Proceedings of the 6th Győr Symposium and 3rd Hungarian-Polish
and 1st Hungarian-Romanian Joint Conference on Computational Intelligence

Modelling Repeated Measurements Data on the Example of Systolic and Diastolic Blood Pressure

D. Kruszewski

Ph.D. Studies, Polish Academy of Sciences, Systems Research Institute
ul. Newelska 6, 01-447 Warsaw, Poland
E-mail: damian.kruszewski@tlen.pl

Abstract: The relationship between blood pressure and various medical, demographic and socioeconomic characteristics, on the example of the longitudinally collected data, was examined. The correlation between the repeated measurements on the same experimental unit was modelled using random effects. Significant predictors of systolic and diastolic blood pressure were identified.

Keywords: blood pressure, repeated measurements, random effects

1. Introduction

The purpose of this paper is to study, on the example of the longitudinally collected data, the dependence of blood pressure on various factors. The interplay of determinants, such as physiological measurements, medical attributes, demographic and socioeconomic characteristics on systolic and diastolic blood pressure is investigated.

Repeated measures data is very frequently used, especially in the clinical trial studies. Unfortunately, their analysis is often restricted to standard statistical techniques, like simple analysis of endpoints, analysis of covariance, analysis of increments or analysis of the area under the curve. This, in this end, leads to loss of information, which could be used to explore not only cross-sectional but also longitudinal trends. Very often, longitudinal data is the only way to discover the evolution over time and the changes pertinent to the analyzed phenomena.

In this paper, the practical application of methods accounting for the presence of the correlation in the longitudinally collected data is presented. For these purposes, Generalized Linear Mixed-Effects Model (GLMM) is applied to real-life empirical data.

2. Rationale and description of the problem

Hypertension is a chronic health condition prevalent in most developed nations. Its prevalence in the western countries exceeds 20%. Untreated high blood pressure is a major risk factor for coronary heart disease, cardiovascular disease, stroke or diabetes. This risk could be minimized by better knowledge of blood pressure determinants.

Modelling blood pressure and identifying its significant predictors support the decision making process concerning hypertension diagnosis and its treatment.

The precision of a single blood pressure measurement could not be adequate or could change over time. In such settings, it is better to measure such response multiple times. Repeated measurements for each of the experimental units (subjects) are encountered very often and are used especially to assess the evolution of a given endpoint over time. When data is organized in this manner, the realizations are not longer independent. It implies the correlation among the responses on the same subject over time.

The correlations implied by the presence of repeated measurements on the same experimental unit seem to have two aspects and two sources of variability. The first one is that measures on the same subject are correlated simply because they share common contribution from the subject. The second aspect is that measures on the same subject close in time are often more highly correlated than measures far apart in time. The common mistake is that the variability within subjects is incorrectly assumed to represent the variability among subjects. The repeated measures are not independent samples from the population of interest. They are repeated measurements on the same experimental unit. The response of a specific experimental unit at measurement can provide information about the response of the same experimental unit at re-measurement. Therefore, the analysis of the repeated measures data should consider the presence of correlation between the measurements obtained on the same subject and for possible non-constant variability. Of note is here that also variances of repeated measures often change in time. These potential patterns of correlation and variation may combine to produce a complicated covariance structure of repeated measures. This means that for analyzing such data, a special methodology needs to be applied.

3. Description of NHANES data used for analysis

The data set used in this paper is obtained from the National Health & Nutrition Examination Survey (NHANES). NHANES is an ongoing program designed to assess the health status of patients in the United States. The data comes from 2009 and involves 5430 patients.

The data used for the modelling of blood pressure exhibits following features:

- 3 repeated measurements for both systolic and diastolic blood pressure obtained on the same experimental unit (subject);
- balanced nature of the design, i.e. measurements are taken at fixed time points, and there is an equal number of measurements available for all subjects;
- the response variable of continuous nature.

The consecutive measurements of systolic and diastolic blood pressure were taken in sitting position at an interval of five minutes (after allowing the person to rest for a period of five minutes). The kernel density plots for 3 consecutive blood pressure measurements are presented below. For kernel density estimation, please refer to [1].

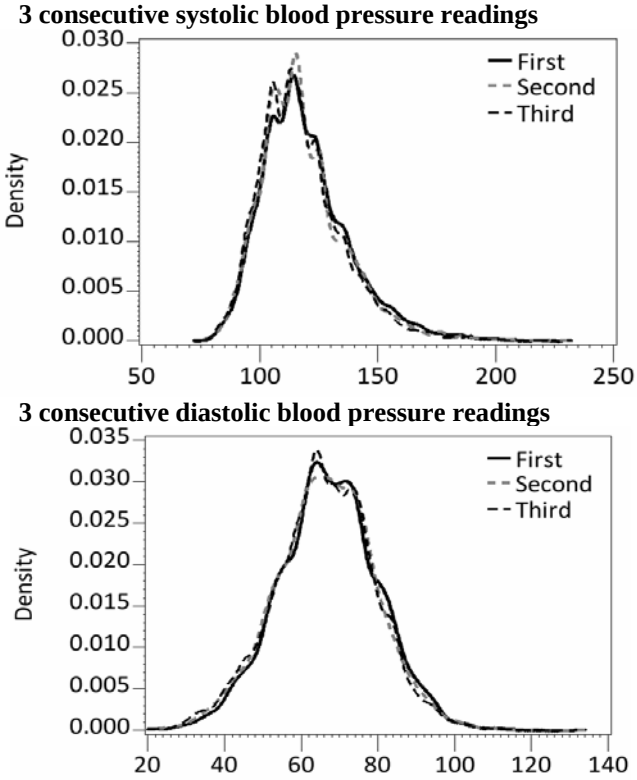

Figure 1. Kernel Density for 3 consecutive readings of blood pressure

Predictor measurements are mainly variables representing medical, demographic and socioeconomic characteristics. Ratio of Income to Poverty compares a family's income to their appropriate poverty threshold. Below explanatory variables are used for modelling systolic and diastolic blood pressure:

- categorical (classification) variables: age cohort (years), gender (male, female), Body Mass Index (BMI) group (kg/m²);
- continuous variables: uric acid (mg/dL), HDL-cholesterol (mmol/L), gamma glutamyl transferase (U/L), family poverty income ratio[1], glucose (mmol/L), creatinine (umol/L).

Age at screening is classified into 4 age cohorts: below 20 years, 20-40 years, 40-60 years and above 60 years, and Body Mass Index (BMI) into 3 BMI groups: below 24 kg/m², 24-30 kg/m², above 30 kg/m². Note that the associations observed between

[1] Ratio of '1' means living right at the poverty line (income at 100% of poverty level), ratio above '1' indicates living above the official definition of poverty).

continuous explanatory variables do not indicate any specific trends. Only a weak correlation exists between explanatory variables. The impact of multicollinearity (concurvity) on parameter estimates should not be an issue.

4. General Linear Mixed-Effects Models (GLMM)

General Linear Mixed-Effects Models (GLMM) extend the ordinary linear regression model by allowing one to incorporate the lack of independence between observations and to model more than one error term. They allow one to derive population estimates of intercept and slope as fixed effects with individual estimates of intercept and slope entered as random effects. The model assumes a continuous outcome variable which is linearly related to a set of explanatory variables – both fixed and random.

General Linear Mixed-Effects Models (GLMM) result from combing a two-stage approach, following below steps:

- stage 1: regression model for each subject separately. A straight line fits the observed responses for each subject – studying variability within subjects;
- stage 2: regression model relating the mean of the individual intercepts (and slopes) estimated in stage 1 to subject-specific covariates – studying variability between subjects [2].

After combining the two stages, General Linear Mixed-Effects Model (GLMM) is defined as below [3]:

$$\begin{cases} Y_i = Z_i\,\beta_i + \varepsilon_i \\ \beta_i = K_i\beta + b_i \end{cases} \Rightarrow Y_i = Z_i\,K_i\beta + Z_i\,b_i + \varepsilon_i \tag{1}$$

or

$$Y_i = X_i\beta + Z_i\,b_i + \varepsilon_i, \tag{2}$$

where

- Y_i is a n_i-dimensional vector of the observed continuous responses for the i-th subject, i.e. $Y_i = \begin{bmatrix} Y_{i,1} \\ Y_{i,2} \\ \vdots \\ Y_{i,n_i} \end{bmatrix}$ and (N = number of subjects);
- $X_i = Z_i\,K_i$ is a $(n_i \times p)$ dimensional fixed effects design matrix (p = number of fixed effects);
- β is a p-dimensional vector of fixed effects parameters used to model $E(Y_i)$;

- Z_i is a $(n_i \times q)$ dimensional random effects design matrix (q = number of random effects) – contains known values for q variables for random effects;
- b_i is q -dimensional vector of random effects (latent variables) used to model the within-subject correlation structure, $b_i \sim N(0, D)$;

n_i

- ε_i is a -dimensional vector of residual components.

In terms of the analyzed blood pressure NHANES data, the fixed effects design matrix, X_i, is a 3×11 dimensional matrix ($n_i = 3, p = 11$), which represents 11 covariates corresponding to the fixed effects for each observation of the i -th subject. The fixed effects design matrix could be defined as:

$$X_i = \begin{bmatrix} x_{i,1}^{(1)} & x_{i,1}^{(2)} & \dots & x_{i,1}^{(11)} \\ x_{i,2}^{(1)} & x_{i,2}^{(2)} & \dots & x_{i,2}^{(11)} \\ x_{i,3}^{(1)} & x_{i,3}^{(2)} & \dots & x_{i,3}^{(11)} \end{bmatrix} \tag{3}$$

where the fixed effects defined by $x_{i,n_i}^{(p)}$ refer to population intercept, time point, age cohort, gender, BMI group, uric acid, HDL-cholesterol, gamma glutamyl transferase, family poverty income ratio, glucose, creatinine respectively. As among the explanatory variables in the design matrix X_i there is time point, this matrix links both within- and between-subject covariates to the fixed effects parameters. The fixed effects vector, β, consists of 11 unknown regression coefficients associated with the covariates from the design matrix X_i and is defined as $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{11} \end{bmatrix}$.

For the analyzed NHANES data, it is assumed that random intercept is the only random effect which might make the variability to vary over time. Thus, the random effects design matrix, Z_i, which represents covariates corresponding to the random effects for each observation of the i -th subject, is just 3×1 dimensional matrix ($n_i = 3, q = 1$):

(4)

The random effects are effects that vary randomly across subjects. Hence, they include the individual differences for the subjects. In case of the model proposed for

NHANES data, the random effects vector, b_i consists only of one random effect which is associated with the covariates from the design matrix Z_i and is defined by $b_i = [b_{1i}]$. Note that in General Linear Mixed-Effects Models (GLMM) the number of observations, n_i, for each subject may differ – n_i is a subject-specific – and consequently each subject could have different number of random effects. In the case of NHANES data, this is simplified to balanced design.

It is assumed that the random effects vector, b_i, follows a multivariate normal distribution, $b_i \sim N(0, D)$ where D is a $(q \times q)$ dimensional positive definite symmetric covariance matrix. As random intercept is the only random effect used to model the variability, D matrix is simplified to 1×1 form,

i.e. .

The residual vector ε_i for data with 3 repeated occasions ($n_i = 3$) is defined by $\varepsilon_i = \begin{bmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \varepsilon_{i,3} \end{bmatrix}$, where each element represents the residual associated with each response for the i-th subject. Unlike the residuals in standard linear models, the residuals associated with repeated observations on the same subject in a linear mixed effects model can be correlated. It is assumed that the residuals follow a multivariate normal distribution, $\varepsilon_i \sim N\left(0, \sum_i \square\right)$ where $\sum_i \square$ is a $(n_i \times n_i)$ dimensional positive definite symmetric covariance matrix. The vectors of residuals ε_i and random effects b_i are independent on each other [4]:

$$\sum_i \square = Var(\varepsilon_i) = \begin{bmatrix} Var(\varepsilon_{i,1}) & Cov(\varepsilon_{i,1}, \varepsilon_{i,2}) & Cov(\varepsilon_{i,1}, \varepsilon_{i,3}) \\ Cov(\varepsilon_{i,1}, \varepsilon_{i,2}) & Var(\varepsilon_{i,2}) & Cov(\varepsilon_{i,2}, \varepsilon_{i,3}) \\ Cov(\varepsilon_{i,1}, \varepsilon_{i,3}) & Cov(\varepsilon_{i,2}, \varepsilon_{i,3}) & Var(\varepsilon_{i,3}) \end{bmatrix} \quad (5)$$

General Linear Mixed-Effects Models (GLMM) for systolic/diastolic blood pressure boil down to:

$$Y_i = \begin{bmatrix} Y_{i,1} \\ Y_{i,2} \\ Y_{i,3} \end{bmatrix} = \begin{bmatrix} x_{i,1}^{(1)} & x_{i,1}^{(2)} & \dots & x_{i,1}^{(11)} \\ x_{i,2}^{(1)} & x_{i,2}^{(2)} & \dots & x_{i,2}^{(11)} \\ x_{i,3}^{(1)} & x_{i,3}^{(2)} & \dots & x_{i,3}^{(11)} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{11} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad (6)$$

where Y_i is a vector of systolic/diastolic blood pressure readings for the i -th subject and all other components are as defined earlier. Note that, the vector of random effects is restricted to the single intercept b_{i1}, thus General Linear Mixed-Effects Model (GLMM) for blood pressure data could be referred to as Random-Intercept Model. The random component b_{i1} is subject i 's predicted deviation from the population intercept – this model is therefore subject-specific and is meaningful under hierarchical model interpretation. Components b_i and ε_i are the random error terms, all the other terms give the regression model for the mean response implied by the two-stage model.

The proposed model with both fixed effects β and with subject-specific effects b_i assumes that the vector of repeated measurements follows a linear regression model for each subject where some of the regression parameters are population-specific (the same for all subjects) whereas other parameters are subject-specific. For the estimation of this model, the restricted log likelihood of the data is formed, given the fixed-effects matrix. The methods of parameter estimation of General Linear Mixed-Effects Models (GLMM) have been widely addressed in the literature, with the most commonly used approaches being restricted maximum likelihood (REML) estimation or maximum likelihood (ML) estimation [3].

Fitting General Linear Mixed-Effects Models (GLMM) requires specification of a mean structure (covariates, time effects, interactions), as well as covariance structure (random effects, serial correlation). Both components affect each other. Unless robust inference is used, an appropriate covariance model is essential to obtain valid inferences for the parameters in the mean structure, which is usually of primary interest [5]. Too restrictive specifications invalidate inferences when the assumed structure does not hold (invalid inferences for the mean structure), whereas overparametrization of the covariance structure could lead to inefficient estimation and poor assessment of standard errors (inefficient inferences for mean). Note that in this paper, due to limited space, only a brief summary of these points with regards to analyzed NHANES data is made. The proposed model includes age cohort, gender, BMI group, uric acid, HDL-cholesterol, gamma glutamyl transferase, family poverty income ratio, glucose, creatinine. As checked, this structure could not be reduced to more parsimonious one. Exclusion of any of the predictors results in significant reduction of -2 log likelihood, what in the end gives sufficient evidence to reject the null hypothesis of the reduced model in favor of the most general or saturated structure. It is believed that, apart from random intercept, there are no other random effects which could make the variability to vary over time. There is no need to include slope for time effects (neither linear nor quadratic). Higher order random effects are also not needed – only random intercept is statistically significant.

Below tables present the estimation results of General Linear Mixed-Effects Models (GLMM) for systolic and diastolic blood pressure data.

Table 1. Solution for Fixed Effects – systolic and diastolic blood pressure

	Systolic blood pressure		***Diastolic blood pressure***	
Effect	***Estimate (Std. Error)***	***t-value (p-value)***	***Estimate (Std. Error)***	***t-value (p-value)***
Intercept	94,50 (1,48)	63,98 (<0,01)	55,28 (1,09)	50,43 (<0,01)
BMI (above 30 vs below 24)	3,51 (0,59)	5,90 (<0,01)	2,98 (0,44)	6,75 (<0,01)
BMI (24-30 vs below 24)	2,54 (0,53)	4,74 (<0,01)	1,14 (0,39)	2,89 (<0,01)
Gender (male vs female)	3,44 (0,48)	7,19 (<0,01)	2,93 (0,35)	8,25 (<0,01)
Age (above 60 vs below 20)	20,82 (0,71)	29,53 (<0,01)	6,32 (0,52)	12,08 (<0,01)
Age (40-60 vs below 20)	10,60 (0,68)	15,71 (<0,01)	13,25 (0,50)	26,48 (<0,01)
Age (20-40 vs below 20)	4,22 (0,64)	6,59 (<0,01)	8,51 (0,48)	17,91 (<0,01)
Uric acid	0,75 (0,18)	4,27 (<0,01)	0,18 (0,13)	1,37 (0,17)
HDL-cholesterol	2,11 (0,55)	3,86 (<0,01)	0,52 (0,41)	1,28 (0,20)
Gamma glutamyl transferase	0,03 (0,01)	4,78 (<0,01)	0,02 (0,00)	4,63 (<0,01)
Family poverty income ratio	-0,71 (0,13)	-5,53 (<0,01)	0,01 (0,09)	0,07 (0,94)
Glucose	0,68 (0,12)	5,72 (<0,01)	-0,07 (0,09)	-0,84 (0,39)
Creatinine	0,02 (0,01)	2,97 (<0,01)	-0,01 (0,00)	-1,54 (0,12)
Time point (first vs third measurement)	2,30 (0,08)	28,74 (<0,01)	1,24 (0,08)	15,95 (<0,01)
Time point (second vs third measurement)	0,92 (0,08)	11,45 (<0,01)	0,29 (0,08)	3,74 (<0,01)

Note: Estimation methods: Restricted Maximum Likelihood (REML), Covariance parameters: 3, Columns in X: 19, Columns in Z per subject: 1, Number of subjects: 5430, Max obs. per subject: 3.

Table 2. Covariance Parameter Estimates – systolic and diastolic blood pressure

	Systolic blood pressure	***Diastolic blood pressure***
Covariance Parameter	***Estimate***	***Estimate***
Intercept	215,40	116,25
Time point (factor)	16,4476	15,4450
Residual	0,9998	0,9995

The estimated model could be perceived in both conditional and unconditional sense. If the interest is in quantities averaged over all possible values of the random effects, then the focus is on the marginal formulation. Although in practice one is usually interested in estimating the parameters in marginal linear mixed-effects model, it is often useful to calculate estimates for the random effects b_i. Under the hierarchical model interpretation, estimation of the random effects b_i is helpful for detecting outlying profiles [6]. This approach is particularly useful in clinical trials where interest is in drug efficacy for a particular patient (conditional formulation). It needs to be pointed out that for the analyzed blood pressure data, huge between subject variability in comparison to the within subject variability is observed. In such settings, hierarchical model interpretation, i.e. subject-specific interpretation rather than marginal one, is more reasonable.

The estimate of subject-specific blood pressure could be seen as a weighted average of the estimated overall effect averaged over all subjects combined and the observed effect of a particular subject. The weight is determined by the relative sizes of the between-subject and within-subject variances. In case of the NHANES data, the estimated overall average effect size is not a good summary measure for all the subjects combined. The between-subject variance is relatively large as compared to the within-subject variances meaning that there are large deviations from the population mean.

In order to identify outlying profiles or group of individuals evolving differently in time, which deserve better attention by the researcher, one needs to focus on the deviations from the overall mean - how much the subject-specific profiles deviate from the overall average profile. Random effects, b_i, reflect how the evolution for the i th subject deviates from the expected evolution $X_i\beta$. These estimates could be interpreted as marginal residuals, i.e. $Y_i - X_i\hat{\beta}$. For the identification of outlying observations marginal studentized residuals were used.

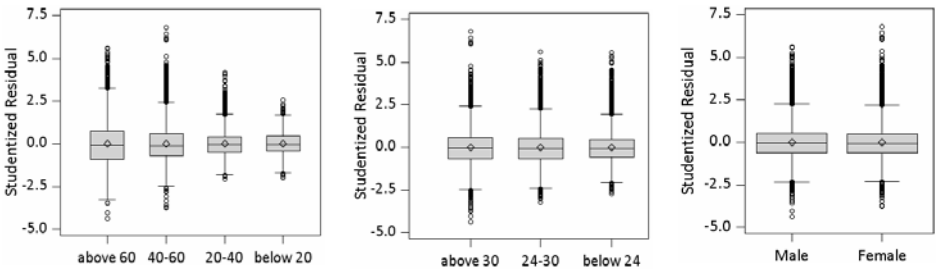

Figure 2. Box Plot of Studentized Marginal Residuals – on the example of systolic blood pressure

Systolic blood pressure

Quantile	***Estimate***
99%	3,1268
95%	1,7181
90%	1,1985
75% Q3	0,5162
50% Median	-0,0684
25% Q1	-0,6269
10%	-1,1256
5%	-1,4631
1%	-2,1964

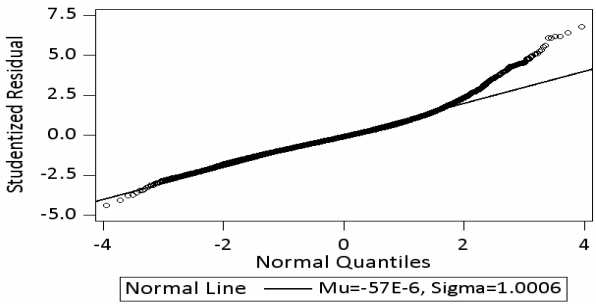

Figure 3. Q-Q Plot for Studentized Marginal Residuals - on the example of systolic blood pressure

5. Discussion – interpretation of the results

The relationship between blood pressure and analyzed explanatory variables is different for systolic and diastolic blood pressure. Model for systolic blood pressure indicates that all the explanatory variables influence the response. They are statistically significant at the assumed significance level of 0,01. This relationship is not so clear in case of diastolic blood pressure where only one continuous predictor - gamma glutamyl transferase - significantly impacts blood pressure.

The estimates are consistent with previous studies concerning systolic blood pressure involving the application of nonlinear splines in the response-predictor relationship over the averaged responses obtained using Generalized Additive Models [7]. It is expected however, that the analysis accounting for the presence of correlation and avoiding the averaging of the responses, which leads to loss of information, is more powerful to detect differences and relationships present in the data.

The risk of high systolic blood pressure is higher among older patients. On average, the systolic blood pressure for subjects of age above 60 is more than 20 mmHg higher than for subjects of age below 20 years. Overweight subjects (those with BMI 24-30 kg/m²) and obese (those with BMI above 30 kg/m²) have significantly higher blood pressure than subjects with BMI below 24 kg/m². On average, systolic blood pressure for men (as compared to females) is greater by 3.5 mmHg and diastolic blood pressure greater by 3.

People with elevated uric acid level are at greater risk of high systolic blood pressure. It is also supported by other studies on uric acid [8]. Effective drugs already exist which lower the level of uric acid and, by these means, are intended to lower blood pressure level. HDL-cholesterol increases systolic blood pressure by about 2 mmHg for every 1 unit increase (mmol/L). This link is not significant for diastolic blood pressure readings. Gamma glutamyl transferase, a marker of oxidative stress, is positively related to both systolic and diastolic blood pressure. More susceptible to high systolic blood pressure are also subjects in low socioeconomic status i.e. with lower income as assessed by family poverty income ratio. These findings seem to be pretty worrying, as most researches on hypertension are focused on developed urban countries. Very little is known about hypertension treatment and its diagnosis outside high-income areas. In low-income regions, high blood pressure is the major risk factor for cardiovascular diseases. Higher blood pressure is associated with higher serum creatinine level (an indicator of chronic renal disease) and an increase in glucose level. Elevated glucose increases the chance of having diabetes and this probably leads to higher systolic blood pressure.

For subjects with high random effects values (in magnitude), the probability of having high systolic/diastolic blood pressure is much more due to the patient’s uncharacterized "frailty" than to fixed effects. Large number of outlying values and large variability among random effects indicate that there is wide between-individual heterogeneity. This additionally speaks for the application of random effects approach and allowing them to represent natural heterogeneity between subjects.

6. Concluding remarks

In this paper, the practical use of General Linear Mixed-Effects Model (GLMM) for analysis of the longitudinal data was demonstrated. Modelling correlation among measurements made on the same experimental unit was performed using random effects. The real-life empirical example showed the robustness and flexibility of General Linear Mixed-Effects Models (GLMM) for the analysis of repeated measurements data with continuous response variable. Used methodology is much more appropriate as compared to traditional methods like separate ANOVA at each time point, or multivariate approach employing unstructured-covariance matrix. The use of General Linear Mixed-Effects Models (GLMM) in clinical trials should be encouraged, given their capability to model the within-cluster correlation.

References

[1] Kulczycki, P.: *Estymatory jadrowe w analizie systemowej*, WNT, Warszawa, 2005
[2] Verbeke, G., Molenberghs, G.: *Linear Mixed Models for Longitudinal Data*, Springer, New York, 2000
[3] Laired, N., Ware, J.: *Random-effects models for longitudinal data*, Biometrics, pp. 963-974, 1982
[4] West, B., Welch, K., Galecki, A.: *Linear Mixed Models - A Practical Guide Using Statistical Software*, Chapman & Hall/CRC Taylor & Francis Group, London, 2007
[5] Molenberghs, G., Verbeke, G.: *Meaningful statistical model formulation for repeated measures*, Statistica Sinica, pp. 989-1020, 2004
[6] Molenberghs, G., Verbeke, G.: *Models for Discrete Longitudinal Data*, Springer, New York, 2005
[7] Kruszewski, D.: *Nonparametric modelling of medical scheme data*, Technical Transactions, Automatic Control, vol. 110 (1-AC/2013), pp. 93-117, 2013
[8] Feig, D., Kang, D., Nakagawa, T., Mazzali, M., Johnson, R.: *Uric acid and hypertension*, Curr Hypertens Rep., pp. 111-115, 2006
[9] Official page of National Health & Nutrition Examination Survey (data source): http://www.cdc.gov/nchs/nhanes.htm