# Part 1 – The prothrombin data

## 1. Data description

The data comes from a longitudinal, 2 arm RCT of 383 patients. 189 patients were randomised to the active treatment and 194 to placebo. Prothrombin measurements were taken at baseline (month 0) and at 3, 6, 12, and 24 months after randomisation. There is missing data, only 53 patients in each arm (27.3% of placebo group, 28.0% of active) have prothrombin measurements at all 5 times. Table 1 shows the number of patients measured at each time point & the mean prothrombin measurement.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Measure | Month 0 | Month 3 | Month 6 | Month 12 | Month 24 |
| **Active** | Patients | 189 | 147 | 152 | 131 | 108 |
| Patients with full data | 189 | 147 | 112 | 83 | 53 |
| Prothrombin: mean (SD) | 13.9 (2.1) | 13.9 (2.4) | 14.4 (2.4) | 14.5 (2.3) | 14.8 (2.6) |
| **Placebo** | Patients | 194 | 158 | 161 | 115 | 98 |
| Patients with full data | 194 | 158 | 125 | 78 | 53 |
| Prothrombin: mean (SD) | 14.3 (2.3) | 15.4 (2.3) | 15.4 (2.9) | 15.4 (2.6) | 15.5 (2.6) |
| *Table 1: Missing data patterns & prothrombin measurements. Prothrombin measured in seconds* | | | | | | |

Patients can miss appointments and attend later appointments – fewer patients attend the month 3 measurement compared to month 6 for example. The distribution of prothrombin measurement times is approximately normal at each time point for both the active and placebo groups. The average trajectory for the two groups is shown in the third & sixth lines of the table. Prothrombin time increases over the course of the study, and increases faster in the placebo group compared to active. In the placebo group there is a large increase between month 0 and 3 followed by a plateau. The active group sees a smaller increase which starts in month 6, followed by a gradual rise over time. These trajectories are affected by the missing data mentioned earlier, so may not be representative. More representative trajectories – which account for the missing data – are shown in the results section.  
Person level data on sex is also recorded and has no missing data.

## 2. Methods

All analysis was done in Stata 17. Mixed (conditional) models were fit using the mixed command, and marginal models were fit using the xtgee command. Some data processing – converting implicit missing values to explicit and adding the baseline prothrombin measurement as a person level variable – was done before model fitting and analysis.  
Plotting individual trajectories from both groups shows fanning, so a random coefficient model was used. As there are a small number of variables, we begin by saturating the mean structure and try a variety of variance structures. An initial model used an unstructured covariance matrix, and the residuals were allowed to vary by sex. Simpler covariance structures were then fitted and assessed through likelihood ratio tests (LRTs). There was no evidence that allowing residuals to vary with sex resulted in a better fit compared to a constant level-1 variance model (p-value = 0.12), and there is strong evidence that an unstructured covariance matrix has a better fit compared to an independent or exchangeable structure (p-values 0.03 and <0.0001 respectively). As such, an unstructured covariance matrix with constant level-1 variance was chosen as the variance structure for the model.  
 With the variance structure chosen we next start to simplify the mean structure by removing terms, starting with the terms with highest Wald statistic p-values. The effect of removing each term was assessed using LRTs. This process removed almost all interaction terms and kept all linear terms. There was no evidence that the simplified mean structure had worse fit compared to the saturated model (LRT p-value = 0.62).   
 All models were fit using maximum likelihood rather than restricted maximum likelihood, which allows us to use LRTs to test both the mean and variance structures. Once the final model was chosen, it was fit to the data using restricted maximum likelihood to reduce bias.

The final model is a random coefficient model with random effects on the intercept and time, and fixed effects for time, treatment group, sex, and the interaction of baseline with time & treatment group. The formula for the model is:

where is the prothrombin measurement of person at time , is the time of measurement for person , the are fixed effects, and are random effects for the intercept and slope respectively, is the residual error of the measurement for person , and the square brackets are indicator variables which are equal to 1 if the condition inside is true and zero otherwise. The time index is for months 3, 6, 12, and 24 respectively.  
Conditional on the covariates in the formula above, the residual error is assumed to be independent of the covariates with mean zero and constant variance , and the random effects are assumed to be independent of the covariates, independent of the residual error, and be distributed as

For the marginal model, we keep the same mean structure as we used in the conditional mixed effect model and use an unstructured working matrix and robust standard errors. To estimate population average trajectories [DESCRIBE STEPS]

## 3. Results

Words

## 4. Summary

Words

# Part 2 – Novel research setting