slightly higher than that of the females, and the drug seems to be removed from the system at a faster rate - such that the females' drug levels are consistently higher than the males from day 7 onwards. In terms of Sulfadoxine levels, there does not seem to be a difference in the average peak levels but again it seems that the drug is eliminated from the system at a faster rate for the males.

Figure 7 similarly plots the mean drug levels for the sample, stratified by weight quartile. Considering Pyrimethamine, the subjects in the first and last weight quartile seem to have similar average peak concentrations which are markedly higher than that for the subjects in the middle two weight quartiles. This may imply that participants with more extreme weights tend to have slower rates of absorption. In terms of clearance of the drug from the system, those in the first and second quartile seem to eradicate the drug much faster than those with a higher relative mass: by day 14 there is on average much lower concentrations seen in the first two weight quartiles than the last two. Considering the Sulfadoxine levels, once again the highest average peaks are seen with respect to the middle two quartiles - however the peaks are notably different for all four weight quartiles. The patterns of elimination/clearance rates correspond with those seen for Pyrimethamine - where the subjects with relatively less mass appear to remove the drug from their systems at a faster rate - such that by day 28 the average drug levels seem to increase with weight quartile.

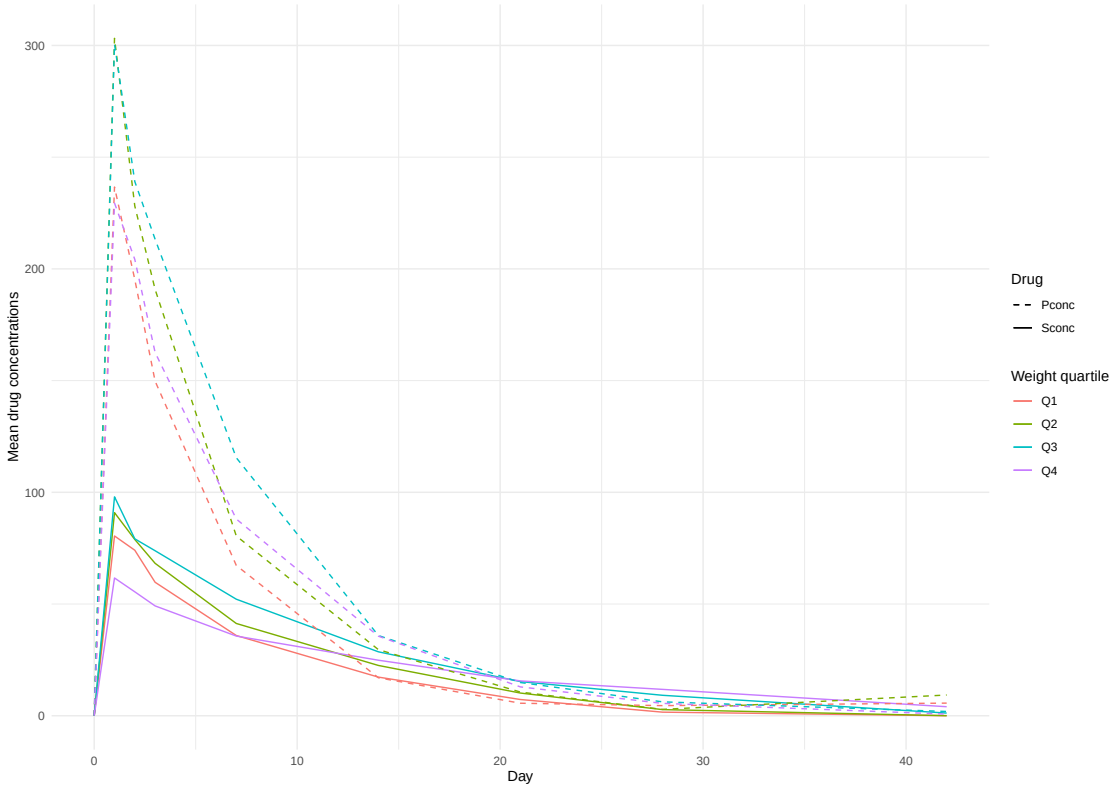


Figure 7: Mean longitudinal profiles of drug levels - by weight quartile

Lastly, Figure 8 plots the sample variance for each drug at each visit day - stratified by study arm. Firstly, as was noted from Figure 4, there is relatively more variability when the drug concentrations peak (day 1), which diminishes afterwards. This shows that the drug concentrations are certainly heteroskedastic, and thus may need to be accounted for in the modeling procedure to avoid yielding biased parameter estimates (by explicitely modeling the variance as a function of study day). What can be noted here also is that there is much more variability around the Pyrimethamine concentrations than the Sulfadoxine concentrations. This is not surprising, since the Pyrimethamine concentrations are measured on a larger scale than the other drug and thus the scale of the variance is larger. Lastly, the plots may suggest a degree of dependence between the study arms with respect to the variability - since there are subtle differences in the variance profiles when comparing the two arms. This may necessitate stratifying the variance function by st§udy arm.

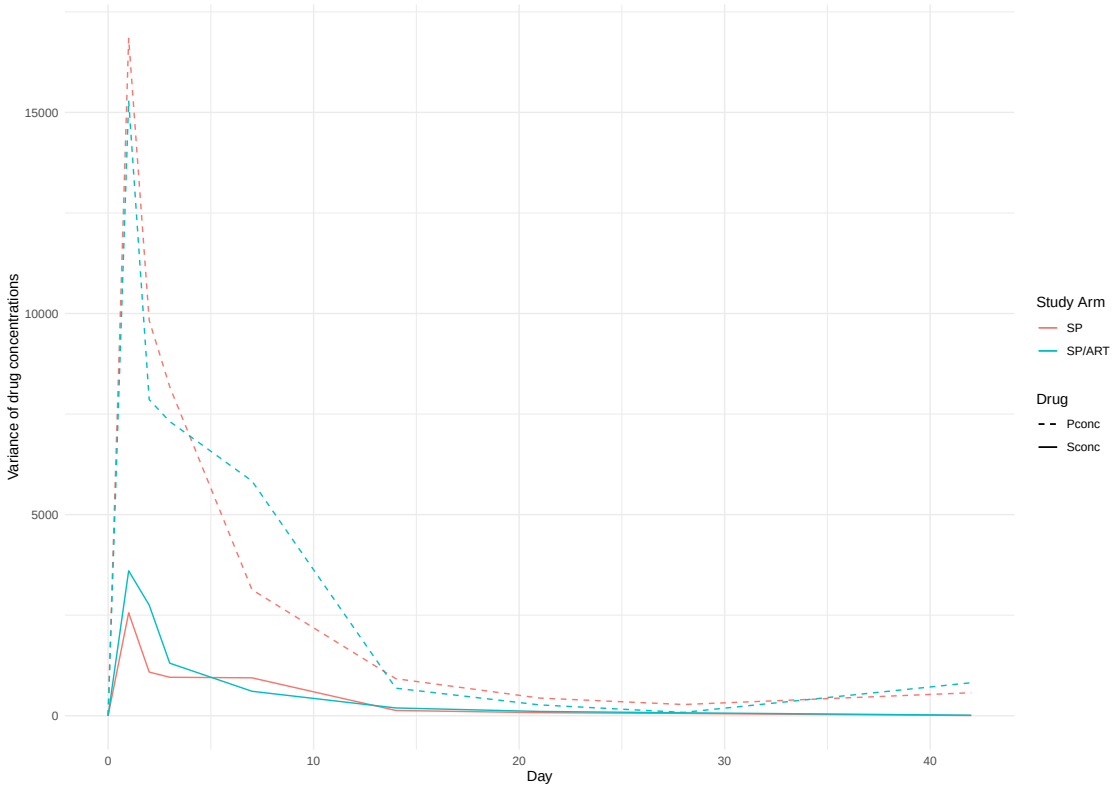


Figure 8: Longitudinal variance profiles of drug levels - by arm

4 Methods and Results

Prior to building the model, a final data management step was taken. This was to convert all zero values to $1e^{-4}$. Without this step, the model estimation procedure would not converge. This is deemed an acceptable modification which represents reality, since it is likely that the zero values represent a present but undetectable level of the drugs in the circulatory system. $1e^{-4}$ was chosen as an appropriately small value since it is much less than the actual minimum detected values of 0.146 (Pyrimethamine) and 0.125 (Sulfadoxine).

In this report, the exploratory analysis is used to guide the model building procedure to an extent. Firstly, the nature of the data and the shape of the longitudinal profiles informs the type of non-linear model which is most appropriate. In this case, a first-order compartment model should be sufficient to capture the absorption of the drugs from the digestive system and the subsequent elimination from the circulatory system. More specifically, this mechanistically models the drug concentration with an absorption, elimination and clearance rate. Clearance is a concept distinct from but connected to elimination, and is defined as the volume of blood plasma cleared of the drug per unit time - and thus accounts for the elimination of the drug and the rate of blood flow between organs [3]. This first-stage model thus captures the within-subject variation. This model is fit to each subject's measurements, and the resulting parameter estimates form the input into the second-stage, where the between-subject variance is modeled as a function of fixed and random effects.

The full stage-1 model formulation is thus given by:

$$y_{ij} = D\frac{\exp(\phi_1 + b_{1i})\exp(\phi_2 + b_{2i})}{\exp(\phi_3 + b_{3i})[\exp(\phi_2 + b_{2i}) - \exp(\phi_1 + b_{1i})]}[\exp(-\exp(\phi_1 + b_{1i}) * x_j) - \exp(-\exp(\phi_2 + b_{2i}) * x_j)] + \epsilon_{ij} \tag{1}$$

where y_{ij} represents the i-th subject's drug concentration on day j; the dosage, D, is constant and set to 1 in this case; x represents the study day; ϕ_1 is the logarithm of the mean elimination rate; ϕ_2 is the logarithm of the mean absorption rate; ϕ_3 is the logarithm of the mean clearance rate; $\boldsymbol{b_i}$ are the subject-specific random effects - and thus $\boldsymbol{\phi} + \boldsymbol{b_i} = \boldsymbol{\phi_i}$ is the subject-specific parameter vector; and $\epsilon_{ij} \sim N(0, \sigma^2)$; $\boldsymbol{b_i} \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$.

More generally, this can be expressed as:

$$\boldsymbol{y_i} = \boldsymbol{f_i}(\boldsymbol{\phi_i}, \boldsymbol{v_i}) + \boldsymbol{\epsilon_i} \tag{2}$$

where $\boldsymbol{v_i}$ is a vector of between-subject varying covariates; and $\boldsymbol{f_i}$ defines the first-order compartment model in equation (1).

The parameter vector, $\boldsymbol{\phi_i}$, is then modeled as:

$$\boldsymbol{\phi_i} = \boldsymbol{A_i}\boldsymbol{\beta} + \boldsymbol{B_i}\boldsymbol{b_i} \tag{3}$$

where $\boldsymbol{A_i}$ and $\boldsymbol{B_i}$ are the design matrices for the fixed and random effects, $\boldsymbol{\beta}$ and $\boldsymbol{b_i}$, respectively; and $\boldsymbol{b_i} \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$.

The model is fit to the data by the iterative likelihood linear mixed effects approximation method proposed by Lindstrom and Bates [5], which alternates between a penalized non-linear least squares (PNLS) and LME step to estimate the parameters. Going forward this estimation procedure will be referred to simply as the LME approximation, or the restricted LME (RLME) approximation. There are a number of control parameters within the PNLS and LME steps which can be manipulated where necessary to ensure convergence in this estimation procedure. Those utilised in this report are: the minimum factor to shrink the step size in the sum of squares minimization in the PNLS step ($minScale$); and the overall tolerance for the convergence criteria.

These models can be further extended to incorporate multiple levels of nesting. For example, some variation in the response may be attributable to differing characteristics between the study sites - implying the necessity of additional random effects. The vector of drug concentrations for the i-th subject in study site k ($\boldsymbol{y_{ik}}$) could thus be modeled in two stages as:

$$\begin{aligned} \boldsymbol{y_{ik}} &= \boldsymbol{f_{ik}}(\boldsymbol{\phi_{ik}}, \boldsymbol{v_{ik}}) + \boldsymbol{\epsilon_{ik}} \\ \boldsymbol{\phi_{ik}} &= \boldsymbol{A_{ik}}\boldsymbol{\beta} + \boldsymbol{B_{i,k}}\boldsymbol{b_i} + \boldsymbol{B_{ik}}\boldsymbol{b_{ik}} \end{aligned} \tag{4}$$

where $\boldsymbol{B_{i,k}}$ and $\boldsymbol{B_{ik}}$ are the design matrices for the random effects at the subject-within-site level ($\boldsymbol{b_i} \sim N(\boldsymbol{0}, \boldsymbol{\Psi_1})$) and the site-level ($\boldsymbol{b_{ik}} \sim N(\boldsymbol{0}, \boldsymbol{\Psi_2})$), respectively; and $\boldsymbol{\epsilon_{ik}} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

The assumption of homoskedasticity among the residual errors can be relaxed by explicitly modeling the variance as a function of the mean ($\mu ij = E[y_{ij}|b_i]$), a vector of covariates (v_{ij}), and a vector of variance parameters (δ):

$$var(\epsilon_{ij}|b_i) = \sigma^2 g^2(\mu_{ij}, v_{ij}, \delta) \tag{5}$$

where $g()$ models the standard deviation of the residuals with some variance function.

Before fitting the nlme model, the structure of the random effects component needs to be determined. It is therefore beneficial to first model to each subject's concentration curves separately as:

$$y_{ij} = D\frac{\exp(\phi_{1i})\exp(\phi_{2i})}{\exp(\phi_{3i})[\exp(\phi_{2i}) - \exp(\phi_{1i})]}[\exp(-\exp(\phi_{1i}) * x_j) - \exp(-\exp(\phi_{2i}) * x_j)] + \epsilon_{ij} \tag{6}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$.

4.1 Sulfadoxine Results

4.1.1 Model building

The above model was fit to the Sulfadoxine data by non-linear least squares, as proposed by Bates and Chambers [1], using the *nlsList* R function, where starting values are produced automatically within the

'self-starting' function. This allows one to examine the between-subject variability of the parameter estimates - to guide which parameters likely need random effects.

The estimation procedure managed to fit this model to 30 out of the 270 individual curves, yielding 30 subject-specific estimates for $\boldsymbol{\phi_i}$. These estimates along with their associated 95% confidence intervals are displayed in Figure 9. The correspondence with the parameters in equations (1) and (6) is: $\phi_1 = \text{lKe}$ (log-elimination rate); $\phi_2 = \text{lKa}$ (log-absorption rate); and $\phi_3 = \text{lCl}$ (log-clearance rate). These estimates suggest that there is much between-subject variability in the log-clearance rates - where the intervals do not overlap much. Similar observations are made for the log-elimination rate. There is more overlap between the log-absorption rates, but still many estimates which do not overlap. Altogether, there is enough between-subject variability among all three coefficients to suggest that random effects may initially be necessary for all of them - and certainly for the clearance parameter.

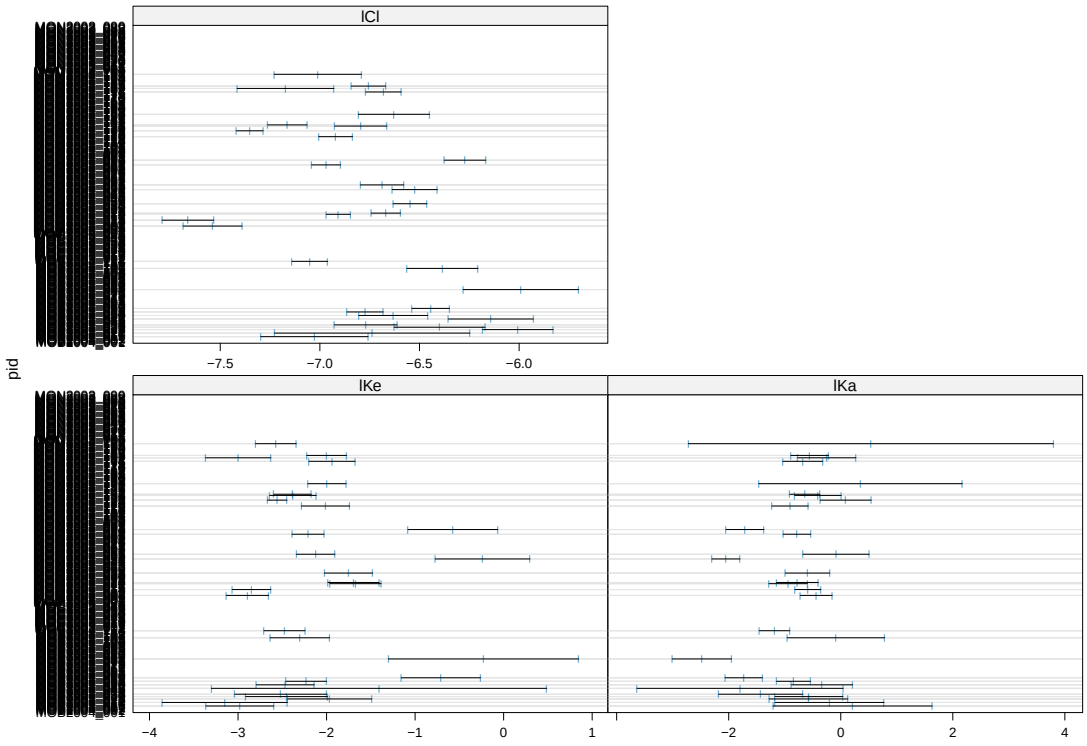


Figure 9: Parameter estimates for individual sulfadoxine non-linear models

Another use of this modelling step is to investigate the structure of the random-effects covariance matrix, $\boldsymbol{\Psi}$. This can be explored by looking at the pairwise correlations between the estimates for $\boldsymbol{\phi_i}$, which are plotted in Figure 10. These suggest that there may be a negative association between the log-absorption rate and both the other parameters, and a positive association between the log-clearance and log-elimination rates. The most evident correlation seems to be that between the absorption and elimination parameters, which perhaps reflects the fact that the elimination rate is mechanistically a function of the absorption rate.

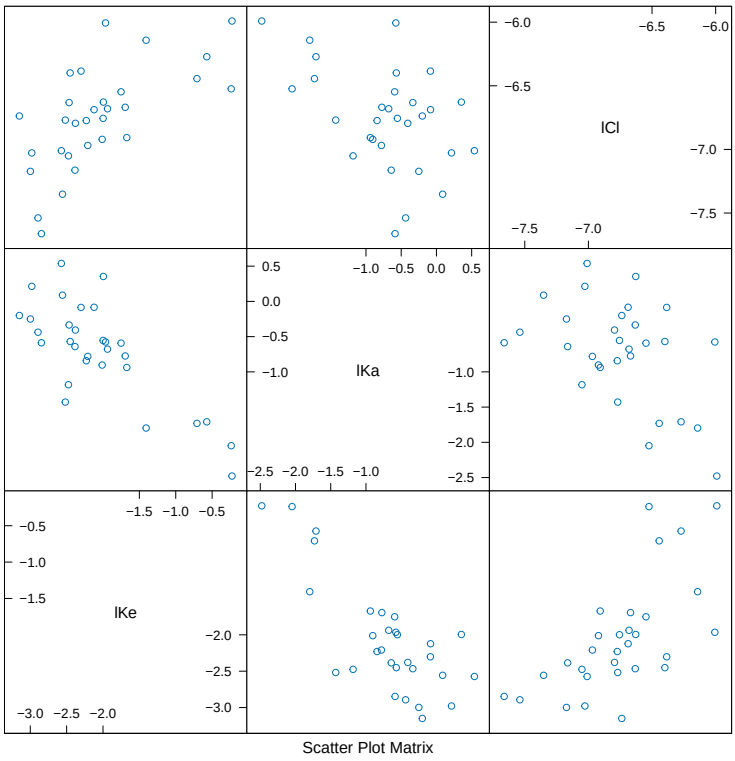


Figure 10: Pairwise correlations between sulfadoxine individual parameter estimates

Lastly, the residuals of these models should be briefly investigated to confirm the underlying assumption of homoskedastic, normally distributed residual errors with mean zero. Figure 18 in Appendix A shows that the residuals are indeed symmetrical and centered around zero, although with a number of outlying values and much less dispersion than a normal distribution.

Overall, the individual models indicate that it is wise to first specify the model with random effects on all three parameters, and to also allow these parameters to be correlated with a general positive-definite structure. This allows each of the off-diagonal entries of the $\mathbf{\Psi}$ matrix to be unique and non-zero. This model was fit to the data using the RLME approximation procedure, with $minScale = 1e^{-20}$ and $tolerance = 0.1$. The model was then refit with a diagonal covariance structure (i.e. zero values for the off-diagonals in $\mathbf{\Psi}$). The two models were then compared according to the goodness-of-fit metrics in Table 3 - where S1a refers to the model with the positive-definite structure for $\mathbf{\Psi}$. This table suggests that the simpler model, with the diagonal random effects structure, provides a better fit to the data along both the information criteria and the log-likelihood.

Table 3: Model Comparison: Positive-definite vs. Diagonal RE Covariance

Model	df	AIC	BIC	logLik	L.Ratio	p-value
S1a	10	15988.15	16043.11	-7984.08	-	-
S1b	7	15720.23	15758.70	-7853.11	261.92	<.0001

The estimated diagonal elements of $\mathbf{\Psi}$ associated with model S1b were $(lKe, lKa, lCl) = ((8.56e^{-5})^2, 0.81^2, 0.46^2)$. The relatively small variance of the lKe random effect suggests that this may not be necessary. The model was thus refit with random effects only on the log-absorption and log-clearance parameters (and a diagonal covariance structure). This is compared to model S1b in Table 4. Here the reduction of the log-likelihood

is not conclusive, however the likelihood ratio test is borderline significant. This, along with the smaller information criteria values, subjectively provides sufficient evidence to reject model S1b in favour of the model with no lKe random effect (S1c).

Table 4: Model Comparison: Random effects on all parameters vs. no lKe random effect

Model	df	AIC	BIC	logLik	L.Ratio	p-value
S1b	7	15720.23	15758.70	-7853.11	-	-
S1c	6	15714.62	15747.59	-7851.31	3.61	0.057

This concludes the determination of the random effects part of the model. The next step is to determine the structure of the fixed effects component. Since it is of primary concern to quantify differences between the study arms, this is automatically included as a fixed effect. Furthermore, the exploratory analysis revealed that there may be between-sex differences in the elimination/clearance of the drug, and a weight-effect which may modify both the absorption and elimination/clearance rates. Additionally, the association between the random effects from model S1c and covariates can be explored - since if the random effects are strongly associated with a covariate, this implies that some variation which is being attributed to subject-specific effects should actually be attributed to between-subject differences along these covariates. The Figures 19 and 20 in Appendix A suggest that there are likely no differences along the covariates with respect to the absorption rates, however there does appear to be a strong weight and age effect with respect to the clearance parameter, and perhaps a sex-effect too. There was much less evidence for the sex-effect and thus a decision was taken to focus on the weight-effect. Since age and weight display a strong linear association, with a Pearson correlation coefficient of about 0.83, it is hypothesized that adjusting for weight should remove much of the age-effect.

Models with different fixed effects should only be compared if fitted by the LME approximation (not RLME), and thus model S1c is first refit and all subsequent models in this section of fit with the LME algorithm. The model is first fit with a study-arm fixed effect on all three parameters (S2a). This model is then adjusted by including a weight effect on all three parameters (S2b). These models are compared sequentially in Table 5, which shows that the arm fixed effects in S2a significantly improve on the model S1c, and the weight fixed effects in S2b significantly improve on S2a.

Table 5: Model Comparison: S1c, S2a, S2b, S2c

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
S1c	6	15708.35	15741.34	-7848.18	-	-	-
S2a	9	15694.33	15743.81	-7838.16	S1c vs S2a	20.03	0.0002
S2b	12	15591.66	15657.63	-7783.83	S2a vs S2b	108.67	<.0001
S2c	15	15570.36	15652.83	-7770.18	S2b vs S2c	27.30	<.0001

The association of the random effects from model S2b with weight and age were subsequently explored again, and it was seen that some association still remained. Therefore, a fixed effect for the square of weight was also also added (model S2c). As shown in Table 5, this significantly improved the model fit along all metrics and thus model S2c was selected. The effect of incorporating weight and weight^2 fixed effects can be seen in Figure 21 in Appendix A, where it is clear that this removed the association from the random effects.

Model S2c was thus refit by the RMLE approximation, and the associated output is summarised in Table 6. It must firstly be noted that the intercepts are not necessarily interpretable, since they define the sample averages with $weight = 0$ and there are no subjects with zero weight. This is not necessarily a problem since the objective of the analysis is to investigate/quantify between-arm differences and not to quantify the value of the rates of absorption, etc. With that said, the first notable observation is that there appear to be significant between-arm differences with respect to the logarithms of the rates of elimination and clearance. When comparing the participants in the two arms, those in the SP/ART arm had a log-elimination rate

about 0.18 units greater ($p = 0.0011$) and a log-clearance rate 0.24 units greater ($p = 0.0007$) than those in in the SP arm - ceteris paribus. There was no significant difference in the absorption rates. As suspected, the participants' weights do not seem to significantly influence the absorption rates but perhaps do modify the rates of elimination and clearance. Specifically, an additional unit of weight is associated with: a 0.021 unit reduction in the log-elimination rate ($p = 0.0035$), which is changed by a factor of 0.00012 times the age ($p = 0.1955$); and a 0.026 unit reduction in the log-elimination rate ($p = 0.0021$), which is changed by a factor of 0.00028 times the age ($p = 0.0085$).

Table 6: Output for Model S2c

Parameter	Value	Std.Error	DF	t-value	p-value
lKe.(Intercept)	-1.82	0.12	1523	-15.32	<0.0001
lKe.armSP/ART	0.18	0.06	1523	3.26	0.0011
lKe.weight	-0.021	0.01	1523	-2.92	0.0035
lKe.I(weight^2)	0.00012	0.00	1523	1.30	0.20
lKa.(Intercept)	0.77	0.28	1523	2.72	0.0065
lKa.armSP/ART	0.023	0.14	1523	0.16	0.87
lKa.weight	-0.0039	0.02	1523	-0.23	0.82
lKa.I(weight^2)	0.0000	0.00	1523	-0.00	0.99
lCl.(Intercept)	-6.29	0.14	1523	-45.17	<0.0001
lCl.armSP/ART	0.24	0.07	1523	3.39	0.0007
lCl.weight	-0.026	0.01	1523	-3.09	0.0021
lCl.I(weight^2)	0.00028	0.00	1523	2.64	0.0085

The estimates associated with the random effects covariance matrix were:

$$\mathbf{\Psi} = \begin{bmatrix} 0.622^2 & \\ 0 & 0.412^2 \end{bmatrix}$$

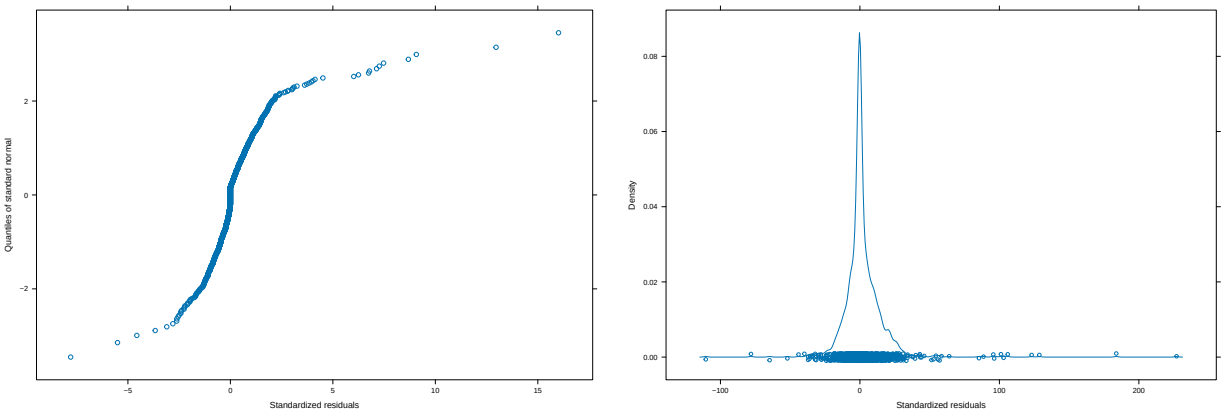
where the first diagonal represents the variance of the random effect on the log-absorption rate and the latter is associated with the log-clearance rate.

The estimate for the variance of the within-subject residuals was very large, at $\hat{\sigma}^2 = 14.18$. This is significantly larger than that of the random effects, and this may suggest that there is an important source of variation in the drug concentrations which was not accounted for in this model.

4.1.2 Model Diagnostics

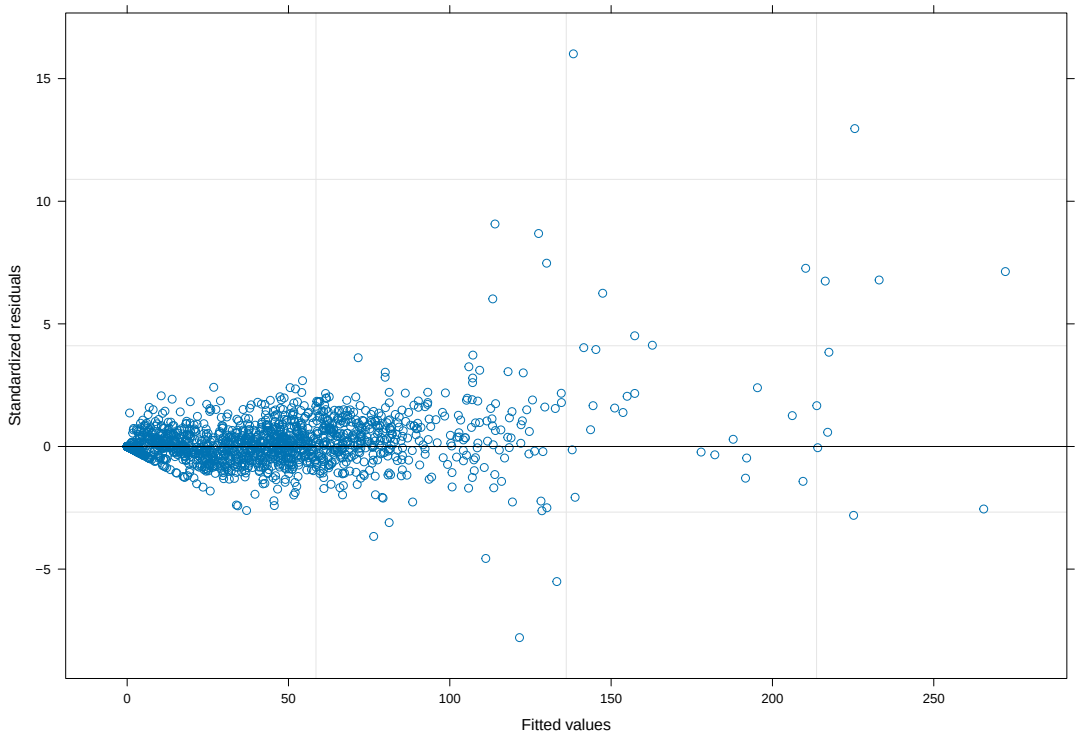
The final fitted model, S2c, must be validated to assess the underlying assumptions: (a) the residual errors are independently distributed and normally distributed around mean zero, with constant variance; (b) the random effects are independent between groups and normally distributed around mean zero with covariance matrix $\mathbf{\Psi}$ [7].

To assess assumption (a), one can inspect the plots in Figure 11. The top panel suggests that while the within-subject residuals are centered around zero, they exhibit less dispersion than that of a normal distribution. The bottom panel shows the variance of the residuals tends to increase with the fitted values, which defies the assumptions of homoskedasticity and independence.



(a) QQ plot

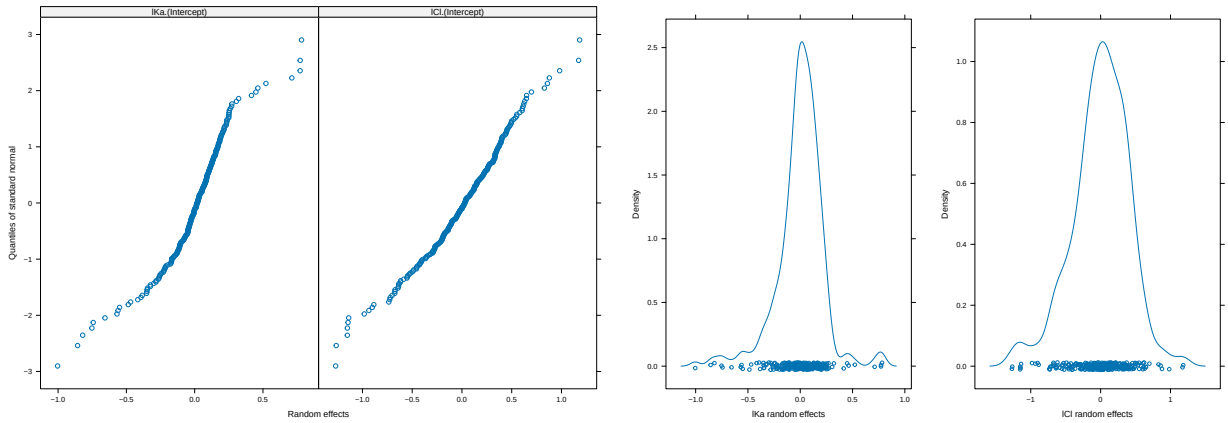
(b) Density plots



(c) Density plot

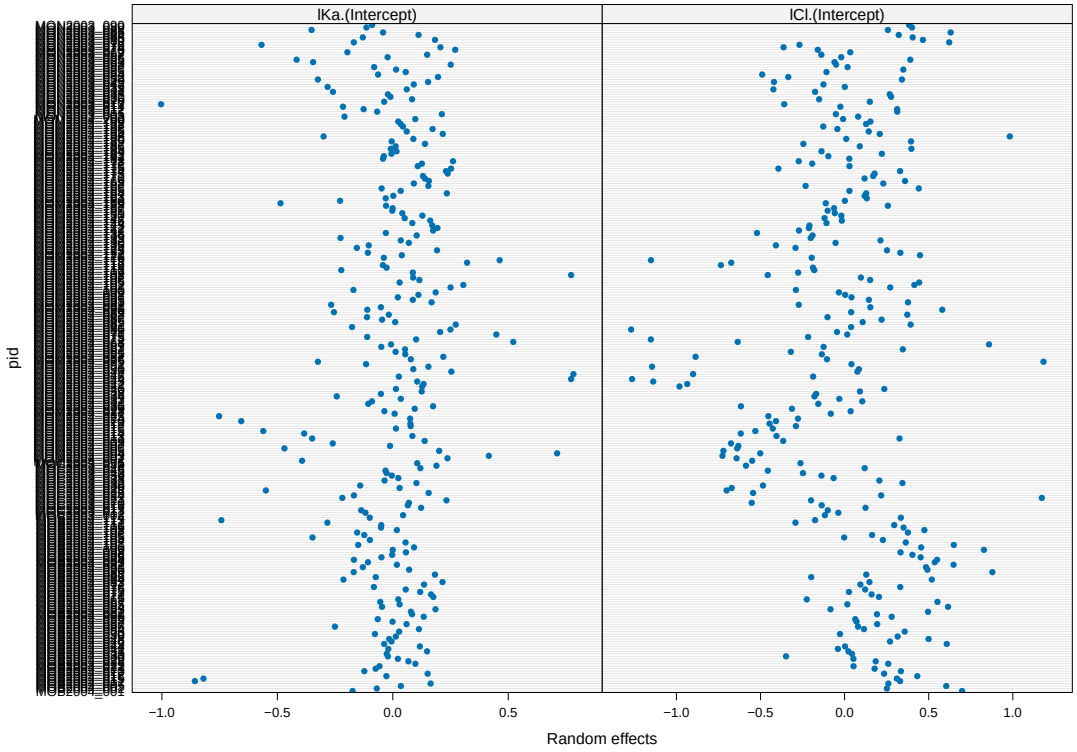
Figure 11: Sulfadoxine: Residual plots for Model S2c

The plots given in Figure 12 help to assess assumption (b). The top panel suggests that the random effects follow a normal distribution imperfectly, but relatively well, and are centered around zero. The bottom plot confirms that the random effects are independent of the subject-level grouping.



(a) QQ plots

(b) Density plots



(c) Residuals against subject ID

Figure 12: Sulfadoxine: Random effects plots for Model S2c

4.1.3 Variance Functions

The heteroskedasticity among the residuals seen in the above model diagnostics suggest that it is necessary to explicitly model the variance of the residual errors as a function of the fitted values, as generalised by equation (5).

The residual errors are thus modeled with an exponential function as:

$$var(\epsilon_{ij}) = \sigma^2 e^{2\delta\mu_{ij}} \tag{7}$$

where $\mu_{ij} = E[y_{ij}|v_{ij}, b_{ij}]$.

Model S2c was thus updated to incorporate the above and fit to the data by the RMLE approximation. The effect on the distribution of the residuals can be seen in Figure 13. The 'fan' shape has been removed, but now the residuals exhibit a funnel shape - where the variance reduces with increasing fitted values. This does not seem to be an valuable improvement on model S2c. An attempt was made to fit other variance functions to improve the model fit, however this was unsuccessful.

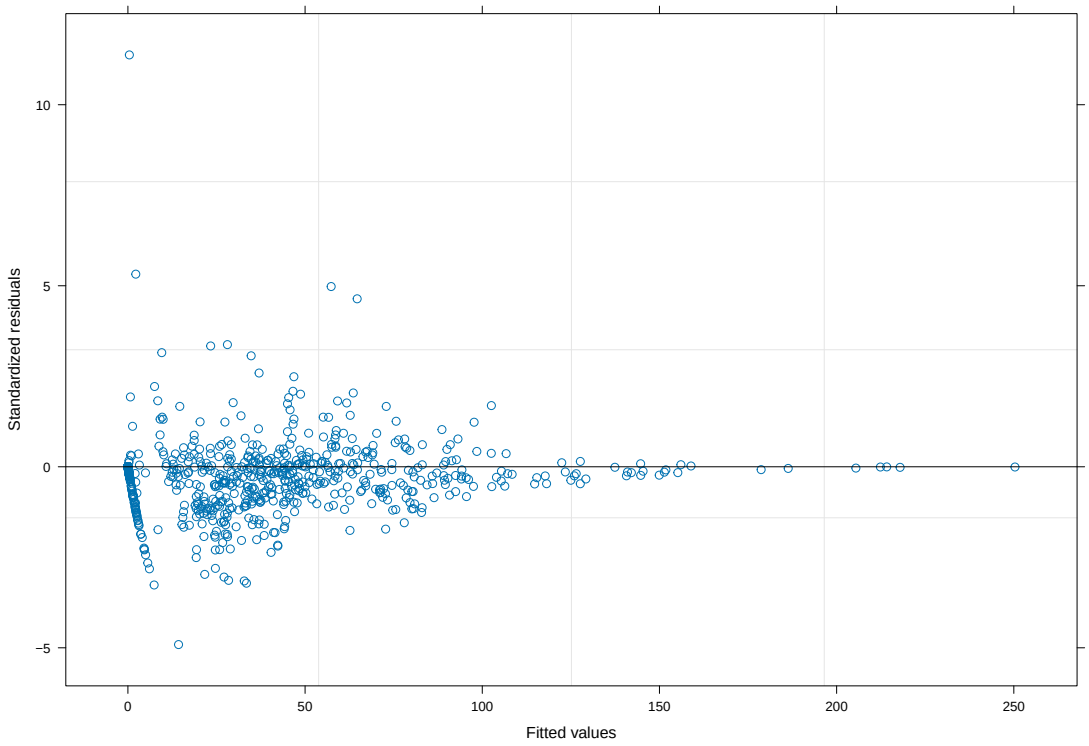


Figure 13: Sulfadoxine: Model S3a residuals vs. fitted values

4.2 Pyrimethamine Results

In this section the results associated with the Pyrimethamine concentrations are presented. The analysis follows a parallel procedure to that shown for the Sulfadoxine modeling, where the random effects structure is determined first, followed by the fixed effects, model diagnostics and then extensions of the model if necessary.

4.2.1 Model Building

As an initial step, the first order compartment model is fit to the individual concentration curves independently. This was successful for 15 out of the 270 participants. The estimates of the log absorption, elimination and clearance rates are shown in Figure 14. There is very little overlap of the estimates for the clearance and elimination parameters, suggesting that random effects will be needed to capture this between-subject variability. The estimates for the log-absorption rate show less variation in this respect. Therefore, if any of these parameters do not need random effects, it would likely be the latter.

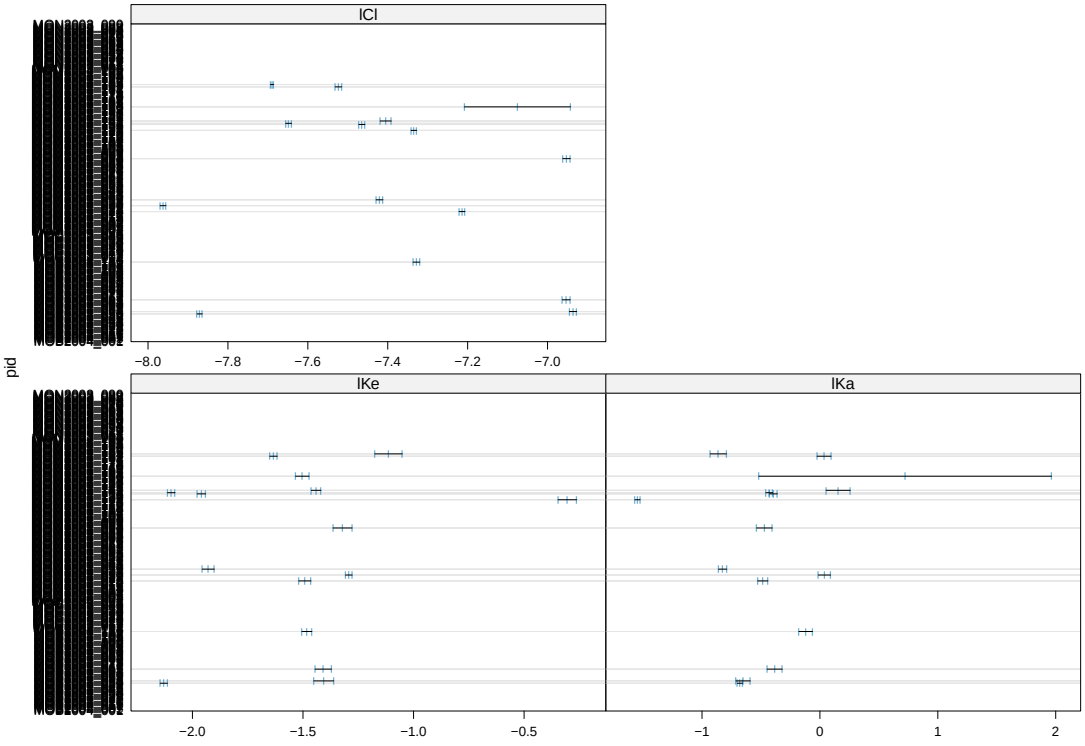


Figure 14: Pyrimethamine: Individual model estimates and 95% intervals

The correlation structure of the random effects is investigated through the pairwise scatterplots in Figure 15. With so few observations, it is difficult to identify a trend. That being said, there does not seem to be any obvious association between the parameter estimates. A diagonal random effects covariance matrix is thus probably the most appropriate structure.

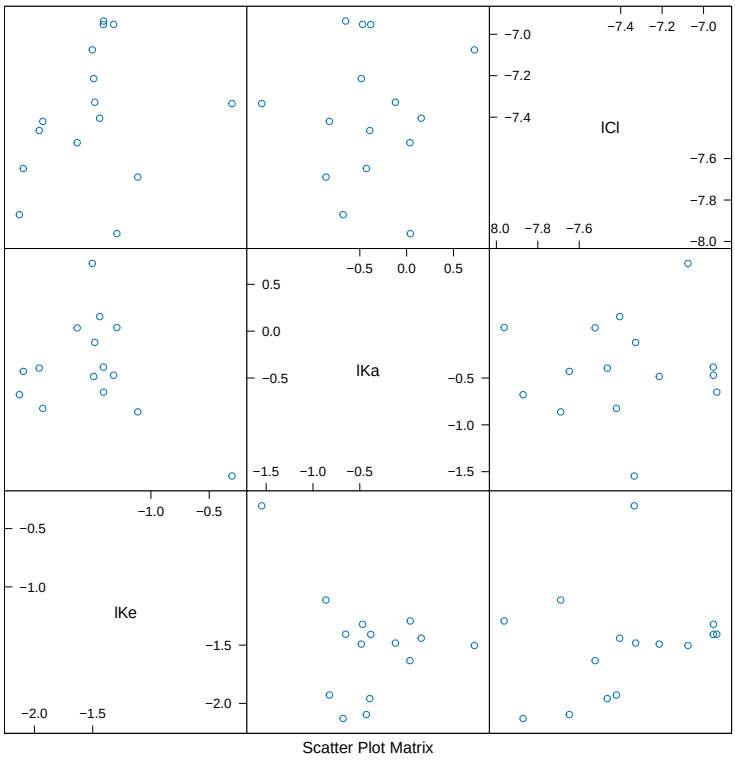


Figure 15: Pyrimethamine: Pairwise parameter scatterplots

Given the above information, the first step was to attempt to fit a model with random effects on all three parameters and a diagonal covariance matrix. This was unsuccessful and the model did not converge, likely because it was over-parameterized. The next logical step was to try each possible combination of two random effects. Of these, the models that successfully fit to the data were: P1a - random effects on log-absorption and log-clearance; P1b - random effects on log-clearance and log-elimination. As can be seen in Table 7, the model fit is better along all three metrics for model P1b when compared to P1a - therefore the model without the absorption random effect is selected. Following this, model P1c was fit - which allows the elimination and clearance random effects to have a positive-definite covariance structure. The likelihood ratio test indicated that there was a significant difference between the models, and thus the simpler model (P1b) was favoured. The information criteria agreed with this result. Lastly, a model with a random effect on only the elimination parameter was fit (P1d). This resulted in an increase in the information criteria and a more negative log-likelihood, and thus the final random effects structure is that represented by model P1b.

Table 7: Model Comparison: P1a, P1b, P1c, P1d, P1e

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
P1a	6	21394.03	21427.52	-10691.02	-	-	-
P1b	6	20714.64	20748.13	-10351.32	-	-	-
P1c	7	20779.68	20818.75	-10382.84	P1b vs P1c	63.04	<.0001
P1d	5	21392.13	21420.03	-10691.06	P1c vs P1d	616.45	<.0001

Next, the above model was refit with the MLE approximation and then study arm was brought in as a fixed effect to modify the model parameters (model P2a). Looking at the plots in Figure 22 and 23, which relate to model P1b, it is clear that additional fixed effects are required to capture the variation in the parameters

attributable to weight and age. Therefore, the next step was to bring in a fixed effect to account for the participants' weights. This model, however, could not fit to the data - yielding singularity errors at each attempt despite manipulations of the estimation algorithm and starting values. Unfortunately, this necessary step was thus unable to be executed.

The final model for the Pyrimethamine concentrations was thus model P2a. It's associated output, with respect to the fixed effects is summarized in Table 8. With respect to the objective of the analysis, the important thing to note here is there were no significant differences between the study arms along any of the first-order compartment parameters.

Table 8: Output for Model P2a

Parameter	Value	Std.Error	DF	t-value	p-value
lKe.(Intercept)	-1.71	0.04	1689	-41.13	<0.0001
lKe.armSP/ART	-0.03	0.06	1689	-0.55	0.5799
lKa.(Intercept)	2.11	5.20	1689	0.41	0.6844
lKa.armSP/ART	-0.08	6.52	1689	-0.01	0.9903
lCl.(Intercept)	-7.45	0.13	1689	-57.70	<0.0001
lCl.armSP/ART	0.11	0.17	1689	0.66	0.5072

The estimates associated with the random effects covariance matrix were:

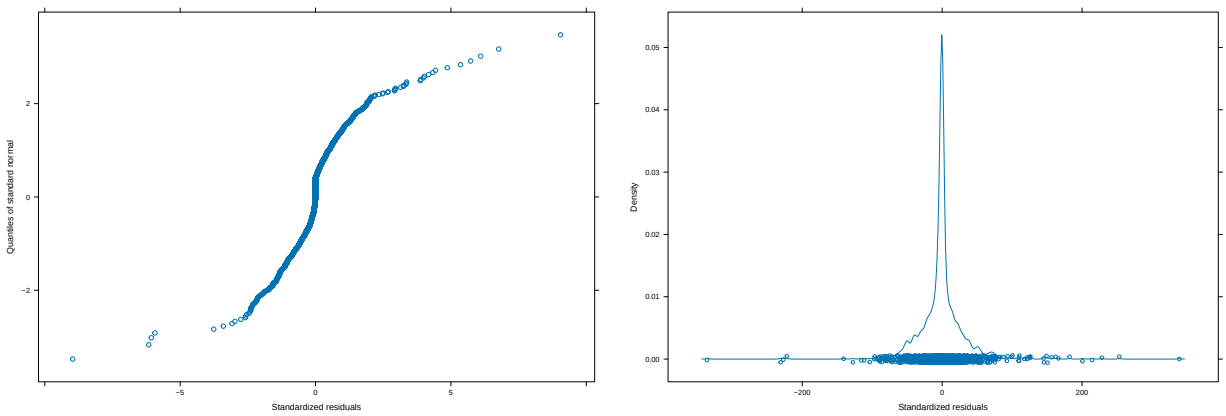
$$\mathbf{\Psi} = \begin{bmatrix} 0.283^2 & \\ 0 & 0.356^2 \end{bmatrix}$$

where the first diagonal represents the variance of the random effect on the log-elimination rate and the latter is associated with the log-clearance rate.

The estimate for the variance of the within-subject residuals was very large, at $\hat{\sigma}^2 = 37.45$. This is significantly larger than that of the random effects, and this may suggest that there is an important source of variation in the drug concentrations which was not accounted for in this model.

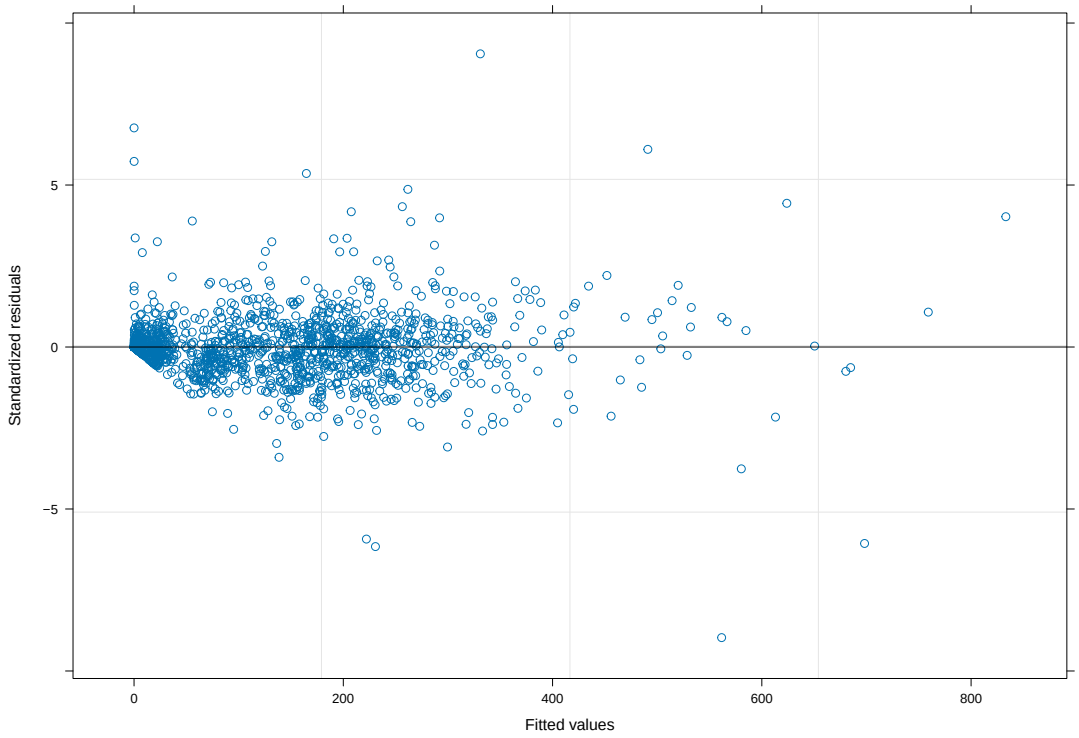
4.2.2 Model Diagnostics

As always, the model assumptions need to be validated. As was seem with the Sulfadoxine model diagnostics, the residuals for this model are centered around zero but not normally distributed (Figure 16). Furthermore, the residuals exhibit heteroskedasticity and dependence along the fitted values. The assumptions surrounding the residuals are thus not satisfied and the model parameter estimates are likely biased to some degree.



(a) QQ plot

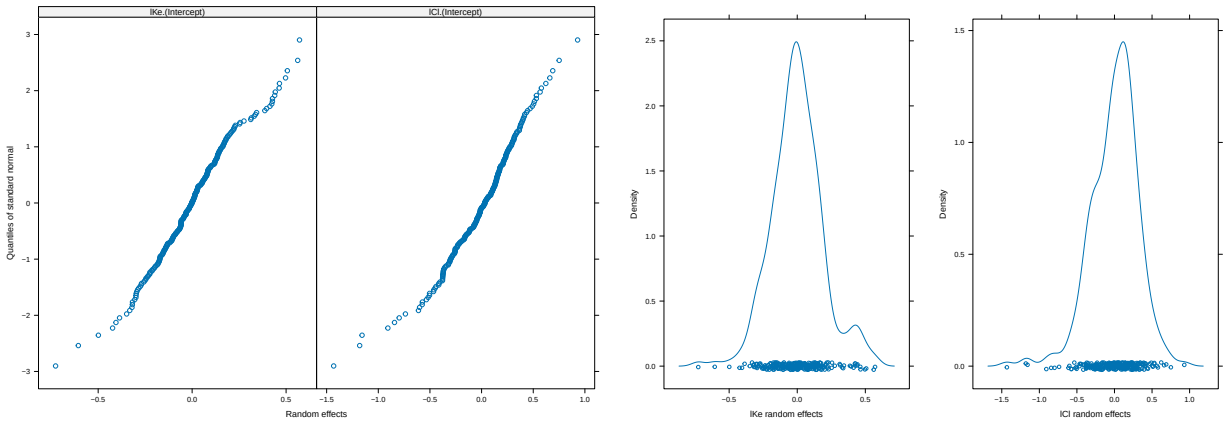
(b) Density plots



(c) Density plot

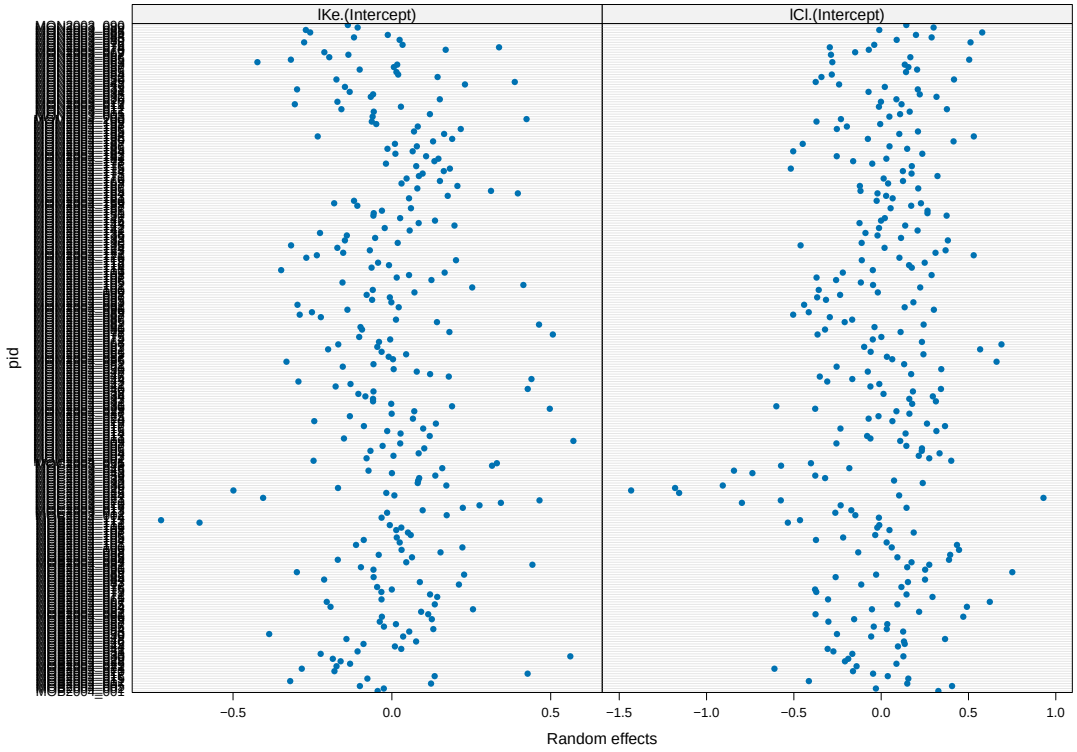
Figure 16: Pyrimethamine: Residual plots for Model P2a

The plots given in Figure 17 help to assess the random effects assumptions. The top panel suggests that the random effects follow a normal distribution imperfectly, but relatively well, and are centered around zero. The bottom plot confirms that the random effects are independent of the subject-level grouping.



(a) QQ plots

(b) Density plots



(c) Residuals against subject ID

Figure 17: Pyrimethamine: Random effects plots for Model P2a

To improve the model fit, attempts were made to model the variance of the within-subject residuals. All of these models were singular, however, and likely over-parameterized. Unfortunately, this necessary step was thus abandoned.

5 Discussion

The above analysis highlighted some strengths and weaknesses of NLME models in capturing complex longitudinal pharmacokinetic trends. The source and nature of the data was presented, and a necessary data manipulation procedure was presented. This reduced the sample size significantly. Following this, the composition of the sample of participants used in the analysis was presented through some descriptive statistics. A thorough exploratory analysis was conducted to demonstrate some of the important considerations when modeling longitudinal data. In this report, there was little clinical input and thus the analysis relied heavily on the data exploration. This revealed that a first-order open compartment model was likely the most appropriate to capture the non-linear pharmacokinetic trajectories. Furthermore, this showed how parts of the average longitudinal profiles differed between different sample strata. Lastly, it suggested that the variance of the response was not constant over time - as an important consideration for the modeling procedure.

Next, the statistical methods were presented - showing the two-stage model formulation used to capture within and between-group sources of variation. The specification, assumptions and interpretation of the first-order open compartment model was given through equation (1), and the full specification of the two-part formulation given in matrix form in equations (2) and (3). The iterative linear mixed effects approximation method used to estimate the models parameters was also outlined. It was then shown how the model can be expanded to incorporate additional levels of nesting, and heteroskedasticity among the within-group residual errors.

Following this, the results associated with the Sulfadoxine measurements were presented. Overall, this showed that by placing random effects on the logarithms of the absorption and clearance rates, significant between-subject variation could be accounted for. The final Sulfadoxine model suggested that there was significant between-study-arm differences in the elimination and clearance of the drug, but not the absorption. There was also evidence that participants with greater mass were associated with lower log-elimination and log-clearance rates, where this association increased quadratically. The model diagnostics suggested that the assumptions related to the random effects were satisfied but the within-subject residual errors were not - and thus the parameter estimates were likely biased to some degree. The assumption of homoskedasticity was relaxed by modeling the variance of the residuals in an exponential framework, but the model residuals were still not normally distributed and this the residual assumptions were still not satisfied.

Next, the Pyrimethamine modeling was presented. It was shown that random effects on the log-elimination and log-clearance parameters were necessary to capture between-subject variation. The final model suggested that there was no evidence to suggest a between-study-arm difference along any of the non-linear model's parameters. It was noted, that the model ideally would have been adjusted for covariates such as weight, but the associated parameters could not be estimated. Similar to the Sulfadoxine model, this model's residuals exhibited non-normality and heteroskedasticity - implying biased parameters. Unfortunately, attempts to model the variance of the residuals were in vain and thus the model was not improved.

A number of major challenges were experienced in, and limitations induced into, this analysis which should be acknowledged. Firstly, the data was processed to manipulate it into a form which would be suitable for the modeling procedure. This presents the first limitation of the analysis. Many participants were removed from the dataset because of missing values, possibly introducing a significant source of bias into the analysis if the observations were not missing completely at random. Besides inducing bias, this also significantly reduced the number of observations and thus limited the complexity of the models which could be fit to the data. The analysis could be strengthened by considering methods for imputation as an alternative to removing missing values.

Furthermore, besides identifiability issues introduced by the combination of relatively few observations and many of those being missing values - the complexity of the fixed effects components of the models was limited to modifying the parameters for participants' weights. This was done to simplify the models, but may have induced bias into the analysis since variation in the responses attributable to other covariates such as sex and age were not accounted for. Similarly, it was shown how an additional level of nesting (subject-within-study-site) could have been introduced to account for this source of variation but was not executed.

6 Conclusion

This report aimed to demonstrate how non-linear mixed effects models can be used to capture complex trends in pharmacokinetic data. The modeling procedure was implemented with the primary aim of evaluating between-study-arm differences in the drug concentration curves of the study's participants. The positive results in this regard suggested that there were significant between-arm differences in the elimination and clearance of Sulfadoxine but no between-arm difference in the processing of Pyrimethamine. It was also seen that the elimination and clearance of Sulfadoxine is, on average, at a slower rate for subjects with relatively larger mass. Unfortunately, the results in this report must be considered with skepticism for a few reasons. The assumptions underlying these models, with respect to the model residuals, were not satisfied - and this the model coefficients were likely biased to some degree. Furthermore, sources of bias were likely introduced into the analysis due to data management procedures and due to the lack of consideration of some covariates.

References

[1] D. M. Bates, J. M. Chambers, and T. Hastie. "Statistical models in S". In: *Computer Science and Statistics: Proceedings of the 19th Symposium on the Interface*. California: Wadsworth & Brooks, 1992.

[2] H. Henry and H. Wickham. *purrr: Functional Programming Tools*. R package version 0.3.4. 2020. URL: https://CRAN.R-project.org/package=purrr.

[3] GW Horde and V Gupta. *Drug Clearance*. StatPearls [Internet]. [Updated 2022 Jun 23]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK557758/. June 2023.

[4] A. Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.6.0. 2023. URL: https://CRAN.R-project.org/package=ggpubr.

[5] M. J. Lindstrom and D. M. Bates. "Nonlinear mixed effects models for repeated measures data". In: *Biometrics* (1990), pp. 673–687.

[6] J. Pinheiro, D. Bates, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-160. 2022. URL: https://CRAN.R-project.org/package=nlme.

[7] J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. New York: Springer, 2000. DOI: 10.1007/b98882.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019. URL: https://www.R-project.org/.

[9] W. Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 2.2.5. Northwestern University. Evanston, Illinois, 2022. URL: https://CRAN.R-project.org/package=psych.

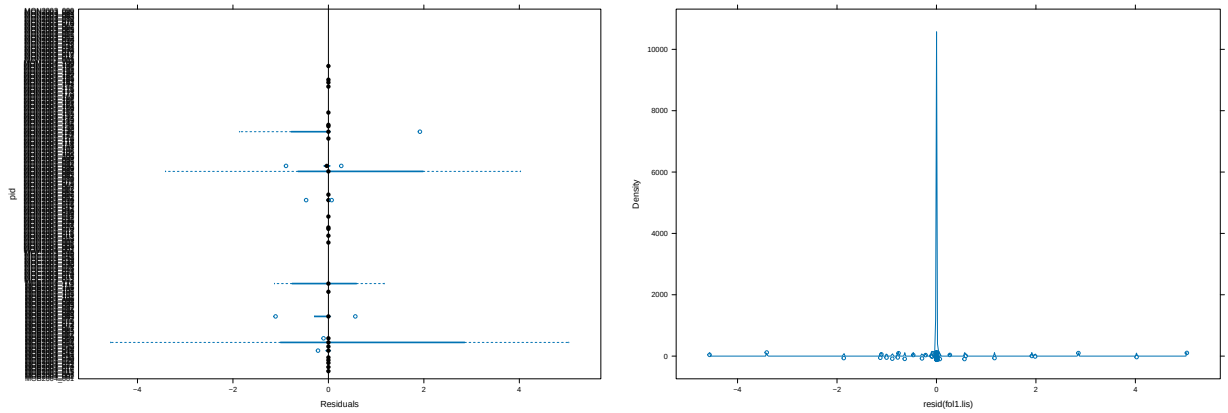
[10] D. Sarkar. *Lattice: Multivariate Data Visualization with R*. ISBN 978-0-387-75968-5. New York: Springer, 2008. URL: http://lmdvr.r-forge.r-project.org.

[11] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[12] H. Wickham et al. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.10. 2022. URL: https://CRAN.R-project.org/package=dplyr.

7 Appendix

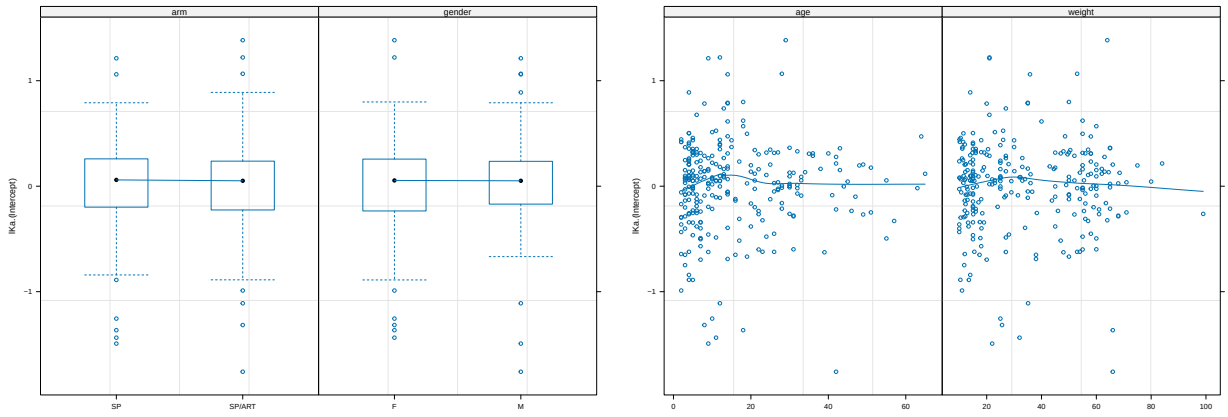
Appendix A: Additional model building figures



(a) Residuals against subject ID

(b) Density plot

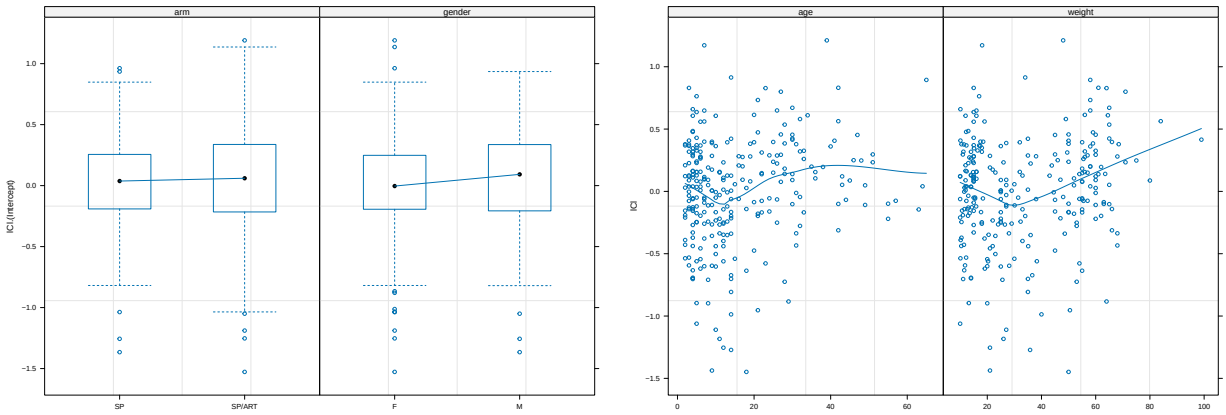
Figure 18: Sulfadoxine: Residual plots for individual fitted models



(a) vs. categorical covariates

(b) vs. continuous covariates

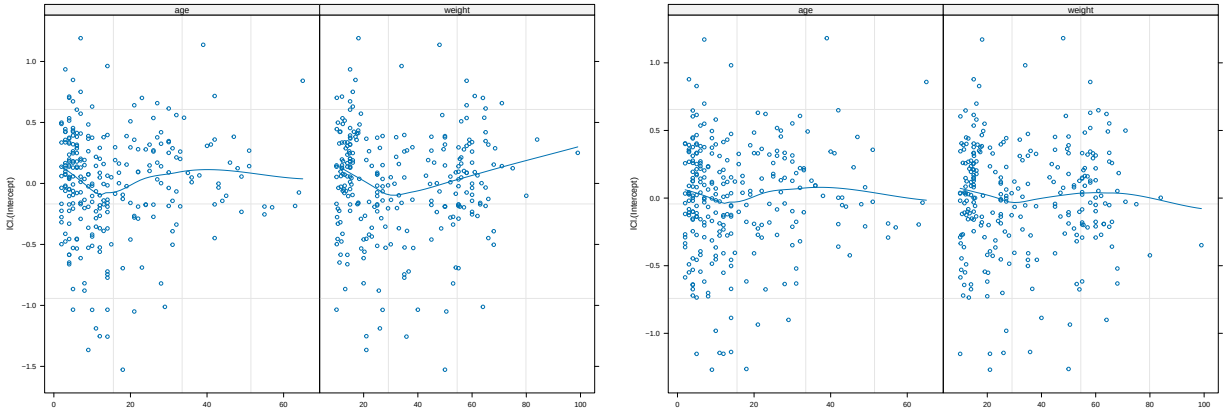
Figure 19: Sulfadoxine: Association between absorption random effects and covariates



(a) vs. categorical covariates

(b) vs. continuous covariates

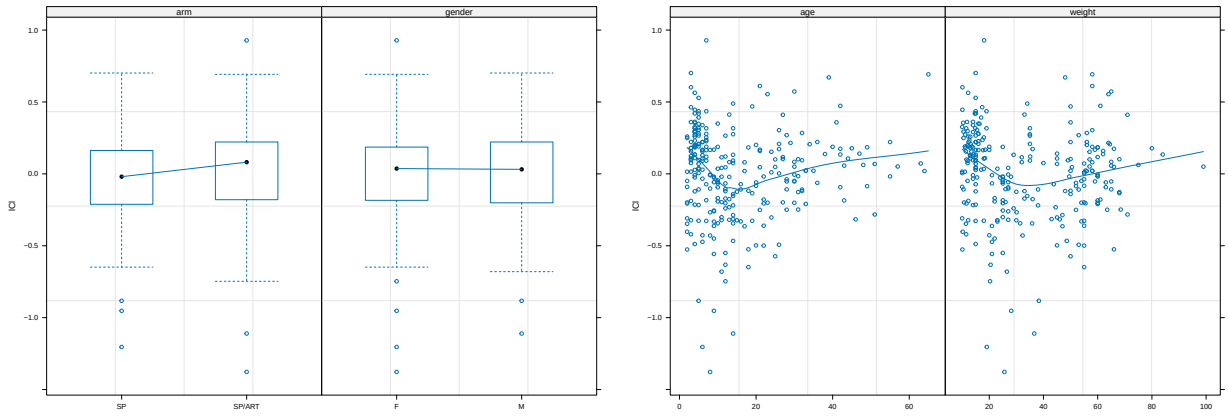
Figure 20: Sulfadoxine: Association between clearance random effects and covariates



(a) Association after including weight

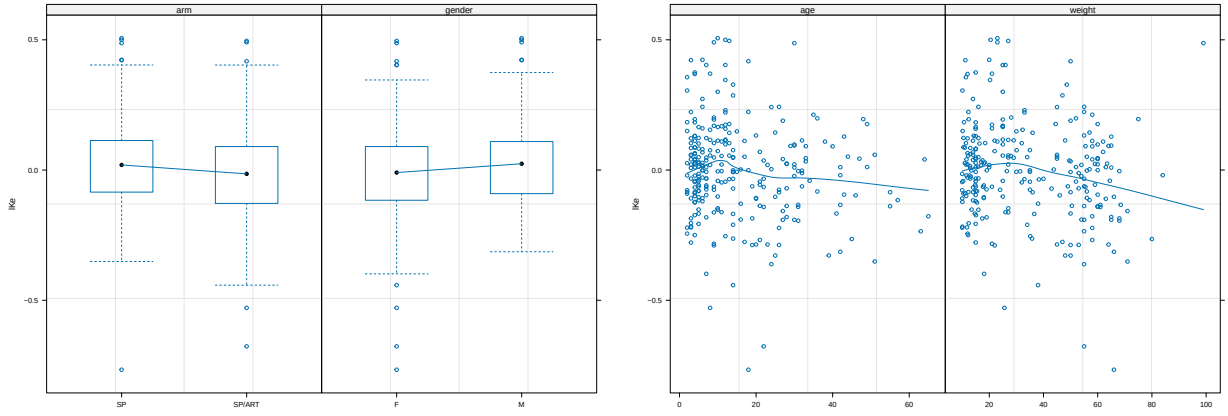
(b) Association after including weight and it's square

Figure 21: Sulfadoxine: Effect of incorporating fixed effects for weight and it's square on clearance



(a) vs. categorical covariates (b) vs. continuous covariates

Figure 22: Pyrimethamine: Association between clearance random effects and covariates



(a) vs. categorical covariates (b) vs. continuous covariates

Figure 23: Pyrimethamine: Association between elimination random effects and covariates

Appendix B: Code Listing

```
#============================ LDA2 - Non-Linear PK Modeling ============
#====== CAMPBELL MCDULING
#====== 06 JUNE 2023
rm(list = ls())
setwd("~/OneDrive - University of Cape Town/2023/MSc Biostatistics/Coursework Semester 1/
    LDA/A2")

# LOAD PACKAGES
library(dplyr)
library(nlme)
library(psych)
library(ggplot2)
library(lattice)
# library(ggeffects)
```

```
# library(emmeans)
library(ggpubr)

#==================== DATA MANAGEMENT =============
dat0 = read.csv("malariadata.csv")
head(dat0)
summary(dat0)

describe(dat0)
str(dat0)

#define variable types accordingly
dat0$site = as.factor(dat0$site)
dat0$arm = as.factor(dat0$arm)
dat0$pid = as.factor(dat0$pid)
dat0$gender = as.factor(dat0$gender)
dat0$country = as.factor(dat0$country)
dat0$PIoutcome = as.factor(dat0$PIoutcome)
summary(dat0)

names(dat0)[8] = "Pconc" #shorten names for convenience
names(dat0)[9] = "Sconc"

#create logged drug concentrations
dat0$l.Sconc = log(dat0$Sconc+0.000001)
dat0$l.Pconc = log(dat0$Pconc+0.000001)

#create weight and age quartiles
quart <- quantile(dat0$weight, probs = c(0, 0.25, 0.5, 0.75, 1), na.rm = T)
dat0$weight_quart <- cut(dat0$weight, breaks = quart, labels = c("Q1", "Q2", "Q3", "Q4"),
     include.lowest = T)
quart <- quantile(dat0$age, probs = c(0, 0.25, 0.5, 0.75, 1), na.rm = T)
dat0$age_quart <- cut(dat0$age, breaks = quart, labels = c("Q1", "Q2", "Q3", "Q4"),
    include.lowest = T)


#==================== DATA VALIDATION =============

#--------- create wide df
dat0.wide = reshape(dat0, timevar = "pday", idvar="pid",direction="wide", v.names = c("
    Pconc", "Sconc"),
                drop=c("Hb", "pardens", "gamedens", "Studyyear"))

#REMOVE ALL PARTICIPANTS WITH NON-ZERO VISIT 1 DRUG CONC
#BOTH DRUGS
index.zer = which(dat0.wide$Pconc.0 == 0 & dat0.wide$Sconc.0 == 0) #retain only those
    with zero for both conc at visit 0
pid.zer = dat0.wide[index.zer,]$pid
index.zer.long = which(dat0$pid %in% pid.zer)
dat0 = dat0[index.zer.long,] #remove from long DF
dat0.wide = dat0.wide[index.zer,] #remove non-zeros from wide DF
n_distinct(dat0$pid) # SAMPLE SIZE REDUCED to 296 participants
n_distinct(dat0.wide$pid)
```

```
#--------- MISSING VALUES
#missing drug concentration values
sum(is.na(dat0$Pconc)) #308 missing Pconc observations
n_distinct(dat0[is.na(dat0$Pconc),]$pid) #over 180 subjects
sum(is.na(dat0$Sconc)) #474 missing Sconc observations
n_distinct(dat0[is.na(dat0$Sconc),]$pid) #over 216 subjects

#--------- MISSING BY TIME POINT

#missings per time point
sum(is.na(dat0.wide$Sconc.0))
sum(is.na(dat0.wide$Sconc.2))
sum(is.na(dat0.wide$Sconc.3))
sum(is.na(dat0.wide$Sconc.7))
sum(is.na(dat0.wide$Sconc.14))
sum(is.na(dat0.wide$Sconc.21))
sum(is.na(dat0.wide$Sconc.28))
sum(is.na(dat0.wide$Sconc.42))
sum(is.na(dat0.wide$Pconc.0))
sum(is.na(dat0.wide$Pconc.2))
sum(is.na(dat0.wide$Pconc.3))
sum(is.na(dat0.wide$Pconc.7))
sum(is.na(dat0.wide$Pconc.14))
sum(is.na(dat0.wide$Pconc.21))
sum(is.na(dat0.wide$Pconc.28))
sum(is.na(dat0.wide$Pconc.42))


#--------- which participants have very few drug concentration measurements
#REMOVE ALL WHO HAVE LESS THAN 4 measurements for each drug
index.missmoreP = which(rowSums(!is.na(dat0.wide[,seq(14, 31, 2)]))<4) #those with less
    than four S after base
dat0.wide[index.missmoreP,]
index.missmoreS = which(rowSums(!is.na(dat0.wide[,seq(15, 31, 2)]))<4) #those less than
    four P after base
dat0.wide[index.missmoreS,]

index.missmore = c(index.missmoreP, index.missmoreS)

pid.missmore = dat0.wide[index.missmore,]$pid
index.missmore.long = which(dat0$pid %in% pid.missmore)
dat0 = dat0[-index.missmore.long,]
dat0.wide = dat0.wide[-index.missmore,]
n_distinct(dat0$pid) #sample size reduced to 271
n_distinct(dat0.wide$pid)


dat0 %>% filter(is.na(weight))
dat0 = dat0 %>% filter(!is.na(weight)) #remove this participant - no weight measurement
dat0.wide = dat0.wide %>% filter(!is.na(weight))
n_distinct(dat0$pid)
```

```
n_distinct(dat0.wide$pid) # sample size reduced to 270

# write.csv(dat0, file="dat0.csv", row.names = F)
# write.csv(dat0.wide, file="dat0.wide.csv", row.names = F)

#==================== DATA EXPLORATION =============
#exploratory data:for descriptive stats - as original

dat0 = read.csv("dat0.csv")
dat0.wide = read.csv("dat0.wide.csv")

dat0$site = as.factor(dat0$site)
dat0$arm = as.factor(dat0$arm)
dat0$pid = as.factor(dat0$pid)
dat0$gender = as.factor(dat0$gender)
dat0$weight_quart = as.factor(dat0$weight_quart)
dat0$age_quart = as.factor(dat0$age_quart)
summary(dat0)

#------- DESCRIPTIVE STATS (gender, age, weight, site, arm, study year)

#baseline statistics by study arm
dat.baseline = dat0 %>% filter(pday==0)

#continuous variables
library(purrr)
dat.baseline[,c(2,8, 9, 11, 12)] %>% split(.$arm) %>% map(summary)

#discrete variables
dat.baseline[,c(1,2, 10)] %>% split(.$arm) %>% map(summary)

#descriptive plots - continuous covars by categorical
ggplot(dat.baseline, aes(x=age, col=arm)) + geom_density()
ggplot(dat.baseline, aes(x=weight, col=arm)) + geom_density()
ggplot(dat.baseline, aes(x=age, col=gender)) + geom_density()
ggplot(dat.baseline, aes(x=weight, col=gender)) + geom_density()

#correlations between continuous variables
#cor.plot(dat.baseline[,c(5:9, 11:12)])

#------- OUTCOMES (all zero at baseline)
### BIVARIATE
dat.baseline[,c(2,8,9, 17,18)] %>% split(.$arm) %>% map(summary) #baseline drug vs arm

#dist of drug conc disregarding timepoint
hist(dat0$Pconc)
hist(dat0$Sconc)
ggplot(dat0, aes(x=Pconc, fill=arm)) + geom_histogram() + theme_minimal()
ggplot(dat0, aes(x=Sconc, fill=arm)) + geom_histogram() + theme_minimal()
table(dat0$arm)

#by visit - densities
p1=ggplot(dat0.wide, aes(x=Pconc.1, col=arm)) + geom_density() + theme_minimal()
```

```
p2=ggplot(dat0.wide, aes(x=Pconc.2, col=arm)) + geom_density() + theme_minimal()
p3=ggplot(dat0.wide, aes(x=Pconc.3, col=arm)) + geom_density() + theme_minimal()
p4=ggplot(dat0.wide, aes(x=Pconc.7, col=arm)) + geom_density() + theme_minimal()
p5=ggplot(dat0.wide, aes(x=Pconc.14, col=arm)) + geom_density() + theme_minimal()
p6=ggplot(dat0.wide, aes(x=Pconc.21, col=arm)) + geom_density() + theme_minimal()
p7=ggplot(dat0.wide, aes(x=Pconc.28, col=arm)) + geom_density() + theme_minimal()
p8=ggplot(dat0.wide, aes(x=Pconc.42, col=arm)) + geom_density() + theme_minimal()
ggarrange(p1,p2,p3,p4,p5,p6,p7,p8,ncol=4,nrow=2)

s1=ggplot(dat0.wide, aes(x=Sconc.1, col=arm)) + geom_density() + theme_minimal()
s2=ggplot(dat0.wide, aes(x=Sconc.2, col=arm)) + geom_density() + theme_minimal()
s3=ggplot(dat0.wide, aes(x=Sconc.3, col=arm)) + geom_density() + theme_minimal()
s4=ggplot(dat0.wide, aes(x=Sconc.7, col=arm)) + geom_density() + theme_minimal()
s5=ggplot(dat0.wide, aes(x=Sconc.14, col=arm)) + geom_density() + theme_minimal()
s6=ggplot(dat0.wide, aes(x=Sconc.21, col=arm)) + geom_density() + theme_minimal()
s7=ggplot(dat0.wide, aes(x=Sconc.28, col=arm)) + geom_density() + theme_minimal()
s8=ggplot(dat0.wide, aes(x=Sconc.42, col=arm)) + geom_density() + theme_minimal()
ggarrange(s1,s2,s3,s4,s5,s6,s7,s8,ncol=4,nrow=2)



#---------------------- LONGITUDINAL EXPLORATORY ----
#===SPAGHETTI PLOTS

xyplot(Sconc ~ pday, dat0, type="l", groups = pid)
#by study arm
xyplot(Sconc ~ pday|arm, dat0, type="l", groups = pid)
#by study site
xyplot(Sconc ~ pday|site, dat0, type="l", groups = pid)

xyplot(Pconc ~ pday, dat0, type="l", groups = pid)
#by study arm
xyplot(Pconc ~ pday|arm, dat0, type="l", groups = pid)
#by study site
xyplot(Pconc ~ pday|site, dat0, type="l", groups = pid)



#mean log S concentrations
ggplot(dat0, aes(x = pday, y = l.Sconc, color = arm, group = arm)) +
  stat_summary(fun = "mean", geom = "line") +
  labs(x = "Day", y = "Mean log(S + 1e-06)", color = "Study Arm") +
  theme_minimal()

#Mean drug conc by study arm (both drugs)
ggplot(dat0, aes(x = pday, color = arm, group=arm)) +
  geom_line(aes(y = Sconc, linetype = "Sconc"), stat = "summary", fun = "mean") +
  geom_line(aes(y = Pconc, linetype = "Pconc"), stat = "summary", fun = "mean") +
  labs(x = "Day", y = "Mean drug concentrations", color = "Study Arm", linetype = "Drug")
       +
  scale_linetype_manual(values = c("Sconc" = "solid", "Pconc" = "dashed")) +
  theme_minimal()
#above but on log scale
```

```
ggplot(dat0, aes(x = pday, color = arm, group=arm)) +
  geom_line(aes(y = l.Sconc, linetype = "log(Sconc)"), stat = "summary", fun = "mean") +
  geom_line(aes(y = l.Pconc, linetype = "log(Pconc)"), stat = "summary", fun = "mean") +
  labs(x = "Day", y = "Mean log drug concentrations", color = "Study Arm", linetype = "
      Drug") +
  scale_linetype_manual(values = c("log(Sconc)" = "solid", "log(Pconc)" = "dashed")) +
  theme_minimal()
# BY SEX
ggplot(dat0, aes(x = pday, color = gender, group=gender)) +
  geom_line(aes(y = Sconc, linetype = "Sconc"), stat = "summary", fun.y = "mean") +
  geom_line(aes(y = Pconc, linetype = "Pconc"), stat = "summary", fun.y = "mean") +
  labs(x = "Day", y = "Mean drug concentrations", color = "Sex", linetype = "Drug") +
  scale_linetype_manual(values = c("Sconc" = "solid", "Pconc" = "dashed")) +
  theme_minimal()
ggplot(dat0, aes(x = pday, color = gender, group=gender)) +
  geom_line(aes(y = l.Sconc, linetype = "log(Sconc)"), stat = "summary", fun.y = "mean")
      +
  geom_line(aes(y = l.Pconc, linetype = "log(Pconc)"), stat = "summary", fun.y = "mean")
      +
  labs(x = "Day", y = "Mean log drug concentrations", color = "Sex", linetype = "Drug") +
  scale_linetype_manual(values = c("log(Sconc)" = "solid", "log(Pconc)" = "dashed")) +
  theme_minimal()
# BY SITE
ggplot(dat0, aes(x = pday, color = site, group=site)) +
  geom_line(aes(y = Sconc, linetype = "Sconc"), stat = "summary", fun.y = "mean") +
  geom_line(aes(y = Pconc, linetype = "Pconc"), stat = "summary", fun.y = "mean") +
  labs(x = "Day", y = "Mean drug concentrations", color = "Site", linetype = "Drug") +
  scale_linetype_manual(values = c("Sconc" = "solid", "Pconc" = "dashed")) +
  theme_minimal()
ggplot(dat0, aes(x = pday, color = site, group=site)) +
  geom_line(aes(y = l.Sconc, linetype = "log(Sconc)"), stat = "summary", fun.y = "mean")
      +
  geom_line(aes(y = l.Pconc, linetype = "log(Pconc)"), stat = "summary", fun.y = "mean")
      +
  labs(x = "Day", y = "Mean log drug concentrations", color = "Site", linetype = "Drug")
      +
  scale_linetype_manual(values = c("log(Sconc)" = "solid", "log(Pconc)" = "dashed")) +
  theme_minimal()

#by weight quartile
ggplot(dat0, aes(x = pday, color = weight_quart, group=weight_quart)) +
  geom_line(aes(y = Sconc, linetype = "Sconc"), stat = "summary", fun = "mean") +
  geom_line(aes(y = Pconc, linetype = "Pconc"), stat = "summary", fun = "mean") +
  labs(x = "Day", y = "Mean drug concentrations", color = "Weight quartile", linetype = "
      Drug") +
  scale_linetype_manual(values = c("Sconc" = "solid", "Pconc" = "dashed")) +
  theme_minimal()


# VARIANCE PROFILES
ggplot(dat0, aes(x = pday, color = arm, group=arm)) +
  geom_line(aes(y = Sconc, linetype = "Sconc"), stat = "summary", fun = "var") +
  geom_line(aes(y = Pconc, linetype = "Pconc"), stat = "summary", fun = "var") +
```

```
  labs(x = "Day", y = "Variance of drug concentrations", color = "Study Arm", linetype =
      "Drug") +
  scale_linetype_manual(values = c("Sconc" = "solid", "Pconc" = "dashed")) +
  theme_minimal()
ggplot(dat0, aes(x = pday, color = arm, group=arm)) +
  geom_line(aes(y = l.Sconc, linetype = "Sconc"), stat = "summary", fun = "var") +
  geom_line(aes(y = l.Pconc, linetype = "Pconc"), stat = "summary", fun = "var") +
  labs(x = "Day", y = "Variance of logged drug concentrations", color = "Study Arm",
      linetype = "Drug") +
  scale_linetype_manual(values = c("Sconc" = "solid", "Pconc" = "dashed")) +
  theme_minimal()

#============================ MODELING ========================
dat0 = read.csv("dat0.csv")
dat0.wide = read.csv("dat0.wide.csv")

dat0$site = as.factor(dat0$site)
dat0$arm = as.factor(dat0$arm)
dat0$pid = as.factor(dat0$pid)
dat0$gender = as.factor(dat0$gender)
dat0$weight_quart = as.factor(dat0$weight_quart)
dat0$age_quart = as.factor(dat0$age_quart)


dat0$Dose = 1 #create dose variable

#what is minimum conc, besides zero
temp = dat0 %>% filter(.$Pconc != 0)
min(temp$Pconc) #0.146
temp = dat0 %>% filter(.$Sconc != 0)
min(temp$Sconc) #0.125

dat0$Sconc[which(dat0$Sconc==0)] = 1e-4 #change all zero concentrations to very small +
    ive number
dat0$Pconc[which(dat0$Pconc==0)] = 1e-4

dat0.grp.s = groupedData(Sconc ~ pday|site/pid, dat0)
dat0.grp.p = groupedData(Pconc ~ pday|site/pid, dat0)

#------------- FIRST ORDER COMPARTMENT MODEL
#====================== SULFCONC=======
#----- LEVEL 1 MODEL
#fit for one subject
tst = dat0.grp.s[dat0.grp.s$pid=="MOB2004_015",]
nls(Sconc~SSfol(Dose, pday, lKe, lKa, lCl), data=tst)

#fit for all subjects
#contr = nls.control(minFactor = 1/5000, tol=1e-2)
fol1.lis = nlsList(Sconc ~ SSfol(Dose, pday, lKe, lKa, lCl)|pid, data=dat0.grp.s,
                   na.action = na.exclude)
fol1.lis
summary(fol1.lis)
```

```
#investigate random effects structure
plot(intervals(fol1.lis)) #much between-pid variability in clearance and elimination,
    absorption is similar
#investigate covariance structure
pairs(fol1.lis, id=0.05) #maybe negative relationship between lKe and lKa, positive
    between lKe and lCl
#residuals
plot(fol1.lis, pid~resid(.), abline=0, xlab="residuals")
densityplot(resid(fol1.lis), xlab="residuals")
plot(resid(fol1.lis))

#----- LEVEL 2 MODEL
##### RANDOM EFFECTS
nlm.contr = nlmeControl(minScale = 1e-20, maxIter = 50, tolerance = 1e-1)

#Model 1a - general +ive def
ran.S1a = list(lKe + lKa + lCl ~ 1)
nlm.S1a = nlme(fol1.lis, random = ran.S1a, control = nlm.contr, method="REML")
summary(nlm.S1a) #magnitude of lKa and lCl RE are much smaller than lKe

#Model 1b - diagonal
ran.S1b = pdDiag(lKe + lKa + lCl ~ 1)
nlm.S1b = update(nlm.S1a, random=ran.S1b)
summary(nlm.S1b) #magnitude of lKe is much smaller now

anova(nlm.S1a, nlm.S1b) #loglik is signif reduced, take diagonal model

# #Model 1c - blocked
# ran.S1c = pdBlocked(list(lKe~1, lKa+lCl~1))
# nlm.S1c = update(nlm.S1a, random=ran.S1c) #does not converge
# summary(nlm.S1c)
#
# anova(nlm.S1b, nlm.S1c)

#Model 1d - no elimination RE
ran.S1d = pdDiag(lKa + lCl ~ 1)
nlm.S1d = update(nlm.S1b, random=ran.S1d)
summary(nlm.S1d)

anova(nlm.S1b, nlm.S1d) # no sig diff between models, take simpler model (1d)

#Model 1e - only lKa RE
nlm.S1e = update(nlm.S1d, random=list(lKa~1))
# #Model 1f - only lCl RE
# nlm.S1f = update(nlm.S1d, random=list(lCl~1)) #singularity error

anova(nlm.S1d, nlm.S1e) #sig diff between models, stick with model 1d with two REs


plot(nlm.S1d) #non-constant variance
qqnorm(nlm.S1d) #big deviation from normality for residuals
densityplot(resid(nlm.S1d))
qqnorm(nlm.S1d, ~ ranef(.)) #normality for random effects
```

```
##### FIXED EFFECTS
#investigate fixed eff structure
re.nlm.S1d = ranef(nlm.S1d)
re.nlm.S1d$arm = dat0.wide$arm
re.nlm.S1d$gender = dat0.wide$gender
re.nlm.S1d$weight = dat0.wide$weight
re.nlm.S1d$age = dat0.wide$age
plot(re.nlm.S1d, form=lKa~arm+gender) #no significant deviation between arms or sexes
plot(re.nlm.S1d, form=lKa~age+weight) #appears to be a weight and age effect, weight is
    stronger
plot(re.nlm.S1d, form=lCl~arm+gender) #no significant deviation between arms or sexes
plot(re.nlm.S1d, form=lCl~age+weight) #appears to be a weight and age effect, strong
    weight effect
cor(dat0.wide$age, dat0.wide$weight, method="pearson")


# BUILD MODEL
nlm.contr2 = nlmeControl(minScale = 1e-20, tolerance = 0.2)

nlm.S1d.ML = update(nlm.S1d, method="ML")#create ML version for model comparisons
summary(nlm.S1d.ML)
#extract fixed effects for starting values
nlm.S1d.fix = fixef(nlm.S1d.ML)
fix1 = list(lKe + lKa + lCl ~ arm)
start1 = c(nlm.S1d.fix[1], -0.02, nlm.S1d.fix[2], 0.04, nlm.S1d.fix[3], 0.08)

#Model 2a - fixed effect on study arm
nlm.S2a = update(nlm.S1d.ML, fixed=fix1, start=start1, control=nlm.contr2)
summary(nlm.S2a) #arm effects are sig


anova(nlm.S1d.ML, nlm.S2a) #logLik significantly reduced

#Model 2b - adjust for other covariates - forward selection
nlm.S2a.fix = fixef(nlm.S2a)
fix2 = list(lKe + lKa + lCl ~ arm+weight)
start2 = c(nlm.S2a.fix[1], nlm.S2a.fix[2], 0, nlm.S2a.fix[3], nlm.S2a.fix[4], 0,
          nlm.S2a.fix[5], nlm.S2a.fix[6], 0)
nlm.S2b = update(nlm.S2a, fixed=fix2, start=start2)
summary(nlm.S2b) #weight significantly modifies elimination and maybe clearance, not
    absorption

anova(nlm.S2a, nlm.S2b) #significantly different, 2b is chosen
anova(nlm.S1d.ML, nlm.S2a, nlm.S2b)

#look at RE associations now that weight and arm are adjusted for
re.nlm.S2b = ranef(nlm.S2b)
re.nlm.S2b$arm = dat0.wide$arm
re.nlm.S2b$gender = dat0.wide$gender
re.nlm.S2b$weight = dat0.wide$weight
re.nlm.S2b$age = dat0.wide$age
plot(re.nlm.S2b, form=lKa.(Intercept)~arm+gender) #much the same as before
```

```
plot(re.nlm.S2b, form=lKa.(Intercept)~age+weight) #much of the association across both
    age and weight has been removed
plot(re.nlm.S2b, form=lCl.(Intercept)~arm+gender) #much the same as before
plot(re.nlm.S2b, form=lCl.(Intercept)~age+weight) #still association remaining

#Model 2c - try weight^2 also
fix3 = list(lKe + lKa + lCl ~ arm+weight+I(weight^2))
start3 = c(nlm.S2a.fix[1], nlm.S2a.fix[2], 0, 0, nlm.S2a.fix[3], nlm.S2a.fix[4], 0, 0,
          nlm.S2a.fix[5], nlm.S2a.fix[6], 0,0)
nlm.S2c = update(nlm.S2b, fixed=fix3, start=start3)
#nlm.S2c = nlme(Sconc ~ SSfol(Dose, pday, lKe, lKa, lCl), data=dat0.grp.s,
# fixed=fix3, start=start3, random = ran.S1d, control=nlm.contr2, na.action = na.omit)
summary(nlm.S2c)

anova(nlm.S2b, nlm.S2c) #model significantly improved

anova(nlm.S2a, nlm.S2b, nlm.S2c)

#Model 2d - has weight effects only on elim and clear
fix4 = list(lKa~arm, lKe+lCl~arm+weight+I(weight^2))
nlm.S2c.fix = fixef(nlm.S2c)
start4 = c(nlm.S2c.fix[1], nlm.S2c.fix[2], nlm.S2c.fix[3], nlm.S2c.fix[4],
          nlm.S2c.fix[5], nlm.S2c.fix[6],
          nlm.S2c.fix[9], nlm.S2c.fix[10], nlm.S2c.fix[11], nlm.S2c.fix[12])
#nlm.S2d = update(nlm.S2c, fixed=fix4, start=start4) #singularity error

#=========== MODEL VALIDATION
#fit model with REML
nlm.S2c = update(nlm.S2c, method="REML")
summary(nlm.S2c)
re.nlm.S2c = ranef(nlm.S2c)
re.nlm.S2c$arm = dat0.wide$arm
re.nlm.S2c$gender = dat0.wide$gender
re.nlm.S2c$weight = dat0.wide$weight
re.nlm.S2c$age = dat0.wide$age

plot(re.nlm.S2c, form=lKa.(Intercept)~age+weight)
plot(re.nlm.S2c, form=lCl.(Intercept)~age+weight) #association removed

#residuals
plot(nlm.S2c) #definitely heteroskedasticity
qqnorm(nlm.S2c) #definitely non-normal
densityplot(resid(nlm.S2c), xlab="Standardized residuals")

#random effects
qqnorm(nlm.S2c, ~ ranef(.)) #normality for random effects
d1=densityplot(unlist(ranef(nlm.S2c)[1]), xlab="lKa random effects")
d2=densityplot(unlist(ranef(nlm.S2c)[2]), xlab="lCl random effects")
ggarrange(d1,d2)
plot(ranef(nlm.S2c), xlab="")


#### MODEL VARIANCE
```

```
nlm.S3a = update(nlm.S2c, weights=varExp())
summary(nlm.S3a)
plot(nlm.S3a)
qqnorm(nlm.S3a, ~resid(.))
densityplot(resid(nlm.S3a))
plot(nlm.S2c)

#nlm.S3b = update(nlm.S2c, weights=varPower()) #error thrown

#nlm.S3c = update(nlm.S2c, weights=varConstPower()) #error thrown

nlm.S3d = update(nlm.S2c, weights=varIdent(~pday))
summary(nlm.S3d)
plot(nlm.S3d) #not fixed

#====================== PYRCONC =======
#----- LEVEL 1 MODEL
#fit for one subject
tst = dat0.grp.p[dat0.grp.p$pid=="MOB2004_015",]
nls(Pconc~SSfol(Dose, pday, lKe, lKa, lCl), data=tst)

#fit for all subjects
#lis.contr = nls.control(maxiter = 200, minFactor = 1/1024, tol=1e-2)
fol2.lis = nlsList(Pconc ~ SSfol(Dose, pday, lKe, lKa, lCl)|pid, data=dat0.grp.p,
                 na.action = na.exclude)
fol2.lis

#investigate random effects structure
plot(intervals(fol2.lis)) #much between-pid variability in all three parameters
plot(fol2.lis, pid~resid(.), abline=0)

#investigate covariance structure
pairs(fol2.lis, id=0.05) #maybe negative relationship between lKe and lKa

#----- LEVEL 2 MODEL
##### RANDOM EFFECTS
nlm.contr = nlmeControl(minScale = 1e-6, maxIter = 50, tol=1e-1)

#Model 1 - diagonal with all three parameters as RE
ran.P1 = pdDiag(lKa + lKe + lCl ~ 1)
#nlm.P1 = nlme(fol2.lis, random=ran.P1, control=nlm.contr, method = "REML") #cannot fit
    all three - singularity error (unidentifiable model)


#Model 1a - diagonal with two as RE
ran.P1a = pdDiag(lKa + lKe ~ 1)
nlm.P1a = nlme(fol2.lis, random = ran.P1a, control = nlm.contr, method = "REML")
summary(nlm.P1a) #magnitude of lKa RE is small

#Model 1b
ran.P1b = pdDiag(lKe + lCl ~ 1)
nlm.P1b = update(nlm.P1a, random=ran.P1b)
summary(nlm.P1b)
```

```
anova(nlm.P1a, nlm.P1b) #take model 1b on information criteria and mag of REs

#Model 1c
ran.P1c = pdDiag(lKa + lCl ~ 1)
#nlm.P1c = update(nlm.P1a, random=ran.P1c) #does not converge
#summary(nlm.P1c)


#Model 1c - +ive definite
ran.P1d = list(lKe + lCl ~ 1)
nlm.P1d = update(nlm.P1a, random=ran.P1d)
summary(nlm.P1d)

anova(nlm.P1a,nlm.P1b, nlm.P1d) # sig diff, take simpler model

#Model 1e - RE only on elimination
ran.P1e = list(lKe~1)
nlm.P1e = update(nlm.P1a, random=ran.P1e)
summary(nlm.P1e)



# #Model 1g - RE on clearance only
# ran.P1g = list(lCl~1)
# nlm.P1g = update(nlm.P1a, random=ran.P1g) #singularity error
# summary(nlm.P1g)


anova(nlm.P1a, nlm.P1b, nlm.P1d, nlm.P1e) #best model is 1b



plot(nlm.P1b) #fairly random scatter
qqnorm(nlm.P1b) #big deviation from normality for residuals
densityplot(resid(nlm.P1b))
qqnorm(nlm.P1b, ~ ranef(.)) #reasonable normality for random effects
densityplot(unlist(ranef(nlm.P1b)[1]), xlab="lKe")
densityplot(unlist(ranef(nlm.P1b)[2]), xlab="lKe")


##### FIXED EFFECTS
#investigate fixed eff structure
re.nlm.P1b = ranef(nlm.P1b)
re.nlm.P1b$arm = dat0.wide$arm
re.nlm.P1b$gender = dat0.wide$gender
re.nlm.P1b$weight = dat0.wide$weight
re.nlm.P1b$age = dat0.wide$age
plot(re.nlm.P1b, form=lKe~arm+gender) #no significant deviation between arms or sexes
plot(re.nlm.P1b, form=lKe~age+weight) #appears to be a weight and age effect, strong
    weight effect
plot(re.nlm.P1b, form=lCl~age+weight) #appears to be a weight and age effect, strong
    weight effect
```

```
plot(re.nlm.P1b, form=lCl~arm+gender) #appears to be a weight and age effect, strong
    weight effect

# BUILD MODEL
nlm.P1b.ML = update(nlm.P1b, method="ML")#create ML version for model comparisons
summary(nlm.P1b.ML)
#extract fixed effects for starting values
nlm.P1b.fix = fixef(nlm.P1b.ML)
fix1 = list(lKe + lKa + lCl ~ arm)
start1 = c(nlm.P1b.fix[1], 0, 2, 0, nlm.P1b.fix[3], 0)
nlm.contr2 = nlmeControl(minScale = 1e-20, tol=0.1) #reduce minimum halving factor

#Model 2a - fixed effect on study arm
nlm.P2a = update(nlm.P1b.ML, fixed=fix1, start=start1, control=nlm.contr2)
summary(nlm.P2a) #none of additional effects are sig


anova(nlm.P1b.ML, nlm.P2a)

#adjust for other covariates - forward selection
#weight
nlm.P2a.fix = fixef(nlm.P2a)
fix2 = list(lKe + lKa + lCl ~ arm+weight)
start2 = c(nlm.P2a.fix[1], nlm.P2a.fix[2], 0, nlm.P2a.fix[3], nlm.P2a.fix[4], 0,
          nlm.P2a.fix[5], nlm.P2a.fix[6], 0)
#start2 = c(-1, -0.02,0.001, 2, -0.3, 0.0001, -7, 0.15, 0.0001)
#nlm.P2b = update(nlm.P1b.ML, fixed=fix2, start=start2, control=nlm.contr2) #
    AAAAAHAHHHHHHHHHHHHH SINGULARITY AHHHHHHH
#summary(nlm.P2b) #ANGRY, fury, mania

#anova(nlm.P1b.ML, nlm.P2a, nlm.P2b) #frustration



#look at RE associations now that weight and arm are adjusted for
# re.nlm.P2b = ranef(nlm.P2b)
# re.nlm.P2b$arm = dat0.wide$arm
# re.nlm.P2b$gender = dat0.wide$gender
# re.nlm.P2b$weight = dat0.wide$weight
# re.nlm.P2b$age = dat0.wide$age
# plot(re.nlm.P2b, form=lKe.(Intercept)~arm+gender) #much the same as before
# plot(re.nlm.P2b, form=lKe.(Intercept)~age+weight) #the association across both age and
    weight has been removed
# plot(re.nlm.P1b, form=lKe~age+weight) #before correction

#=========== MODEL VALIDATION
nlm.P2a.reml = update(nlm.P2a, method="REML")
summary(nlm.P2a.reml)
#residuals
plot(nlm.P2a.reml) #non-constant variance
qqnorm(nlm.P2a.reml) #large deviation from normality
densityplot(resid(nlm.P2a.reml), xlab="Standardized residuals")
```

```
#random effects
qqnorm(nlm.P2a.reml, ~ ranef(.)) #normality for random effects
d1=densityplot(unlist(ranef(nlm.P2a.reml)[1]), xlab="lKe random effects")
d2=densityplot(unlist(ranef(nlm.P2a.reml)[2]), xlab="lCl random effects")
ggarrange(d1, d2)
plot(ranef(nlm.P2a.reml))


#---------------------- MODEL VARIANCE ---------
varmod1 = varPower()
# nlm.P3a = update(nlm.P2a.reml, weights=varPower()) #singular
# nlm.P3b = update(nlm.P2a.reml, weights=varExp()) #singular
# nlm.P3c = update(nlm.P2a.reml, weights=varConstPower()) #singular
```