

UNIVERSITY OF TARTU
FACULTY OF SCIENCE AND TECHNOLOGY
INSTITUTE OF MATHEMATICS AND STATISTICS

Maia Arge
Generalized Estimating Equations: an overview and
application in IndiMed study

Mathematical statistics
Master's thesis (30 EAP)

Supervisors:
Kristi Läll, MSc
Pasi Korhonen, PhD
Krista Fischer, PhD

TARTU
2016

Generalized Estimating Equations: an overview and application in IndiMed study

Master's thesis

Maia Arge

Abstract. Generalized Estimating Equations (GEE), developed by (Zeger & Liang 1986), is a method of estimation that accounts for correlations among repeated measurements and is widely used in longitudinal analysis. The purpose of this master's thesis is to provide an overview of GEE and use this approach on a cluster-randomized study called IndiMed. The data are from the Estonian Genome Center of University of Tartu. The clusters are doctors who were randomized to intervention or control group and subjects are patients with high blood pressure. All subjects in one cluster are either in intervention (received genetic risk information on the 2nd visit) or control group (received risk information on the 4th visit). Systolic and diastolic blood pressures were measured on all 5 visits for all patients and compared for the two study groups, taking into account the correlation among the repeated measurements.

CERCS research specialisation: P160 Statistics, operations research, programming, actuarial mathematics

Keywords: Generalized Estimating Equations, marginal models, repeated measurements, cluster-randomized trial, hypertension

GEE: ülevaade ja rakendamine IndiMedi uuringus

Magistritöö

Maia Arge

Lühikokkuvõte. GEE, mille pakkusid välja (Zeger & Liang 1986), on longituudsetes uuringutes laialt kasutatav meetod, mis võtab arvesse kordusmõõtmiste omavahelisi korrelatsioone. Magistritöö eesmärgiks on anda GEE meetodist ülevaade ja rakendada seda IndiMedi uuringus, mille andmed pärinevad Tartu Ülikooli Eesti Geenivaramust. Klasterid moodustavad arstid, kes randomiseeriti sekkumis- või kontrollgruppi ning uuringus osalesid kõrge vererõhuga patsiendid. Kõik samas klasteris olevad patsiendid kuuluvad kas sekkumisgruppi (said geneetilise riski info teisel visiidil) või kontrollgruppi (said geneetilise riski info neljandal visiidil). Gruppide võrdluseks kasutatakse kokku viiel visiidil mõõdetud süstoolset ja diastoolset vererõhku, arvestades kordusmõõtmiste omavahelisi korrelatsioone.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Märksõnad: GEE mudelid, marginaalsed mudelid, kordusmõõtmised, klaster-randomiseeritud uuring, hüpertensioon

Table of contents

1. Introduction	4
2. Generalized Estimating Equations (GEE)	5
2.1 Introduction to Generalized Estimating Equations	5
2.2 Quasi-likelihood estimator	6
2.3 Marginal models.....	7
2.4 Correlation structures	9
2.5 Estimation of parameters.....	11
2.6 Sandwich estimator	11
2.7 Model inference and interpretation of parameters	13
2.8 Quasi-likelihood Information Criterion (QIC).....	13
2.9 Summary of GEE	14
3. Using GEE method of estimation in SAS	15
3.1 PROC GENMOD.....	15
3.2 PROC GLIMMIX	16
4. Analysis of the IndiMed trial.....	18
4.1 IndiMed study design	18
4.2 Descriptive statistics.....	20
4.3 Analysis.....	24
4.3.1 Intervention vs control group.....	25
4.3.2 Comparison of intervention group individuals with high genetic risk to individuals at low risk and to the control group	28
4.4 Discussion	30
5. Conclusion	31
6. References	32
7. Appendix	34
7.1 Correlation structures	34
7.2 PROC GLIMMIX contrasted with PROC GENMOD	35
7.3 IndiMed study design and timeline	36
7.4 Tables of descriptive statistics	37
7.5 Distribution of systolic and diastolic blood pressures.....	38
7.6 Tables of analysis results.....	40

1. Introduction

The purpose of this thesis is to provide an overview of Generalized Estimating Equations (GEE) and apply this method of estimation to a cluster-randomized study called IndiMed.

This master's thesis consists of three parts: overview of Generalized Estimating Equations, a general guideline of how to use GEE in SAS and thirdly a practical part where GEE-type models are applied to data from the Estonian Genome Center of University of Tartu.

Theoretical part provides an overview of the Quasi-likelihood estimator and marginal models which are the basis of GEE. Different correlation structures, estimation of parameters and of their variance is presented. Model inference, interpretation and model selection criteria is also discussed.

In the next section, instructions of using the GEE approach in statistical software SAS are presented for two frequently used procedures: PROC GENMOD and PROC GLIMMIX. Example codes are explained in detail for both procedures.

The applied part of this thesis illustrates the theoretical part by fitting GEE-type models where the data are from a cluster-randomized study called IndiMed. The clusters are doctors who were randomized to intervention or control group and subjects are men with high blood pressure. All subjects in one cluster are either in intervention (received genetic risk information on the 2nd visit) or control group (received risk information on the 4th visit). Systolic and diastolic blood pressures were measured on all of the 5 planned visits. GEE modelling is done to compare the two groups. Finally, results and discussion is presented to conclude the findings.

The author of this thesis would like to express gratitude to the supervisor Kristi Läll for being patient, motivating and guiding throughout the writing of this thesis. Thanks to supervisor Pasi Korhonen for critical comments and overall ideas and thanks to supervisor Krista Fischer from the Estonian Genome Center of University of Tartu for general comments and helping to direct. Final thanks to the whole IndiMed team for the data.

2. Generalized Estimating Equations (GEE)

2.1 Introduction to Generalized Estimating Equations

Longitudinal designs, which involve repeated measurements of a response variable in each subject, are often used in clinical trials and epidemiology. In clinical trials, repeated measurements are usually observed over a short period of time, whereas studies in epidemiology may last for many years. The interest in analysing longitudinal data is to describe the dependence of the outcome on predictor variables (marginal expectation of the outcome Y as a function of covariates X , i.e. $E(Y|X)$). Repeated measurements tend to be correlated since they are made on the same subject. For example, two measurements from the same subject are likely to be more correlated than two measurements from different subjects. Since most statistical tests assume independence of observations, it is crucial to take the within-subject correlation into account to obtain correct statistical analysis. If we do not take the correlation into account, it can lead to a wrong test statistic and inference, for example standard errors will likely be too small. (Burton et al. 1998) (Zeger & Liang 1986)

There are many techniques for analysis when outcome variable is approximately normally distributed (e.g. fitting growth curves for each subject using repeated measurements). When the outcome variable is not approximately normally distributed, there are fewer analysis methods available. Difficulty in analysis comes from the lack of multivariate joint distribution of the outcome variable, hence likelihood methods are not available or are difficult to compute. (Liang & Zeger 1986)

Linear models for normally distributed data have been expanded to non-normal data using generalized estimation methods and quasi-likelihood when there is a single observation for each subject (no repeated measurements). Quasi-likelihood approach does not assume distribution, it only specifies a linear function between marginal expectation of the outcome variable and covariates and assumes that variance (of the outcome variable) is a known function of its expectation. (Zeger & Liang 1986)

Generalized Estimating Equations are based on quasi-likelihood method of estimation. In addition to previously mentioned assumptions (expectation of outcome variable to be a linear function of covariates and that variance is a known function of the mean), one needs

to specify correlation structure between repeated measurements for each subject. The general idea is to incorporate the correlation structure between repeated measurements to get consistent estimators of coefficients and of their variances. (Zeger & Liang 1986)

Before introducing GEE-type models, the next section will give the general idea of quasi-likelihood estimator of which GEE is based on.

2.2 Quasi-likelihood estimator

This section is a short overview of quasi-likelihood used in GEE based on (Zeger & Liang 1986). Quasi-likelihood is a methodology for regression that requires the specification of relationships between mean response and covariates and between mean response and variance. Thus it does not assume a probability distribution as in the case of full likelihood.

Let Y_i be the response variable for each subject $i = 1, \dots, N$ and X_i be $p \times 1$ vector of covariates. Let β be $p \times 1$ vector of regression parameters to be estimated. Define $\mu_i = E(Y_i|X_i)$ to be the conditional expectation of Y_i and a function of covariates and regression parameters, so that $\mu_i = h(X_i^T \beta)$. The inverse of h is the link function which relates the mean response to the linear predictor $X_i^T \beta$. For quasi-likelihood, variance of each Y_i , denoted as v_i , is a known function of the expectation μ_i , so that $v_i = f(\mu_i) \cdot \Phi$. The scale parameter Φ is treated as a nuisance parameter. The quasi-likelihood estimator is the solution to the equations:

$$S_k(\beta) = \sum_{i=1}^N \frac{\partial \mu_i}{\partial \beta_k} v_i^{-1} (Y_i - \mu_i) = 0, k = 1, \dots, p.$$

Estimators of regression parameters, $\hat{\beta}$, are obtained by iteratively reweighted least squares method.

2.3 Marginal models

A standard GEE is known as a marginal model. Marginal models extend generalized linear models to longitudinal data and are typically used when the inference is population-based, rather than individual-based. The term “marginal” means that in the model specification the expected value of the response variable Y , depends only on covariates (fixed effects) and does not depend on subject specific random effects nor directly on previous responses of the subject. Since the purpose is to describe the changes in population mean rather than changes within subjects, within-subject correlation is regarded as a nuisance characteristic. Regression parameters and within-subject correlation is modelled separately. (Fitzmaurice et al. 2004)

Let's introduce some notation for the repeated measurements. We have N subjects who are measured repeatedly. Y_{ij} denotes the response variable for the i^{th} subject on the j^{th} measurement occasion. A realisation of each Y_{ij} is observed at time t_{ij} . The response variable can be continuous, binary, multinomial or a count. We assume the data are unbalanced (the number of repeated measurements can be different for subjects and/or they can be measured at different occasions) and that there are n_i repeated measurements for the i^{th} subject.

The response variable is a $n_i \times 1$ vector

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}, i = 1, \dots, N.$$

Y_i are assumed to be independent, but observations within the subject are not assumed to be independent. Associated with each response at a given time point j , there is a $p \times 1$ vector of covariates

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, i = 1, \dots, N; j = 1, \dots, n_i$$

which can be collected into a $n_i \times p$ matrix of covariates

$$X_i = \begin{pmatrix} X_{i1}^T \\ X_{i2}^T \\ \vdots \\ X_{in_i}^T \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{pmatrix}, i = 1, \dots, N.$$

They can be either time-invariant or time-dependent. Time-invariant variable is fixed within a subject at the same value irrespective of time point j , whereas time-dependent variable is varying with time for each subject.

The GEE requires the following specifications for a marginal model:

- 1) Conditional mean of Y_{ij} is related to covariates by a known link function,

$$g(\mu_{ij}) = \eta_{ij} = X_{ij}^T \beta,$$

where $\mu_{ij} = E(Y_{ij}|X_{ij})$ is a conditional expectation (or mean) of the response variable and β is a $p \times 1$ vector of regression parameters.

- 2) The conditional variance of each Y_{ij} may depend on the mean response, given the effects of covariates, as

$$Var(Y_{ij}) = \phi v(\mu_{ij})$$

where the variance function $v(\mu_{ij})$ is a known function of the mean μ_{ij} and ϕ is a scale parameter that may be known or may need to be estimated. For balanced longitudinal designs, a separate scale parameter, ϕ_j , could be estimated at the j^{th} occasion. Alternatively, ϕ could depend on the times of measurement, where $\phi(t_{ij})$ is some parametric function of t_{ij} . When the response is a continuous variable, then variance of each Y_{ij} does not depend on mean response and is $Var(Y_{ij}) = \phi v(\mu_{ij}) = \phi$. Note that this assumes homogeneity of variance over time, which is often too strong of an assumption.

- 3) Correlation among repeated measurements is a function of the means, μ_{ij} , and a set of parameters, α , which characterize the within-subject correlation and need to be estimated. The “working” covariance matrix for Y_i is given by

$$V_i = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}} / \phi.$$

Correlation matrices $R_i(\alpha)$ can be different for subjects, however, it is fully specified by α , which is the same for all subjects. A_i is a $n_i \times n_i$ diagonal matrix with $g(\mu_{ij})$ on the diagonal. “Working” covariance means that we do not know the true correlation structure between repeated measurements and are not assuming we are specifying it correctly. We would like to get consistent estimates of regression parameters regardless of the chosen structure. (Fitzmaurice et al. 2004) (Zeger & Liang 1986)

2.4 Correlation structures

Here we present five basic correlation structures (see Figure 7.1-1 in the appendix for illustration).

Independence

The most basic structure is independence, where each observation within a subject is uncorrelated with another observation. An example of a 4x4 correlation matrix with independence structure is the following

$$R_{IN} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \text{ where } Cor(Y_{ij}, Y_{ik}) = \begin{cases} 1, & \text{if } j = k \\ 0, & \text{if } j \neq k \end{cases}$$

Exchangeable

In exchangeable correlation structure, responses are assumed to be equally correlated within an individual.

$$R_{EX} = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}, \text{ where } Cor(Y_{ij}, Y_{ik}) = \begin{cases} 1, & \text{if } j = k \\ \alpha, & \text{if } j \neq k \end{cases}$$

Autoregressive AR-1

Two observations closer in time are correlated more than two observations more further in time. This structure is often used in longitudinal designs. Note that n_i in this example is 4.

$$R_{AR-1} = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{pmatrix}, \text{ where } Cor(Y_{ij}, Y_{i,j+t}) = \alpha^t, \text{ for } t = 0, \dots, n_i - j.$$

Toeplitz

Correlation is the same for any two observations that have the same distance in time. Note that n_i in this example is 4.

$$R_{TOEP} = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1 & 1 & \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_1 & 1 & \alpha_1 \\ \alpha_3 & \alpha_2 & \alpha_1 & 1 \end{pmatrix}, \text{ where } Cor(Y_{ij}, Y_{i,j+t}) = \begin{cases} 1, & \text{if } t = 0 \\ \alpha_t, & \text{if } j = 1, \dots, n_i - t \end{cases}$$

Unstructured

There is no assumption made about any two observations within a subject, so correlation can take a value between -1 and 1. This type of correlation is the most flexible one, but the number of parameters can become too high very quickly.

$$R_{UN} = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{12} & 1 & \alpha_{23} & \alpha_{24} \\ \alpha_{13} & \alpha_{23} & 1 & \alpha_{34} \\ \alpha_{14} & \alpha_{24} & \alpha_{34} & 1 \end{pmatrix}, \text{ where } Cor(Y_{ij}, Y_{ik}) = \begin{cases} 1, & \text{if } j = k \\ \alpha_{jk}, & \text{if } j \neq k \end{cases}$$

2.5 Estimation of parameters

The GEE estimator is the solution of the system

$$\sum_{i=1}^N D_i^T V_i^{-1} S_i = 0,$$

where $S_i = Y_i - \mu_i$ with $\mu_i = (\mu_{i1}, \dots, \mu_{in_i})^T$ and $D_i = \frac{\partial \mu_i}{\partial \beta}$. Equations reduce to quasi-likelihood equations when there is only one measurement for the outcome variable ($n_i = 1$) for each subject. Since estimating equations depend on β and correlation parameter α , and generally have no closed-form solution, solving is based on a two-stage iterative method. (Zeger & Liang 1986)

Parameters of β are estimated by solving the equation starting with $\phi = 1$ and R_i as an identity matrix. Estimates are used in calculating fitted values $\hat{\mu}_i = g^{-1}(X_i^T \beta)$ and residuals $y_i - \hat{\mu}_i$. These are used to estimate A_i , R_i and ϕ . Then these steps are repeated to improve $\hat{\beta}$ until convergence.

One of the properties of GEE estimator is that the estimate of β is consistent even when assumed correlation structure among repeated measurements is not correct. That is, $\hat{\beta}$ has the property of being close to the true parameter β in large samples. The unbiasedness of the parameter estimates only depends on the correct specification of the model for mean response.

2.6 Sandwich estimator

The following section is based on (Fitzmaurice et al. 2004).

In the GEE context, usual standard errors of β are not valid when correlation structure among repeated measurements is misspecified (even though the estimate of β is consistent). The variance of the GEE estimator of β can be made robust to the misspecification of correlation structure (in a sense that it provides valid standard errors when correlation structure is not correct) by incorporating the empirical variance estimator, called “sandwich” covariance estimator (Diggle et al. 2002). The robustness of sandwich variance estimator is asymptotic meaning that it is a large sample property.

If the sample size is large, the sampling distribution of $\hat{\beta}$ is multivariate normal distribution with mean β and covariance in the form of

$$Cov(\hat{\beta}) = B^{-1}MB^{-1}$$

where $B^{-1} = (\sum_{i=1}^N D_i^T V_i^{-1} D_i)^{-1}$ and $M = \sum_{i=1}^N D_i^T V_i^{-1} Cov(Y_i) V_i^{-1} D_i$.

The covariance of $\hat{\beta}$ can be estimated by the ‘‘sandwich’’ estimator

$$\hat{\Sigma}_S := \widehat{Cov}(\hat{\beta}) = \left(\sum_{i=1}^N \hat{D}_i^T \hat{V}_i^{-1} \hat{D}_i \right)^{-1} \sum_{i=1}^N \hat{D}_i^T \hat{V}_i^{-1} (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^T \hat{V}_i^{-1} \hat{D}_i \left(\sum_{i=1}^N \hat{D}_i^T \hat{V}_i^{-1} \hat{D}_i \right)^{-1},$$

where α , ϕ and β have been replaced by their estimates in B and M , and $Cov(Y_i)$ has been replaced with $(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^T$ in M .

The model-based covariance estimator is

$$\widehat{Cov}(\hat{\beta}) = \hat{B}^{-1} = \left(\sum_{i=1}^N \hat{D}_i^T \hat{V}_i^{-1} \hat{D}_i \right)^{-1}$$

and it yields valid standard errors when the working correlation is correctly specified.

The use of sandwich estimator ensures that parameter estimates are consistent even though the working correlation structure may not be correct. However, a correct structure can increase efficiency (Burton et al. 1998). It has been shown that misspecification of the correlation structure can reduce efficiency and inflate variance estimates (Wang & Carey 2003). Therefore, it is important to try to choose the structure which is close to the true correlation matrix $R_i(\alpha)$. Since one of the aims of fitting a model is to estimate the parameters which describe the correlation structure, one should choose it before the analysis (Burton et al. 1998).

GEE approach assumes data to be missing completely at random (MCAR, (Rubin 1976), (Laird n.d.)), meaning that the missingness does not depend on observed or unobserved responses. If the data are missing at random (MAR), then GEE method can lead to biased results.

2.7 Model inference and interpretation of parameters

Interpretation of regression coefficients depend on the link function used in modelling. Since within-subject correlation is estimated from residuals, parameter estimates are interpreted in a population-average manner. Formal inferences (statistical significance tests) and parameter estimates' confidence intervals are based upon Wald test. Regression coefficient estimates are divided by their robust (empirical) standard error and the results are treated as Normal standardized deviates (Z statistic). (Burton et al. 1998). It can underperform when sample size is small. Contrasts for a type III test of hypothesis can be tested with generalized score tests.

2.8 Quasi-likelihood Information Criterion (QIC)

Well-known model selection criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) cannot be applied here since these are likelihood-based and full multivariate likelihood is not used in GEE modelling. However, there is a criterion proposed by (Pan 2001) which is based on AIC but with quasi-likelihood approach, constructed under independence working correlation structure. This quasi-likelihood information criteria (QIC) for GEE models is in the form of

$$QIC(R) = -2Q(\hat{\beta}(R), I, \Delta) + tr(\hat{\Sigma}_{M(IN)}^{-1} \hat{\Sigma}_{S(R)}),$$

where $Q(\hat{\beta}(R), I, \Delta)$ is the quasi-likelihood under the assumption of working correlation matrix R , $\hat{\Sigma}_{M(IN)}^{-1}$ is the model-based covariance estimator under independence working correlation structure assumption, $\hat{\Sigma}_{S(R)}$ is the sandwich covariance estimator under the working correlation structure R and tr is the notation for matrix trace. (Jang 2011)

Different models are comparable in the same way as when using AIC – model with smaller QIC value is preferred.

2.9 Summary of GEE

In summary, the GEE approach is a widely used method of estimation in longitudinal designs. Instead of modelling the covariance among repeated measurements, GEE treats it as a nuisance and models the mean response. Covariances and standard errors are valid under the assumption of data missing completely at random.

The GEE has many appealing properties and is often the preferred method for the following reasons. Firstly, even when within-subject correlation structure is not close to the true underlying correlation structure, the GEE estimator $\hat{\beta}$ still gives a consistent estimate. Secondly, the usual standard errors are not valid when correlation structure is misspecified, therefore sandwich estimator of $Cov(\hat{\beta})$ can be used to get valid standard errors. Finally, it is not crucial to assume multivariate normal or any other distribution for the outcome. One needs to specify only the first two moments. The validity of GEE estimates, $\hat{\beta}$, rely only on specifying the correct model for mean response.

3. Using GEE method of estimation in SAS

In this section, two procedures for estimating parameters with GEE in SAS: PROC GENMOD and PROC GLIMMIX are presented. The following chapter is based on SAS User's Guide (SAS Institute Inc. 2011), where more options are available.

3.1 PROC GENMOD

The GENMOD procedure fits generalized linear models, where it is needed to specify the link function and the probability distribution of the response variable. The following example code will be discussed in detail.

```
PROC GENMOD DATA=data DESCENDING;  
CLASS x1(REF='1') id /PARAM=ref;  
MODEL y=x1 x2 x1*x2 /LINK=identity DIST=normal TYPE3 WALD;  
REPEATED SUBJECT=id /WITHINSUBJECT=timepoints TYPE=ar(1) CORRW CORRB;  
RUN;
```

Note that y is the response variable, x_1 and x_2 are the covariates. Each distinct value of the subject-effect (id) identifies a different subject and each distinct value of the within subject-effect ($timepoints$) defines a different outcome for the same subject.

PROC GENMOD statement invokes the procedure. DESCENDING option sorts the response variable from highest to lowest (useful with binary and multinomial outcomes).

CLASS statement lists all categorical variables and must be before the MODEL statement. PARAM=REF and REF='level' need to be specified to be able to use a custom reference level.

MODEL statement specifies the response variable (numeric or character) and covariates (also numeric or character). An intercept is added to the model by default. LINK states the link function and DIST states the probability distribution to be used for Y in the model. If those are omitted, the model defaults to normal distribution with identity link function. TYPE3 computes generalized score statistics for GEEs and WALD calculates Wald statistics for Type III contrasts (only computed when TYPE3 is specified as well) using empirical (robust) variance estimators.

REPEATED statement invokes the GEE method and specifies the covariance structure of repeated measurements in the model. SUBJECT can be a single effect, interaction or nested and must be listed in the CLASS statement. WITHINSUBJECT defines an effect that specifies an order of repeated measurements. If for example there are some missing values, then specifying WITHINSUBJECT effect, the repeated measurements are ordered in a correct way (missing values are not assumed to be the last measurements). TYPE is used to specify the working correlation matrix. CORRW displays the estimated working correlation matrix and CORRB displays the regression parameter estimates' correlation matrix (empirical and model-based matrices are presented).

To address the missing completely at random assumption of GEE-type models, the GENMOD procedure can estimate the working correlation by using all available pairs method, and the obtained covariances and standard errors are valid under the MCAR assumption. (SAS User guide). There is an option to use weighted GEE method available in PROC GEE (in SAS version 13) which provides consistent estimates when data are missing at random (MAR) (SAS Institute Inc. 2014).

3.2 PROC GLIMMIX

This procedure can also fit GEE-type models, but there is an additional option to incorporate random effects. An example of GEE approach in the GLIMMIX procedure:

```
PROC GLIMMIX DATA=data EMPIRICAL;  
CLASS x1 id x3;  
MODEL y=x1 x2 x1*x2 /LINK=identity DIST=normal SOLUTION;  
RANDOM timepoints /SUBJECT=id TYPE=ar(1) RESIDUAL;  
RANDOM intercept /SUBJECT=id(x3);  
RUN;
```

PROC GLIMMIX invokes the procedure. The EMPIRICAL option requests the sandwich estimator to be used in calculating covariances for fixed effects.

CLASS statement like in the previous example lists all the categorical variables and must be before the MODEL statement.

MODEL statement names the response variable and covariates. An intercept is added by default. Adding the option SOLUTION produces the scale parameter and parameter estimates for fixed effects.

RANDOM statement invokes the GEE when used with RESIDUAL option. The *timepoints* effect is specified when there's a need to order the columns of covariance structure by the time variable. SUBJECT option identifies the subjects in the model. All subjects are assumed to be independent from each other. Adding the subject effect nests all the other effects in the RANDOM statement within the subject effect. TYPE specifies the covariance structure in the same way as in the GENMOD procedure.

To incorporate random effects in to the model, one needs to specify it with another RANDOM statement, for example if a variable x_3 is a random effect, specifying

```
RANDOM intercept /SUBJECT=id(x3);
```

will treat x_3 as a random effect and assumes subjects are nested within the random effect. Note that with the EMPIRICAL option, adding any random effects requires that data are grouped by subjects.

For the differences between PROC GENMOD and PROC GLIMMIX, refer to appendix 7.2.

4. Analysis of the IndiMed trial

This section will give an overview of IndiMed study design, some descriptive statistics of the data and results of analysis using GEE method of estimation.

4.1 IndiMed study design

The following section is a summary from an article (Meren et al. 2015).

Cardiovascular diseases (CVD) are the number one cause of death in Estonia. One of the risk factors of cardiovascular diseases is an increased arterial blood pressure which is most often caused by hypertension (HT). The incidence rate of HT has been 1400-1500 per 100 000 persons in Estonia from 2010-2013. Incidence of the disease is significantly higher after turning 35 years old and highest in 45-64-year old men.

HT is treatable by medication when good adherence can be achieved. Unfortunately, 25%-60% of HT patients will stop taking medication within the first 6 months of treatment. Low treatment adherence may be caused by disease's long asymptomatic period, complicated treatment scheme, ignorance of the disease, scepticism of treatment efficacy and high price of medications.

The heritability of cardiovascular disease mortality is estimated to be 38-57%, being different for males and females. Genetic markers tied to risks of CVD have been identified and different risk scores have been calculated using those markers. Results of analysis of the data from the Estonian Genome Center of University of Tartu have been shown that those risk scores are usable in Estonian population. For example, male gene donors in high risk score group have been shown to have 1.9 (95% CL 1.5-2.5) times higher risk for ischemic heart disease than in low risk score group.

It is not known how information about personal genetic risks affect HT patients' behaviour. The primary objective of IndiMed study is to estimate the effect of personal genetic risk information to patients' medication adherence and treatment efficiency.

The study is based on comparison of two groups – intervention and control. Control group consists of patients who get HT treatment and counselling in the standard way, according to the national guidelines. The intervention group patients get the same treatment as

control group, but additionally they receive information about their genetic risks of four diseases (type 2 diabetes, coronary artery disease, stroke, atrial fibrillation) according to the genetic risk scores. The information is communicated to them by their physician. After treatment start, data on treatment adherence, effectiveness of treatment, body weight and smoking habits are collected from both groups. The difference between systolic and diastolic blood pressure in two groups will be used to draw conclusions about effectiveness of receiving information on genetic risks.

This study is cluster-randomized and there was no blinding of patient and physician. The clusters are physicians who were randomized to intervention and control. Informing patients about their personal risk scores was thought to affect physician's behaviour towards the patients who did not receive risk score information and therefore decrease the differences of the study groups. For this reason, all of the patients in a physician's care are in the same study group (intervention or control).

The subjects are 18 to 65-year-old males with 1st or 2nd grade hypertension who did not get medications for HT in the past two months (before entering the study). Duration of patient's participation in the study is 12 months which includes 5 visits. See appendix 7.3 for the study design figure, including the timeline of the visits and what was measured.

Patient's blood sample was taken during the first visit. From the blood sample, DNA was separated and genotyped in EGCUT lab using real-time PCR method. Using known genetic markers (according to published large-scale Genome-Wide Association Studies, GWAS), personal risk scores were calculated for the following diseases: type 2 diabetes, coronary artery disease, stroke and atrial fibrillation (arrhythmia). For each disease state a patient was classified into one of the two categories: high genetic risk group (highest 10% of risk scores) or an average genetic risk group. The choice of these diseases was motivated by the fact that patients with HT are at high risk for these conditions and also by the existing knowledge on genetic markers affecting the risk.

During the next visits, treatment adherence was evaluated using Morisky Medication Adherence Scale (Morisky et al. 1986) and Brief Medication Questionnaire (Svarstad et al. 1999) allowed to describe the problems encountered while taking medication. Change in arterial systolic blood pressure during the study measured treatment effectiveness. Patient's health behaviour change was evaluated from body weight and smoking habits

(number of cigarettes per day). The difference in study groups come from visits 2 and 4 where the physician gave information about personal risk scores (intervention group received risk information on 2nd visit and control on 4th visit). The purpose of last (5th) visit was to evaluate long-term changes in patient's risk behaviours.

4.2 Descriptive statistics

There were 240 patients of whom 127 were in intervention group and 113 were in control group (Table 4.2-1). There were in total 34 doctors of whom 76% had 6 or fewer patients.

Table 4.2-1 *Number of patients in intervention and control group.*

Group	N	%
Intervention	127	52.9
Control	113	47.1
Total	240	100.0

Most of the patients (45.8%) belong to the high risk group for one of the four diseases (Table 4.2-2, note that per cents are based on total *N*). On the other hand, nobody had the high risk for all four diseases.

Table 4.2-2 *Number of conditions with high genetic risk among subjects by study group.*

Risk	No high risks		1 high risk		2 high risks		3 high risks	
Group	N	%	N	%	N	%	N	%
Intervention	39	16.3	61	25.4	25	10.4	2	0.8
Control	42	17.5	49	20.4	17	7.1	5	2.1
Total	81	33.8	110	45.8	42	17.5	7	2.9

All subjects were between 18 to 65 years old at the time of the first visit with mean weight of 94.3 kilograms, mean height of 179.4 centimeters and mean Body Mass Index (BMI, calculated as $BMI = \frac{weight}{height^2/100}$) of 29.2 (Table 4.2-3 Table 4.2-3).

Table 4.2-3 *Descriptive statistics for subjects' age, weight, height and BMI.*

	Age	Weight (kg)	Height (cm)	BMI
N	240	240	239	239
Mean	41.5	94.3	179.4	29.2
Median	42	92	179	28
SD *	12.15	17.51	7.70	4.67
Min	18	56	161	20
Max	65	155	202	43

* Standard deviation calculated as $SD = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$, where N is the number of nonmissing values of x .

For blood pressure related studies, risk factors are typically age, BMI and smoking. It is known that regularly smoking increases blood pressure and for this reason smoking is considered as a risk factor. Most of the patients (61.25%) were non-smokers (Table 7.4-1 in the appendix). Blood pressure could be associated with education level, meaning that subjects with higher education level are following the treatment more carefully and therefore have lower blood pressure by the end of the study. Education is also added into the models in the analysis section. Frequencies of patients' completed education level by study group is displayed in Table 7.4-2 in the appendix.

The following figures (Figure 4.2-1 and Figure 4.2-2) present systolic and diastolic blood pressure box plots by each visit for intervention and control group. In Figure 4.2-1, mean systolic blood pressure in intervention group drops noticeably with three visits. Variability is quite high in both study groups. In Figure 4.2-2, variability and mean diastolic blood pressure in intervention group seem to be a bit smaller than in the control group.

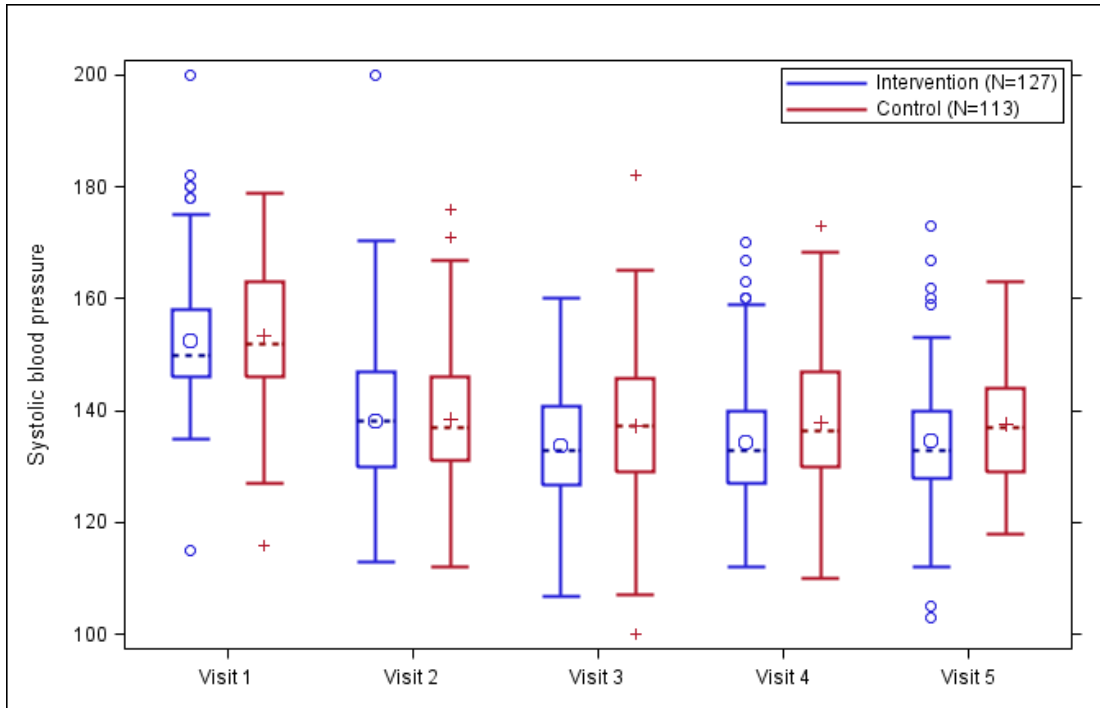


Figure 4.2-1 *Box plots of systolic blood pressure by study group and visit. Note that + and O inside the boxes display the means and outside the boxes display outliers.*

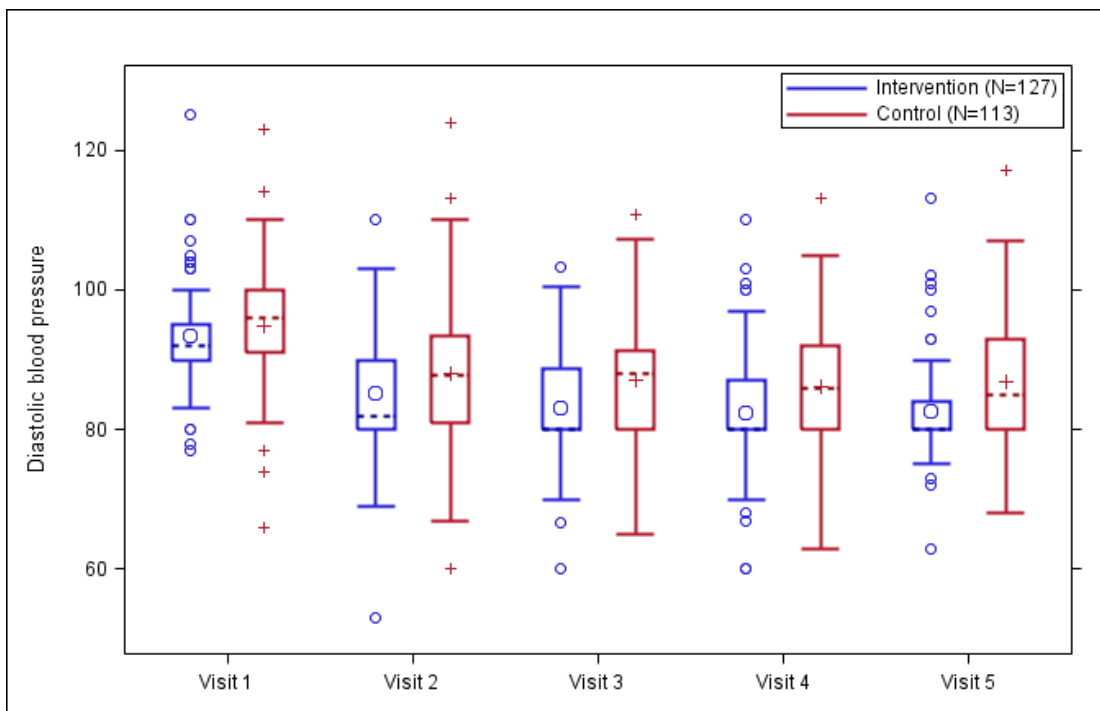


Figure 4.2-2 *Box plots of diastolic blood pressure by study group and visit. Note that + and O inside the boxes display the means and outside the boxes display outliers.*

Response variables, which are systolic and diastolic blood pressures, are assumed to be normally distributed in the population. However, IndiMed data are a sample of subjects with higher blood pressure than in the general population, therefore the distribution is skewed to the right as depicted below.

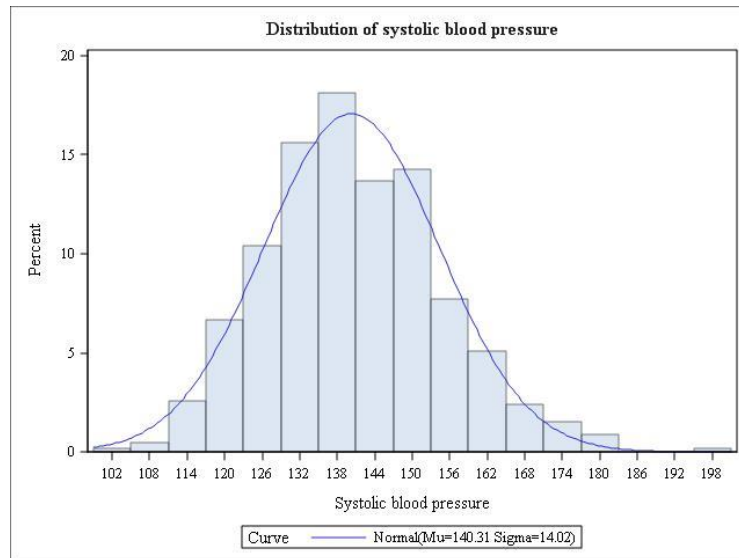


Figure 4.2-3 *Distribution of systolic blood pressure over all visits and subjects.*

Transforming the response variable may help to fix the skewedness. After log-transformation, systolic blood pressure is approximately normally distributed (Shapiro-Wilk statistic p -value 0.23) as shown in Figure 7.5-1 in the appendix.

For diastolic blood pressure (see Figure 7.5-2), neither log or square root transformations were successful according to the Shapiro-Wilk statistic. However, Box-Cox test, which helps to find a power transformation gave log-transformation as the best candidate. Therefore, we will also use log-transformed diastolic blood pressure in analysis (see Figure 7.5-3 for the shape of distribution).

4.3 Analysis

One of the aims of the analysis is to assess the effect of receiving genetic risk information on blood pressure. For this part, intervention and control groups are compared. The hypothesis is that for at least one blood pressure measure (systolic or diastolic), intervention and control groups are statistically different.

Secondly, is there a difference in blood pressure depending on whether the patients who got feedback on genetic risks actually had a high risk for one or more diseases? For this part, intervention group subjects are divided into two groups by the genetic risk: a) no diseases with high genetic risk, b) at least one disease with high genetic risk. The third group will be all patients in the control group. The hypothesis is that there is a difference in mean blood pressure between those three groups.

Since the patients are clustered by doctors, ideally the model should capture a random effect of the doctor, because we want to assume the doctors are randomly chosen from the whole population of doctors. To focus on GEE rather than GEE with additional random effects, the main analysis here will be done assuming a fixed doctor effect by using the GENMOD procedure in SAS. After that, assuming a random doctor effect, the GLIMMIX procedure which allows to add random effects together with GEE, is used.

Taking into account the characteristics of the data and how it is collected, AR-1 type of correlation structure between repeated measurements is assumed. Note that other (independent, exchangeable and Toeplitz) correlation structures were tested, but overall AR-1 gave lower QIC.

The GEE assumes missing values in data to be missing completely at random. There are many subjects with the last visit information missing at the time of the analysis (Table 7.4-3 in the appendix) and therefore only first four visits are used. The procedure GENMOD assumes MCAR data, but the data are more likely to be missing at random. Therefore, the estimates may be biased.

All models are fit for both systolic and diastolic blood pressures, where these variables are log-transformed beforehand to fix skewedness of the distribution.

Let's present an example of a log-linear (response is log-transformed and covariates are in the original scale) model's interpretation of parameter estimates.

Let the model for the blood pressure and estimated parameters be in the form of

$$\log(BP) = 4.49 - 0.03 \cdot \text{group}(\text{intervention}) - 0.08 \cdot \text{visit}(2) - 0.10 \cdot \text{visit}(3) - 0.11 \cdot \text{visit}(4)$$

Then the interpretation of mean blood pressure depends on the reference levels. Here the reference levels are $\text{group} = \text{control}$ and $\text{visit} = 1$. The (geometric) mean blood pressure at the reference levels is $e^{4.49} = 89.12$. Mean blood pressure at visit 1 for intervention group is $e^{(4.49-0.03)} = 86.5$. The means are different for intervention and control depending on the visits, but the ratio stays the same and is $e^{\text{effect}} = e^{(-0.03)} = 0.97$. It is convenient to interpret the differences between the two groups in terms of relative change. In this example, switching from control to intervention, the mean blood pressure is expected to decrease by about 3%.

Lastly, we have to consider multiple comparison since we are testing both systolic and diastolic blood pressures. Using Bonferroni correction, the significance level of parameter estimates is at $\alpha = 0.025$ ($0.05/2 = 0.025$). However, it can be argued that for systolic and diastolic blood pressures this correction is too strict because there is a strong correlation among the response variables ($\text{Cor}(\text{systolic}, \text{diastolic}) = 0.71$).

4.3.1 Intervention vs control group

The following three types of models (of form in section 2.3) for log-transformed systolic and diastolic blood pressures were fitted. Note that group variable is always included in the models, even if nonsignificant, because hypotheses are based on that variable.

1) $\log(Y) \sim \text{group}$

This model compares the overall mean blood pressure (across all visits) of the two study groups. For diastolic blood pressure (DBP), the parameter estimates are in the table below.

Table 4.3-1 Analysis of GEE parameter estimates, diastolic blood pressure, model $\log(Y) \sim \text{group}$.

Response	Parameter	Estimate	SE	Z	p-value
Diastolic	Intercept	4.4890	0.0072	625.89	<.0001
	Group (Intervention)	-0.0297	0.0094	-3.18	0.0015

According to the analysis, there is a significant difference in control and intervention group ($p = 0.0020$, see in Table 7.6-3). The geometric mean of the diastolic blood pressure in the control group is $e^{4.4890} = 89.03$ and for the intervention group it is $e^{4.4890-0.0297} = e^{4.4593} = 86.43$ (over all 4 visits). If the patient receives genetic risk information, we see about 2.9% lower geometric mean of the diastolic blood pressure.

For the systolic blood pressure (see Table 7.6-1 and Table 7.6-2), the group variable is nonsignificant at $\alpha = 0.025$ level ($\hat{\beta} = -0.0136$, $p = 0.0878$).

2) $\log(Y) \sim \text{group} + \text{visit} + \text{doctor}$

This model assumes the mean blood pressure depends on the groups, visits and doctors.

In the model with diastolic blood pressure as response (Table 7.6-6 and Table 7.6-7), the group variable ($p = 0.2694$) is nonsignificant. The effect of visit as well as that of the doctor are significant. Compared to visit 1, we see about 8.1% decrease in the geometric mean of diastolic blood pressure in visit 2 and about 10.5% decrease in blood pressure in visit 4.

For systolic blood pressure (Table 7.6-4 and Table 7.6-5), the doctor's effect ($p = 0.5121$) and group's effect ($p = 0.0559$) were nonsignificant.

3) $\log(Y) \sim \text{group} + \text{visit} + \text{doctor} + \text{BMI} + \text{age} + \text{smoking} + \text{education}$

This model is similar to the above one, but also including some additional risk factors. These are added to account for the associations which may distort the real relationship between response variable and covariates. When modelling blood pressure, often the risk

factors are BMI and age because higher BMI and older age refer to higher risk of HT. In addition, risk of HT can increase due to smoking. It is possible that patients at higher education level are following the treatment better than patients at lower education level and therefore have lower blood pressure by last visits. For this reason, education level is also included as a covariate.

This model will be fit first by treating doctor as a fixed effect and then treating it as a random effect. Note that the other risk factors are not removed from the model, even if nonsignificant, because we want to control for them.

In the model for diastolic blood pressure (Table 7.6-10 and Table 7.6-11), group is again nonsignificant ($p = 0.2045$), therefore there is no evidence for mean blood pressure difference between intervention and control groups. Education is also nonsignificant. BMI, smoking and age all have positive parameter estimates, meaning that high BMI, older age and smoking increase mean blood pressure. For visits the effects are very similar to the previous model 2).

In the random effect model for diastolic blood pressure (Table 7.6-14 and Table 7.6-15), education is nonsignificant as in the above model. However, there is a significant group effect ($p = 0.0024$). Parameter estimates for visit are very similar to the fixed effect model.

In systolic blood pressure model (Table 7.6-8 and Table 7.6-9), there is no difference in mean blood pressure between intervention and control. Only visit and BMI are significant with p -values of < 0.0001 and 0.0004 respectively. Visit effects are similar to the above model 2) and higher BMI has a positive parameter estimate which means that higher BMI refers to higher mean blood pressure.

In random effect model for systolic blood pressure (Table 7.6-12 and Table 7.6-13, group is nonsignificant as it is in the fixed effects model. In addition to visit and BMI, education and age are significant variables.

4.3.2 Comparison of intervention group individuals with high genetic risk to individuals at low risk and to the control group

In the next section, mean blood pressure is assessed in subgroups. The following model is fitted:

$$\log(Y) \sim \text{subgroup}.$$

The subgroup variable used in the model has a value:

- 0, if patient is in control group,
- 1, if patient is in intervention group and has no high genetic risks reported,
- 2, if patient is in intervention group and has high genetic risk for at least one disease reported.

Table 4.3-2 *Subgroup comparison for diastolic blood pressure. Model*

$$\log(Y) \sim \text{subgroup}.$$

Response	Parameter	Estimate	SE	Z	p-value
Diastolic	Intercept	4.4890	0.0072	625.81	<.0001
	Subgroup (1)	-0.0471	0.0117	-4.02	<.0001
	Subgroup (2)	-0.0217	0.0104	-2.08	0.0371

For diastolic blood pressure (Table 4.3-2 above), subgroup effect is significant ($p = 0.0010$, see Table 7.6-17 in the appendix). Compared to control, intervention subgroup 2 does not have statistically different mean diastolic blood pressure. We see about 4.7% decrease in mean blood pressure if the patient receives information on having no high genetic risks for any of the diseases. For subgroup 2, the decrease is about 2.2%. This model has a higher QIC than in model 1) in section 4.3.1 (QICs are 910.8 and 909.1 respectively).

Table 4.3-3 *Subgroup comparison for systolic blood pressure. Model*
 $\log(Y) \sim \text{subgroup}$.

Response	Parameter	Estimate	SE	Z	p-value
Systolic	Intercept	4.9561	0.0061	815.90	<.0001
	Subgroup (1)	-0.0333	0.0102	-3.28	0.0010
	Subgroup (2)	-0.0044	0.0087	-0.51	0.6109

For systolic blood pressure (Table 4.3-3 above and Table 7.6-16 in the appendix), subgroup effect is significant ($p = 0.0079$). The lowest mean systolic blood pressure is (as in the same model with diastolic blood pressure) for patients who received genetic risk information and had no high risks. Compared to control group patients, their mean blood pressure is decreased by about 3.3%. The difference between mean blood pressure between control group and subjects in intervention group who had at least one high disease risk, is nonsignificant. This model also does not have a better fit with the data than model 1) in section 4.3.1 (QIC 908.7 and 907.6 respectively), so it seems that whether or not one receives genetic risk information counts more than what their genetic risks are.

In conclusion for this section, the model which estimates the mean blood pressure for intervention and control group is a better fit to the data than the model where the number of high genetic risks in the intervention group is also taken into account.

4.4 Discussion

The mean blood pressure for intervention and control was not statistically different in most of the fitted models, so it implies that overall, it does not matter whether a patient knows or does not know if they have some high risks for the four diseases (type 2 diabetes, coronary artery disease, stroke and atrial fibrillation). It might be that at first, patients who know their personal risks, are worried and will take medication carefully, but after a while will accept it (or simply do not care) and the effect will fade in time. However, for diastolic blood pressure the effect is significant for model 1 in section 4.3.1. It is intriguing that systolic and diastolic blood pressures seem to respond differently.

Subgroup analysis results indicate that control group (all subjects) and intervention group with at least one high disease risk do not have statistically different mean blood pressures. However, the effect between control and intervention (patients with no high risks) is statistically significant. This is surprising and the reasoning behind that needs further research. It might be that patients who know they have many high disease risks are trying to change their lifestyle and carefully follow medication, but have higher blood pressure for some other reason (for example they are less responsive to treatment or have genetically higher blood pressure). It would also be interesting to know whether mean blood pressures are different for a specific disease risk.

In the near future, when all 5 visits have happened for all patients (at the time of the analysis not all patients had had their last visit), treatment adherence should also be analysed alongside with when and why the subject discontinued the study.

Finally, one should notice that the sample size for this pilot study is rather small. The initial number of patients desired was 300, but only 240 patients joined the study. It is possible that the intervention had an effect, but it was too small to be discovered in this trial. Also the point estimates of the parameters in this study suggest a possibility of a beneficial effect of the intervention. Larger study is necessary to confirm the usefulness of genetic feedback.

5. Conclusion

This thesis gave an overview of Generalized Estimating Equations, which is a method of estimation that accounts for correlations among repeated measurements and is widely used in longitudinal analysis. It has many appealing properties, mainly the robustness to misspecification of correlation structure between the repeated measurements. The validity of GEE parameter estimates and standard errors rely only on specifying the correct model for mean response.

The general guidelines of how to implement GEE-type models was presented in section 3 using SAS procedures PROC GENMOD and PROC GLIMMIX.

The final section implemented GEE-type models on a cluster-randomized study called IndiMed, using data from the Estonian Genome Center of University of Tartu. The clusters are doctors who were randomized to intervention or control group and subjects are men with high blood pressure. All subjects in one cluster are either in intervention (received genetic risk information on the 2nd visit) or control group (received risk information on the 4th visit). Systolic and diastolic blood pressures were compared for the two study groups. Autoregressive AR-1 correlation structure was assumed for the repeated measurements of response variables.

The mean blood pressure was not statistically significant for the two study groups in most of the fitted models which indicates there is no difference if the patient received personal risk information or not. The mean blood pressure was statistically different for patients in control group and patients in intervention group who had no high risks. Control group and intervention group patients with at least one high risk did not have statistically different means for blood pressure. These findings are quite surprising and suggest further analysis should be done. Future research can provide answers to questions whether there are some important factors which determine the blood pressure regarding if patient knows or does not know their risks for diseases.

6. References

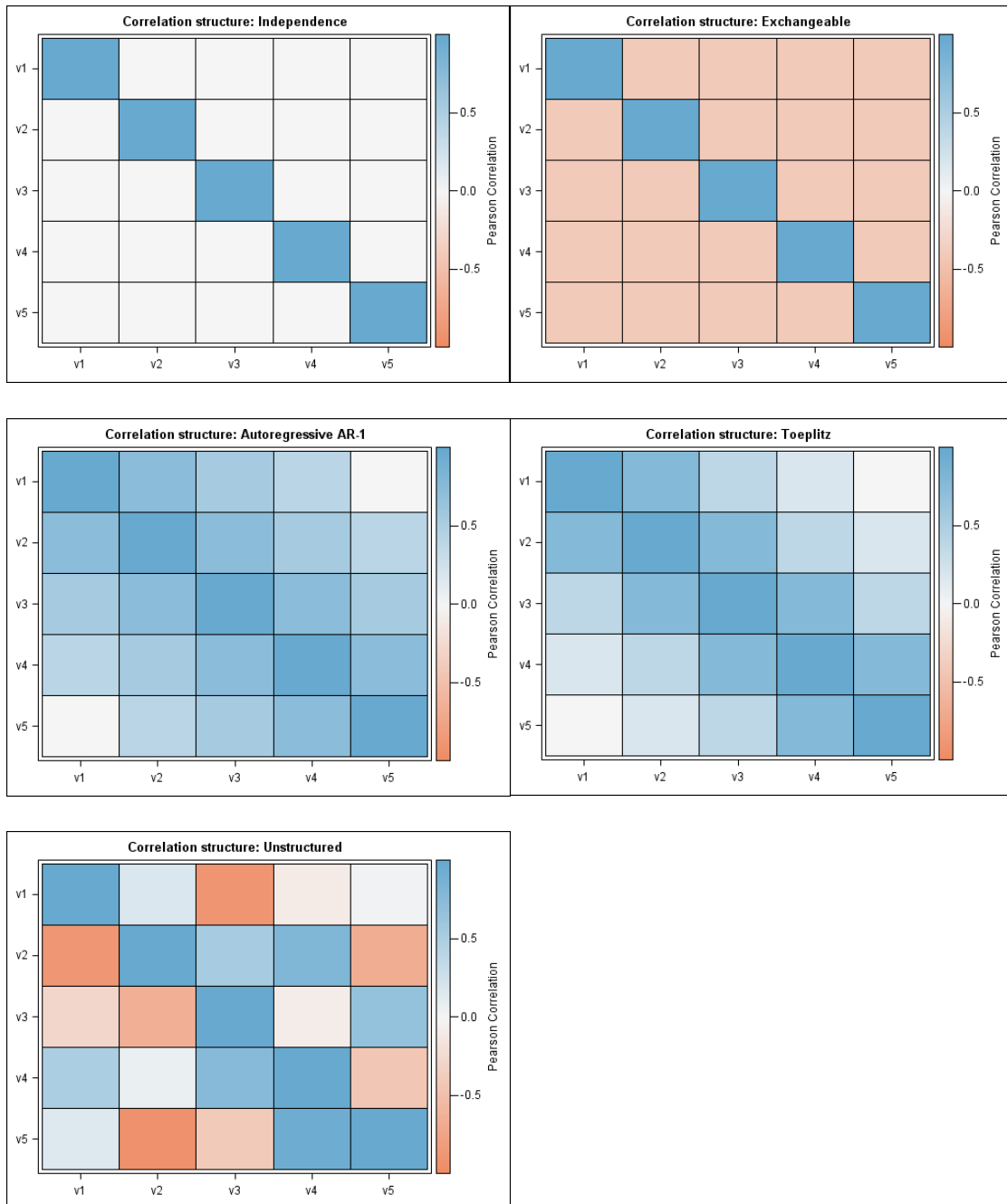
- Burton, P., Gurrin, L. & Sly, P., 1998. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in medicine*, 17(11), pp.1261–91. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/9670414> [Accessed April 9, 2016].
- Diggle, P.J. et al., 2002. *Analysis of Longitudinal Data* Second edi., New York: Oxford University Press.
- Fitzmaurice, G.M., Laird, N.M. & Ware, J.H., 2004. *Applied Longitudinal Analysis*, Hoboken, New Jersey: John Wiley & Sons.
- Jang, M.J., 2011. *Working correlation selection in generalized estimating equations*. University of Iowa. Available at: <http://ir.uiowa.edu/etd/2719> [Accessed April 9, 2016].
- Laird, N.M., Missing data in longitudinal studies. *Statistics in medicine*, 7(1-2), pp.305–15. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/3353609> [Accessed May 9, 2016].
- Liang, K.-Y. & Zeger, S.L., 1986. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73(1), pp.13–22. Available at: <http://links.jstor.org/sici?sici=0006-3444%28198604%2973%3A1%3C13%3ALDAUGL%3E2.0.CO%3B2-D> [Accessed April 9, 2016].
- Meren, Ü.H. et al., 2015. The Effect of the Risk Related Personal Genetic Data on Adherence to and Effectiveness of Hypertension Treatment on Estonian Patients: an Overview of the Methodology of the Randomised Study. *Eesti Arst*, (94), pp.118–122.
- Morisky, D.E., Green, L.W. & Levine, D.M., 1986. Concurrent and predictive validity of a self-reported measure of medication adherence. *Medical care*, 24(1), pp.67–74. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/3945130> [Accessed May 4, 2016].

- Pan, W., 2001. Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics*, 57(1), pp.120–125. Available at: <http://links.jstor.org/sici?sici=0006-341X%28200103%2957%3A1%3C120%3AAICIGE%3E2.0.CO%3B2-Q> [Accessed April 29, 2016].
- Rubin, D.B., 1976. Inference and Missing Data. *Biometrika*, 63(3), pp.581–592.
- SAS Institute Inc., 2014. *SAS/STAT® 13.2 User's Guide*, Cary, NC: SAS Institute Inc. Available at: <http://support.sas.com/documentation/cdl/en/statug/67523/PDF/default/statug.pdf>.
- SAS Institute Inc., 2011. *SAS/STAT® 9.3 User's Guide*, Cary, NC: SAS Institute Inc. Available at: <https://support.sas.com/documentation/cdl/en/statug/63962/PDF/default/statug.pdf>.
- Svarstad, B.L. et al., 1999. The Brief Medication Questionnaire: a tool for screening patient adherence and barriers to adherence. *Patient education and counseling*, 37(2), pp.113–24. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/14528539> [Accessed May 4, 2016].
- Wang, Y.-G. & Carey, V., 2003. Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. *Biometrika*, 90(1), pp.29–41. Available at: <http://biomet.oupjournals.org/cgi/doi/10.1093/biomet/90.1.29> [Accessed May 9, 2016].
- Zeger, S.L. & Liang, K.-Y., 1986. Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, 42(1), p.121. Available at: <http://www.jstor.org/stable/2531248?origin=crossref> [Accessed April 9, 2016].

7. Appendix

7.1 Correlation structures

Figure 7.1-1 *Examples of Independence, exchangeable, autoregressive(1), Toeplitz and unstructured correlation structures.*



7.2 PROC GLIMMIX contrasted with PROC GENMOD

The following is from SAS user guide (SAS Institute Inc. 2011).

SAS procedures GENMOD and GLIMMIX can both fit GEE-type models. However, GENMOD does not allow to add random effects with GEE models (with REPEATED statement).

The GEE estimation in GENMOD procedure uses only R-side covariances whereas GLIMMIX has R-side covariances and also G-side random effects. There is a difference in estimating the covariance parameters: GENMOD estimates these by the method of moments and GLIMMIX by likelihood-based methods.

7.3 IndiMed study design and timeline

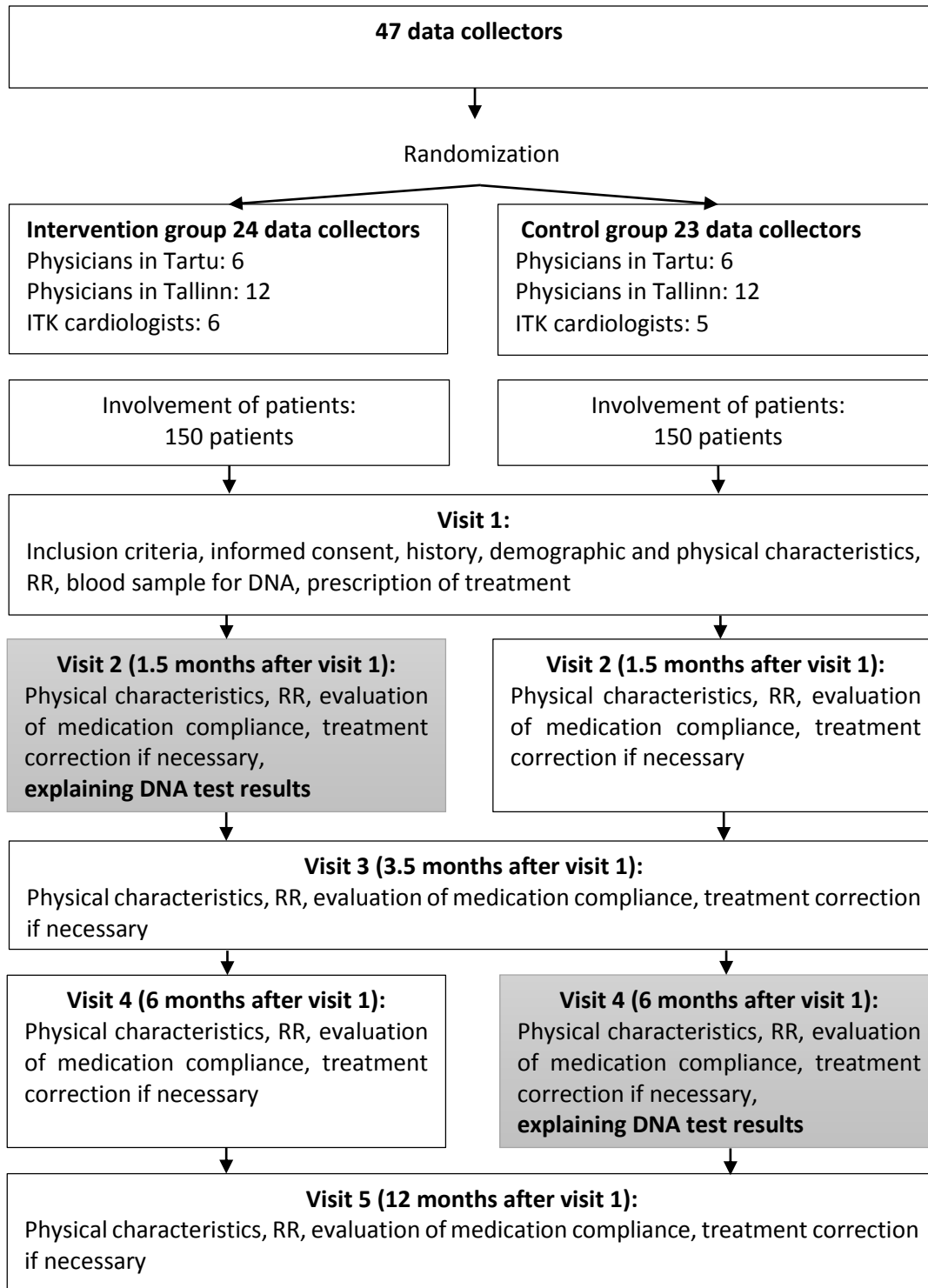


Figure 7.3-1 Study design, explanations are in text.

7.4 Tables of descriptive statistics

Table 7.4-1 *Distribution of patients' smoking by study group.*

Smoking	N	%
Yes	93	38.75
No	147	61.25
Total	240	100.0

Table 7.4-2 *Subjects' completed level of education by study group.*

Level of education	Primary school	Middle school	High or vocational school	Bachelor's degree	Scientific degree MSc/PhD	Missing
Group	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
Intervention	2 (0.8)	17 (7.1)	72 (30.0)	29 (12.1)	5 (2.1)	2 (0.8)
Control	0 (0.0)	19 (7.9)	62 (25.8)	26 (10.8)	5 (2.1)	1 (0.5)
Total	2 (0.8)	36 (15.0)	134 (55.8)	55 (22.9)	10 (4.2)	3 (1.3)

Table 7.4-3 *Missing values for blood pressure by visit and study group.*

Visit	Visit 1		Visit 2		Visit 3		Visit 4		Visit 5	
Group	N (%)	Mis* (%)	N (%)	Mis* (%)	N (%)	Mis* (%)	N (%)	Mis* (%)	N (%)	Mis* (%)
Intervention	127 (52.9)	0 (0.0)	125 (52.1)	2 (0.8)	114 (47.5)	13 (5.4)	106 (44.1)	21 (8.8)	79 (32.9)	48 (20.0)
Control	113 (47.1)	0 (0.0)	111 (46.3)	2 (0.8)	107 (44.6)	6 (2.5)	101 (42.1)	12 (5.0)	53 (22.1)	60 (25.0)
Total	240 (100.0)	0 (0.0)	236 (98.4)	4 (1.6)	221 (92.1)	19 (7.9)	207 (86.2)	33 (13.8)	132 (55.0)	108 (45.0)

* *Mis: Missing*

7.5 Distribution of systolic and diastolic blood pressures

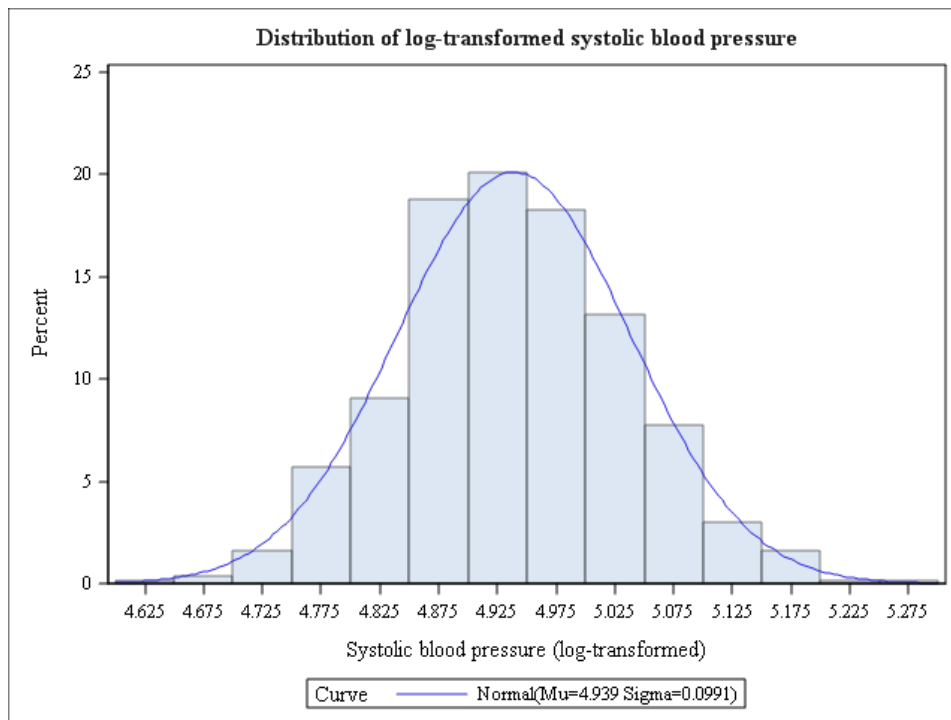


Figure 7.5-1 *Distribution of log-transformed systolic blood pressure over all visits and subjects.*

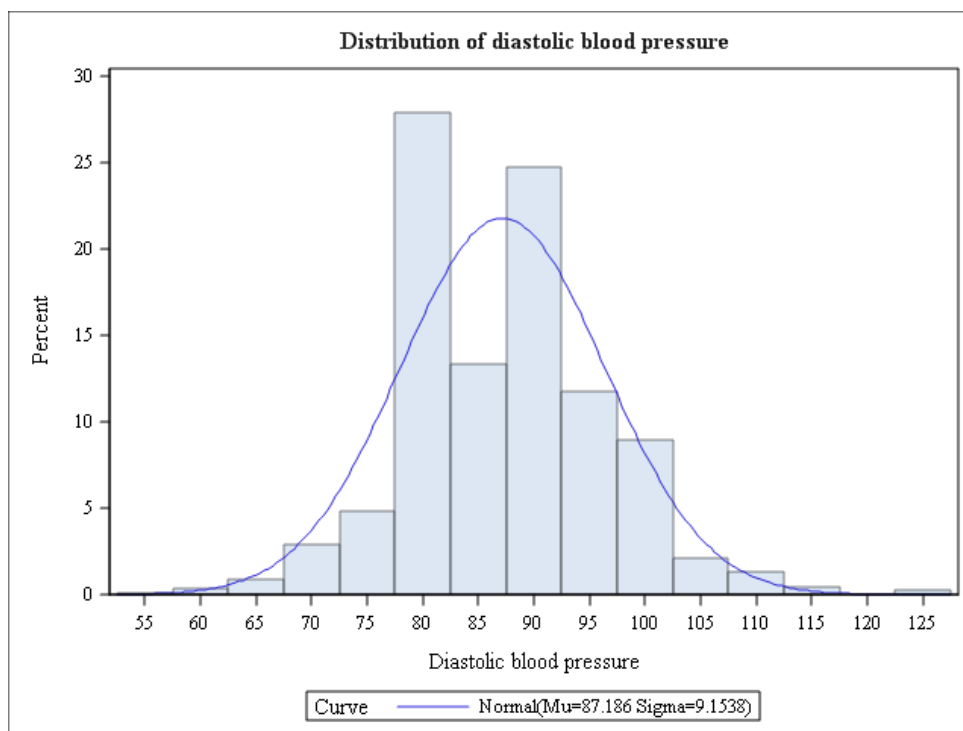


Figure 7.5-2 *Distribution of diastolic blood pressure over all visits and subjects.*

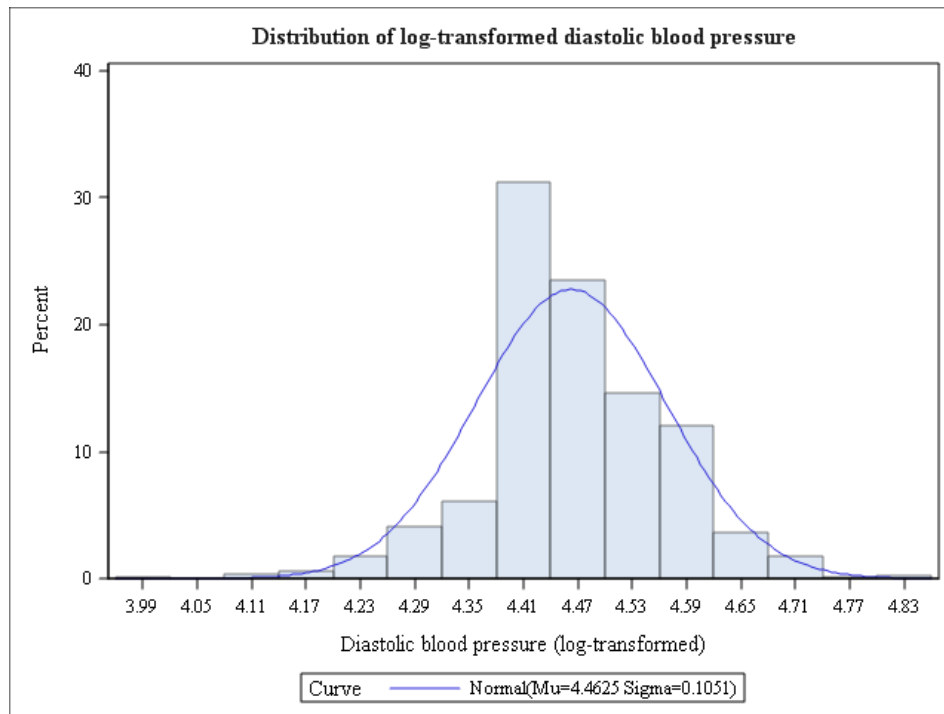


Figure 7.5-3 *Distribution of log-transformed diastolic blood pressure over all visits and subjects.*

7.6 Tables of analysis results

Model 1. $\log(Y) \sim \text{group}$

Table 7.6-1 Analysis of GEE parameter estimates, systolic blood pressure, model $\log(Y) \sim \text{group}$.

Response	Parameter	Estimate	SE	Z	p-value
Systolic	Intercept	4.9562	0.0061	816.17	<.0001
	Group (Intervention)	-0.0136	0.0079	-1.72	0.0859

Table 7.6-2 Score statistics for Type III GEE analysis, systolic blood pressure, model $\log(Y) \sim \text{group}$.

Response	Parameter	DF	ChiSq	p-value
Systolic	Group	1	2.91	0.0878

Table 7.6-3 Score statistics for Type III GEE analysis, diastolic blood pressure, model $\log(Y) \sim \text{group}$.

Response	Parameter	DF	ChiSq	p-value
Diastolic	Group	1	9.59	0.0020

Model 2. $\log(Y) \sim \text{group} + \text{visit} + \text{doctor}$

Table 7.6-4 Analysis of GEE parameter estimates, systolic blood pressure, model $\log(Y) \sim \text{group} + \text{visit} + \text{doctor}$. Note that rows for doctor are omitted.

Response	Parameter	Estimate	SE	Z	p-value
Systolic	Intercept	5.0336	0.0183	274.71	<.0001
	Group (Intervention)	0.0584	0.0106	5.29	<.0001
	Visit (2)	-0.1024	0.0066	-15.43	<.0001
	Visit (3)	-0.1228	0.0068	-17.97	<.0001
	Visit (4)	-0.1201	0.0070	-17.15	<.0001

Table 7.6-5 Score statistics for Type III GEE analysis, systolic blood pressure, model $\log(Y) \sim \text{group} + \text{visit} + \text{doctor}$.

Response	Parameter	DF	ChiSq	p-value
Systolic	Group	1	3.65	0.0559
	Visit	3	150.64	<.0001
	Doctor	33	32.09	0.5121

Table 7.6-6 Analysis of GEE parameter estimates, diastolic blood pressure, model $\log(Y) \sim \text{group} + \text{visit} + \text{doctor}$. Note that rows for doctor are omitted.

Response	Parameter	Estimate	SE	Z	p-value
Diastolic	Intercept	4.5724	0.0190	241.22	<.0001
	Group (Intervention)	-0.0397	0.0315	-1.26	0.2068
	Visit (2)	-0.0842	0.0062	-13.54	<.0001
	Visit (3)	-0.1006	0.0069	-14.52	<.0001
	Visit (4)	-0.1112	0.0071	-15.73	<.0001

Table 7.6-7 Score statistics for Type III GEE analysis, diastolic blood pressure, model $\log(Y) \sim \text{group} + \text{visit} + \text{doctor}$.

Response	Parameter	DF	ChiSq	p-value
Diastolic	Group	1	1.22	0.2694
	Visit	3	139.47	<.0001
	Doctor	33	66.98	0.0004

Model 3. $\log(Y) \sim \text{group} + \text{visit} + \text{doctor} + \text{BMI} + \text{age} + \text{smoking} + \text{education}$

Table 7.6-8 Analysis of GEE parameter estimates, systolic blood pressure, model $\log(Y) \sim \text{group} + \text{visit} + \text{doctor} + \text{BMI} + \text{age} + \text{smoking} + \text{education}$. Note that rows for doctor are omitted.

Response	Parameter	Estimate	SE	Z	p-value
Systolic	Intercept	4.8569	0.0370	131.15	<.0001
	Group (Intervention)	0.0129	0.0248	0.52	0.6020
	Visit (2)	-0.1018	0.0067	-15.30	<.0001
	Visit (3)	-0.1211	0.0068	-17.88	<.0001
	Visit (4)	-0.1196	0.0071	-16.83	<.0001
	BMI	0.0035	0.0009	3.84	0.0001
	Smoking (yes)	0.0101	0.0084	1.21	0.2269
	Education (middle)	0.0163	0.0151	1.08	0.2801
	Education (high)	-0.0028	0.0096	-0.29	0.7724
	Education (BSc)	-0.0116	0.0128	-0.90	0.3675
	Education (MSc/PhD)	-0.0524	0.0205	-2.56	0.0106
	Age	0.0008	0.0004	1.96	0.0495

Table 7.6-9 Score statistics for Type III GEE analysis, systolic blood pressure, model $\log(Y) \sim \text{group} + \text{visit} + \text{doctor} + \text{BMI} + \text{age} + \text{smoking} + \text{education}$.

Response	Parameter	DF	ChiSq	p-value
Systolic	Group	1	0.23	0.6304
	Visit	3	147.53	<.0001
	BMI	1	12.59	0.0004
	Smoking	1	1.45	0.2291
	Doctor	33	26.33	0.7881
	Education	4	7.80	0.0991
	Age	1	3.78	0.0517

Table 7.6-10 Analysis of GEE parameter estimates, diastolic blood pressure, model $\log(Y) \sim \text{group} + \text{visit} + \text{doctor} + \text{BMI} + \text{age} + \text{smoking} + \text{education}$. Note that rows for doctor are omitted.

Response	Parameter	Estimate	SE	Z	p-value
Diastolic	Intercept	4.4737	0.0526	84.98	<.0001
	Group (Intervention)	-0.0608	0.0435	-1.40	0.1625
	Visit (2)	-0.0845	0.0063	-13.45	<.0001
	Visit (3)	-0.1004	0.0070	-14.27	<.0001
	Visit (4)	-0.1109	0.0072	-15.46	<.0001
	BMI	0.0032	0.0011	2.86	0.0042
	Smoking (yes)	0.0196	0.0084	2.33	0.0196
	Education (middle)	-0.0299	0.0333	-0.90	0.3696
	Education (high)	-0.0234	0.0320	-0.73	0.4641
	Education (BSc)	-0.0309	0.0327	-0.94	0.3457
	Education (MSc/PhD)	-0.0313	0.0374	-0.84	0.4031
	Age	0.0008	0.0004	2.34	0.0194

Table 7.6-11 Score statistics for Type III GEE analysis, diastolic blood pressure, model $\log(Y) \sim \text{group} + \text{visit} + \text{doctor} + \text{BMI} + \text{age} + \text{smoking} + \text{education}$.

Response	Parameter	DF	ChiSq	p-value
Diastolic	Group	1	1.61	0.2045
	Visit	3	136.02	<.0001
	Doctor	33	55.61	0.0082
	BMI	1	7.15	0.0075
	Smoking	1	5.24	0.0220
	Education	4	1.45	0.8355
	Age	1	5.14	0.0234

Table 7.6-12 *Systolic blood pressure model with doctor as a random effect. Solution for fixed effects.*

Response	Effect	Estimate	SE	DF	t	p-value
Systolic	Intercept	4.8754	0.0271	228	180.06	<.0001
	Group (Intervention)	-0.0100	0.0073	648	-1.37	0.1712
	Visit (2)	-0.1020	0.0066	648	-15.35	<.0001
	Visit (3)	-0.1214	0.0068	648	-17.95	<.0001
	Visit (4)	-0.1196	0.0071	648	-16.86	<.0001
	BMI	0.0038	0.0008	648	4.83	<.0001
	Age	0.0008	0.0004	648	2.28	0.0227
	Education (middle)	0.0263	0.0124	648	2.12	0.0345
	Education (high)	0.0064	0.0066	648	0.96	0.3360
	Education (BSc)	0.0001	0.0089	648	0.01	0.9908
	Education (MSc/PhD)	-0.0288	0.0169	648	-1.70	0.0897
	Smoking (yes)	0.0143	0.0078	648	1.84	0.0670

Table 7.6-13 *Systolic blood pressure model with doctor as a random effect. Type III tests of fixed effects.*

Response	Parameter	DF	F	p-value
Systolic	Group	1	1.88	0.1712
	Visit	3	135.01	<.0001
	BMI	1	23.38	<.0001
	Age	1	5.22	0.0227
	Education	4	2.28	0.0595
	Smoking	1	3.37	0.0670

Table 7.6-14 *Diastolic blood pressure model with doctor as a random effect. Solution for fixed effects.*

Response	Effect	Estimate	SE	DF	t	p-value
Diastolic	Intercept	4.3889	0.0439	228	99.85	<.0001
	Group (Intervention)	-0.0264	0.0087	648	-3.05	0.0024
	Visit (2)	-0.0848	0.0063	648	-13.45	<.0001
	Visit (3)	-0.1008	0.0070	648	-14.31	<.0001
	Visit (4)	-0.1107	0.0072	648	-15.39	<.0001
	BMI	0.0044	0.0010	648	4.40	<.0001
	Age	0.0011	0.0004	648	3.05	0.0024
	Education (middle)	-0.0340	0.0313	648	-1.08	0.2784
	Education (high)	-0.0113	0.0293	648	-0.39	0.7003
	Education (BSc)	-0.0168	0.0303	648	-0.56	0.5781
	Education (MSc/PhD)	0.0087	0.0339	648	0.26	0.7967
	Smoking (yes)	0.0215	0.0091	648	2.38	0.0177

Table 7.6-15 *Diastolic blood pressure model with doctor as a random effect. Type III tests of fixed effects.*

Response	Effect	DF	F	p-value
Diastolic	Group	1	9.28	0.0024
	Visit	3	107.23	<.0001
	BMI	1	19.35	<.0001
	Age	1	9.31	0.0024
	Education	4	1.21	0.3060
	Smoking	1	5.66	0.0177

Model 4. $\log(Y) \sim \text{subgroup}$

Table 7.6-16 *Score statistics for Type III GEE analysis, systolic blood pressure, model $\log(Y) \sim \text{subgroup}$.*

Response	Parameter	DF	ChiSq	p-value
Systolic	Subgroup	2	9.69	0.0079

Table 7.6-17 *Score statistics for Type III GEE analysis, diastolic blood pressure, model $\log(Y) \sim \text{subgroup}$.*

Response	Parameter	DF	ChiSq	p-value
Diastolic	Subgroup	2	13.80	0.0010

Non-exclusive licence to reproduce thesis and make thesis public

I, Maia Arge (birth date 22.06.1992),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:
 - 1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and
 - 1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Generalized Estimating Equations: an overview and application in IndiMed study,

supervised by Kristi Läll, Pasi Korhonen and Krista Fischer,

2. I am aware of the fact that the author retains these rights.
3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **12.05.2016**