Introduction   Generalized least squares (GLS), linear mixed effect models (LME), and generalized estimating equations (GEE) are commonly used methods for fitting data with correlation in R. However, the performance of these methods can vary depending on factors such as the sample size and model specifications. In this project, we compare the outputs of R packages that implement GLS, LME, and GEE based on simulated scenarios. One scenario involves data generated via a GLS model, while the other involves data generated by an LME model. Specifically, we investigate the performance of these methods and their R implementations to recover the true coefficients and error terms, and examine the heterogeneity captured in the different models. Through our analysis, we aim to provide insights into the strengths and limitations of these methods and inform the selection of appropriate methods for practical applications.

Simulation design and implementation   The data generation process consists of two simulation settings, each generating data for n=10, 50, 100. In the first simulation, a Generalized Least Squares (GLS) model will be fitted on the CD4 data to extract coefficients, which will be used to generate data. The model assumes a compound symmetry structure corresponding to a constant correlation. In the second simulation, coefficients will be estimated via a random intercept linear mixed effect model. Similarly, data will be generated, and GLS, Generalized Estimating Equations (GEE), and Linear Mixed Effects (LME) models will be fitted to recover coefficients. The data generated will be used to compare the models' performance under different sample sizes.

Description of R packages and Metrics used to evaluate performance   Different packages were used to fit the data generated. These packages were `lme4`, `nlme`, `geepack` and `glmmTMB`. We applied the following models implemented in the corresponding R packages: 1) GLS model using `nlme` package, 2) LME model using `nlme` package, 3) LME model using `lme4` package, 4) GEE model using `geepack` package, 5) Generalized linear mixed model using `glmmTMB` package. After fitting the models, we extracted the coefficients for each model, including fixed effects intercept and `Week`, random effect's standard deviation if applicable, and error term's standard deviation. Lastly, the bias of the estimates were calculated by subtracting the estimated value from the true value to evaluate estimation accuracy.

Results and Conclusion   The analysis employed various packages, including `lme4`, `nlme`, `geepack`, and `glmmTMB`, to fit a model with log CD4 counts as the outcome and an intercept and Week as the predictor. Results revealed that the estimated fixed effect was consistent across the different methods, despite slight variations in the standard deviation of the random effects for the same sample size. The low bias of the fixed effects indicated that the models performed well in parameter estimation. However, for models that estimated the random effect, it was evident that the covariance matrix captured a portion of the 'true variance,' highlighting the importance of selecting an appropriate model and considering the data's characteristics to obtain accurate estimates for variances.

Furthermore, the sum of the random effect variance and error variance was uniform across all methods, and the sum was in proximity to the true variance used to generate the data. Notably, the `lme4` and `nlme` packages exhibited a more similar sum of estimated variances than `geepack` and `glmmTMB`, possibly due to the algorithm employed in solving for the solution.

Overall, the findings suggest that the models used in this study effectively estimated the parameters, and the selection of the appropriate model and method is crucial in accurately estimating variances. Future research could explore the performance of different packages and models for other types of data and outcomes.

Detailed analysis can be found via https://chloeyou.github.io/gls-gee-lme-comparisons/analysis/analysis.html