Fall 2012

Linear Mixed Effects Model for a Longitudinal Genome Wide Association Study of Lipid Measures in Type 1 Diabetes

Tao Wang
McMaster University, Wangt7@math.mcmaster.ca

Linear Mixed Effects Model for a Longitudinal Genome Wide Association Study of Lipid Measures in Type 1 Diabetes

Tao Wang

Supervisor: Dr. Angelo J. Canty

Department of Mathematics & Statistics,

McMaster University, Hamilton, Ontario, Canada

September, 2012

Acknowledgements

It is with gratitude that I acknowledge the support and help of all people in last two years in McMaster University. I would like to express the deepest appreciation to my supervisor Dr. Angelo Canty who was abundant helpful and offered invaluable assistance, continuous support, guidance and encouragement. Without his persistent help this thesis would not have been possible.

I would like to thank my thesis committee members, Dr. Jemila Hamid and Dr. Joseph Beyene for agreeing to be my thesis examiners.

I consider it an honor to work with Dr. Andrew Paterson, Dr. Shelley Bull, Dr. Andrew Boright and Dr. Lei Sun in DCCT/EDIC Research Group. Their wealthy comments and valuable feedback were essential to the completion of this thesis.

I am also much appreciated with my family, my father, my husband and my daughter. I hope this thesis could be a gift to my mum who had gone in 2010. They always encourage me and make me feel the spectacular success achievable. I would not have been able to accomplish my studies without them.

Contents

List of Figures

List of Tables

Abstract

Hypercholesterolemia is the presence of high levels of cholesterol in the blood, and it is one of the major factors for the development of long-term complications in T1D patients.

In the thesis, we studied 1303 Caucasians with type 1 diabetes in the Diabetes Control and Complications Trial (DCCT). With the experience of diabetes study, many factors are associated with diabetes complications, they are age, gender, cohort, treatment, diabetes duration, body mass index (BMI), exercise, insulin dose, etc. We mainly focus on which factors are associated with total cholesterol (CHL) analysis in the thesis.

Many measures were collected monthly, quarterly or yearly for average 6.5 years from 1983 to 1993. We used annually lipid measures of DCCT because of their values are sufficient and complete, and they belong to longitudinal data.

Different methods are discussed in the study, and linear mixed effect models are the appropriate approach to the study. The details of model selection with CHL model analysis are shown, which includes fixed effect selection, random effects selection, and residual correlation structure selection. Then the SNPs were added on three models individually in GWAS.

We found locus (rs7412) is not only genome-wide associated with CHL, but also genome-wide associated with LDL.

We will assess whether these SNPs are diabetes-specific in the future, and we will add dietary data in the three models to identify locus are associated with the interaction of diet and SNPs.

Abbreviations Used in the Thesis

AIC——Akaike Information Criterion

BIC——Beyesian Information Criterion

BMI——Body Mass Index

BP——Base Position on the Chromosome

CHL——Total Cholesterol

DCCT ——Diabetes Control and Complications Trial

DNA——Deoxyribonucleic Acid

GWAS ——Genome Wide Association Study

HbA1c——Hemoglobin A1c

HDL——High Density Lipoprotein

LD——Linkage Disequilibrium

LDL——Low Density Lipoprotein

LME—— Linear Mixed Effects

MAF——Minor Allele Frequency

ML——Maximum Likelihood

REML——Restricted (Residual) Maximum Likelihood

RNA——Ribonucleic Acid

SNP——Single Nucleotide Polymorphism

T1D——Type 1 Diabetes

T2D——Type 2 Diabetes

TRG——Triglycerides

VIF——Variance Inflation Factor

Chapter 1

Introduction

1.1 Introduction to the problem

Today, more than 250 million people worldwide are living with diabetes, and approximately seven million people around the world are diagnosed with diabetes every year. In Canada, the number of patients living with diabetes or prediabetes has reached 9 million, so approximately 30 percent of the total Canadian population (around 30 million) (Canada Diabetes Association report, 2012). Diabetes can lead to serious complications, e.g. heart, kidney, eye disease, nerve damage, etc.

There are three main types of diabetes: Type 1 diabetes (T1D), Type 2 diabetes (T2D) and Gestational diabetes. T1D is a disease in which the insulin-producing cells of the pancreas are destroyed by the immune system of the human body, it is usually diagnosed in children and adolescents. Approximately 10 per cent of people with diabetes have T1D. T2D occurs when the cells of human body fail to use the produced insulin properly or the immune system

itself is insulin-resistant. T2D usually develops in adulthood, although increasing numbers of children are being diagnosed. Gestational diabetes is a kind of diabetes found in pregnant women who have never had diabetes before and experience high blood glucose level during pregnancy (JDRF report, 2011). We focus only on T1D in this work.

We focus on four lipid measures as our outcomes, because abnormal lipid levels are an important risk factors for heart disease and nephropathy which are major complications of diabetes. In this work we focus in particular on total cholesterol (CHL).

The Human Genome Projects finished in 2003, the International HapMap finished in 2005 and 1000 Genomes Project are still in progress. These projects have identified approximately 52 million Single Nucleotide Polymorphism (SNPs) on the human genome [NCBI, 2012]. Genome Wide Association Studies (GWAS) are useful in finding common genetic variants that contribute to common, complex diseases. As of 2011, GWAS have examined over 200 diseases and traits, almost 4,000 SNP associations have been found [Trompet & Jukema,2012]. We are doing quantitative trait GWAS to identify novel SNPs which are significantly associated with lipids in T1D.

In the project, we used the Diabetes Control and Complications Trial (DCCT) data set. There are total of 1441 patients followed for on average 6.3 years from 1983 to 1992, of which 1303 Caucasians were studied in the thesis. Many measurements were collected over time, quarterly or annually, for instance, body mass index (BMI) and hemoglobin A1c (HbA1c) were collected each three months and other measures such as CHL, high density lipoprotein (HDL), low density lipoprotein (LDL), triglycerides (TRG), insulin dose, C-peptide, exercise were all collected yearly. We used annual data in the study because our response variables and most

covariates were measured yearly.

Longitudinal data analyses are widely conducted in various applications as diverse as industry, agriculture, biology, economics, etc. There is a big difference with traditional cross-sectional analysis, because observations have correlation within-subject. So some commonly methods, such as ANOVA and MANOVA, cannot solve these problems [Mavidian, 2011]. We will use linear mixed-effects to analysis our longitudinal data set in this thesis.

1.2 Genetics background acknowledge

1.2.1 Gene, DNA and Protein

The genome is made up of 2 copies of 23 pairs of chromosomes of deoxyribonucleic acid (DNA), and each chromosome consists of two long polymers of nucleotides which bind together to form a double-stranded helix. Each nucleotide is one of 4 bases, they are Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). The two strands are complementary and usually bind according to Watson-Crick base-pairing in which A binds with T and G binds with C.

DNA carries the genetic information of a cell and consists of thousands of genes. A gene can be thought of as a section of DNA on a chromosome. Transcription is the process by which this DNA is turned into the related single-stranded ribonucleic acid (RNA). In the process of translation, groups of 3 consecutive RNA nucleotides join together to form an amino acid which join together in various ways to form a protein. In any cell only about 60 percent of the genes produce RNA for translation and production of proteins, and these are called the expressed genes for that cell. For those that are expressed, there can be different amounts of

RNA produced and so different protein levels. Higher gene expression usually results in higher abundance of the protein which has implications for biological functions, including disease [Lodish, 2000].

1.2.2 Single Nucleotide Polymorphism

A single-nucleotide polymorphism (SNP) is a single base pair variation at a specific locus which consist of two alleles in a DNA sequence among individuals, groups, or populations. An allele is an alternative nucleotide located at a specific position on a specific chromosome. Minor allele frequency (MAF) refers to the frequency at which the less common allele occurs in a given population [Pagon R, 1993].

Although more than 99% of human DNA sequences are the same, variations in DNA sequence can have impact on human disease or response to a drug. Scientists believe many SNPs are associated with some human phenotype traits, such as cancer, diabetes, heart disease, etc. SNPs can be used to track the inheritance of disease genes within families, so people have been trying to identify SNPs associated with complex diseases for many years.

SNPs are the most common polymorphism and can occur in coding or non-coding regions of the genome. SNPs occur once in every 300 nucleotides on average.

1.2.3 GWAS

A Genome-wide association study (GWAS) is a study of genetic variation across the entire human genome designed to identify genetic association with observable traits or the presence of a disease. A GWAS uses high-throughout genotyping technologies to assay hundreds of

thousands of SNPs and relate them to clinical conditions and measurable traits [Burton et al., 2007].

By examining the image produced from the microarray, we can deduce the participants genotype at each SNP location. Normally, genotype is coded as 0, 1, or 2, which is a count of the number of copies of the rarer (minor) allele present and describes the genotype, such as T/T, T/C and C/C [Syvanen, 2005].

1.3 Organization of thesis

The thesis is organized as follows: Chapter 2 discusses some general methods for longitudinal data analysis, with emphasis on linear mixed-effect models. Chapter 3 shows the application of linear mixed effects model to the DCCT dataset. The details of analysis were listed for discovering which covariates are associated with CHL. Chapter 4 describes the GWAS for CHL in DCCT. In Chapter 4, some software and applications are introduced, such as PLINK, Haploview which are useful for GWAS analysis. Chapter 5 summaries our findings and discusses ongoing and future work.

Chapter 2

Longitudinal Data Analysis

Longitudinal studies have a lot of advantages compared to cross-sectional designs which study many observations at a given time. Longitudinal studies consider both the between-subject and within-subject time-related variations, and provide more efficient estimators than cross-sectional designs with the same number and patterns of observations [Little, 2000]. Longitudinal studies also can separate aging effects from baseline effects, while cross-sectional designs can not handle this. Furthermore, longitudinal studies provide more information about individual changes over time, since in real world it has more practical meaning for individuals [Davidian, 2011].

2.1 Challenges

There are also some challenges associated with longitudinal data analysis. Observations are made sequentially on the same individual, and are not independent. Therefore these observations must be considered as dependent observations in data analysis. Different correlation

patterns should be considered and compared using Akaike information criteria (AIC) [Sakamoto, 1986] or Bayesian information criteria (BIC) [Schwarz, 1978]. It is very difficult to analyze longitudinal data under some situations, such as, unbalanced designs, missing data, etc. Moreover, in longitudinal studies, we may need to consider some time-varying covariates.

2.2 General Approaches for longitudinal data

Normally a longitudinal data set can be shown as Table 2.1. The meaning of notations are explained below:

i indexes $1, 2, ..., N$ subjects;

j indexes $1, 2, ..., n_i$ observations (e.g. yearly data over time) where n_i is the total number of observations for subject i. Total numbers of observations $= \sum_{i=1}^{N} n_i$.

$\mathbf{y}_i$ indexes $n_i \times 1$ vector of responses for subject i.

$\mathbf{x_{ij}}$ indexes $p \times 1$ covariate vector of subject i at time j where p is the number of covariates. These covariates can be either time-independent (between-subject) or time-dependent (within-subject).

For longitudinal data, the potential sources of variation usually are of two main types:

- Between-subject variation
- within-subject variation

By separating between-subjects variation from within-subject variation, the response can be expressed as below (2.1),

$$\mathbf{y}_{ij} = \mu_{ij} + b_{ij} + e_{ij} \tag{2.1}$$

Table 2.1: *Longitudinal Data Layout*

subject	**observations**	**response**	**covariates**
1	1	y_{11}	x_{111} ... x_{11p}
1	2	y_{12}	x_{121} ... x_{12p}
-	-	-	-
1	n_1	y_{1n_1}	x_{1n_11} ... x_{1n_1p}
-	-	-	-
-	-	-	-
-	-	-	-
N	1	y_{N1}	x_{N11} ... x_{N1p}
-	-	-	-
N	n_N	y_{Nn_N}	x_{Nn_N1} ... x_{Nn_Np}

where we assume E(b_{ij})=0, E(e_{ij})=0. The meaning of notations used in the model is listed below

- b_{ij} is the deviation representing between subject variation at time t_j.
- e_{ij} represents the additional deviation due to within-subject fluctuations about the trend.

There are several commonly used methods for analyzing longitudinal data, which are repeated measures ANOVA, multivariate MANOVA, mixed-effects regression models, covariance pattern models, which are very popular for analyzing binary, count and categorical longitudinal data [Hedeker, 2006]. In the following sections, the methods of repeated measure ANOVA and Multivariate ANOVA will be described and compared to mixed-effects models.

Table 2.2: *Univariate Repeated Data Layout*

	Time point			
Subjects	1	2	...	n
1	y_{11}	y_{12}	...	y_{1n}
2	y_{21}	y_{22}	...	y_{2n}
-	-	-	-	
-	-	-	-	
N	y_{N1}	y_{N2}	...	y_{Nn}

2.2.1 Repeated Measures using ANOVA

First let us look at the univariate ANOVA repeated model, in which each individual observations are considered independent. A even simpler data set is used to explain the repeated measures model and listed in Table 2.2, where t_j is specified in the model as the time point.

$$\mathbf{y}_{ij} = \mu + b_i + t_j + e_{ij} \quad (i = 1, ..., N; j = 1, ..., m) \tag{2.2}$$

where μ represents the grand mean of $N \times n$ observations; b_i is a random effect representing the individual deviation from the grand mean, and assumed to have the following distribution $b_i \overset{\text{iid}}{\sim} N(0, \sigma_b^2)$ with σ_b^2 as a between-subjects variance; t_j represents the effect of time; e_{ij} is the error for subject i at time j and follows $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$ with σ_e^2 as a within-subject variance.

Assumptions:

(1) $\sum_{j=1}^{m} t_j = 0$;

(2) $E(\mathbf{y}_{ij}) = \mu + t_j$;

(3) $\text{Var}(\mathbf{y}_{ij}) = \text{Var}(\mu + t_j + b_i + e_{ij}) = \sigma_b^2 + \sigma_e^2$;

(4) $\text{Cov}(\mathbf{y}_{ij}, \mathbf{y}_{i'j}) = 0$ for $i \neq i'$ and $\text{Cov}(\mathbf{y}_{ij}, \mathbf{y}_{ij'}) = \sigma_b^2$ for $j \neq j'$.

In summary, the fourth assumption shows that the observations from different subjects are independent, but those within the same subject are dependent.

The covariance between observations on the same subject is σ_b^2, which has the following corresponding correlation

$$\text{Cor}(\mathbf{y}_{ij}, \mathbf{y}_{ij'}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}. \tag{2.3}$$

Then we can obtain the variance-covariance matrix Σ for the observations in subject i, that is

$$\Sigma = \begin{pmatrix} \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & ... & \sigma_b^2 \\ ... & ... & ... & ... & ... \\ \sigma_b^2 & ... & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 \end{pmatrix}$$

where this structure is also called **compound symmetry structure**.

ANOVA Table

Later an ANOVA table is need to test the variance components σ_b^2 and σ_e^2:

Table 2.3: *ANOVA Table.* $\bar{y}_{..}$ *indicate total mean of Nn observations,* $\bar{y}_{i.}$ *indicate subject means* $(i = 1, ..., N)$*,* $\bar{y}_{.j}$ *indicate time point mean* $(j = 1, ..., m)$*[Hedeker, 2006]*

Source	df	SS	MS	F
Subjects	N-1	$SS_S = m\sum_{i=1}^{N}(\bar{y}_{i.} - \bar{y}_{..})^2$	$\frac{SS_S}{N-1}$	$\frac{MS_S}{MS_R}$
Time	m-1	$SS_T = N\sum_{i=1}^{N}(\bar{y}_{.j} - \bar{y}_{..})^2$	$\frac{SS_T}{m-1}$	$\frac{MS_T}{MS_R}$
Residual	(N-1)×(m-1)	$SS_R = \sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$	$\frac{SS_R}{(N-1)(m-1)}$	
Total	N×m-1	$SS_y = \sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{..})^2$		

Hypothesis test

In order to perform statistical inference, two hypothesis tests can be set up, and then we have the following results:

(1) When the hypothesis H_s: $\sigma_b = 0$ is true, we have $F_S = \frac{MS_S}{MS_R} \overset{\mathrm{H_S}}{\sim} F_{N-1,(N-1)(m-1)}$;

(2) When the hypothesis H_T: $t_1 = t_2 = ... = t_m$=0 is true, we have $F_T = \frac{MS_T}{MS_R} \overset{\mathrm{H_T}}{\sim} F_{n-1,(N-1)(m-1)}$.

In this kind of univariate model analysis, we note that the analysis of variance is only valid if the assumption of compound symmetry structure holds for the covariance matrix. However it is not a reasonable assumption for most situations. In reality, the variance within subjects often changes over time, and the covariances of two closer observations in time are usually greater than those of two observations that are far away of time point [Frees, 2004].

2.2.2 Multivariate Analysis of Variance Using MANOVA

Compared to the univariate repeated measure ANOVA, using MNOVA to analyze repeated measures requires less restrictive assumptions. However, these two methods have many similarities. For instance, both repeated-measures ANOVA and MANOVA assume that time intervals are equally spaced, the response is normally distributed, but both approaches are robust against violations of normality. Both approaches require complete data for all subjects, i.e. no missing observations for any subjects. The only difference is that repeated-measures ANOVA assumes sphericity, or compound symmetry [Davidian, 2011].

With the same notations used before, the model for MNOVA is written as

$$\mathbf{y}_i = X_i\beta + \epsilon_i, i = 1, ..., N, \epsilon_i \sim N_m(0, \Sigma) \tag{2.4}$$

And Σ is an arbitrary covariance matrix with no particular structure, that is ,

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2 \end{pmatrix}$$

2.3 Limitation of Repeated Measures ANOVA and MANOVA

Based on the discussion with the two classical methods: repeated measures ANOVA and MANOVA, some assumptions and restrictions of these two classical methods are highlighted below [Davidian, 2011].

(1) The basic and characteristic of both ANOVA and MANOVA methods is that all subjects should be observed at the same m time points. But in reality, it is unachievable to have

perfect balance due to various reasons. For instance in DCCT dataset, some patients had participated for 10 years, while some other patients had only participated 3 year.

(2) The repeated measures ANOVA has the assumption of compound symmetry, which implies a very specific pattern of correlation among observations within a subject at different time points. It means that the correlation among all observations on a give subject is the same regardless of how near or far apart the observations are taken in time. The multivariate methods make no assumption about the covariance matrix. They allow the possibility that covariance structure could be anything rather than the random variation between-subject and possible correlation within-subject variation.

(3) Both methods assume that the covariance matrix of observations within each subject is same for all subjects. In DCCT, people were randomly assigned to two different treatment groups, conventional treatment, and intensive treatment. We cannot assume that $\mathbf{y_i}$'s from different groups have the same covariance matrix Σ which would be inappropriate. Even for the subjects within the same treatment groups, assuming the same covariance matrix for each individual may not be appropriate either.

(4) There are also some problems related with other covariates. In DCCT, we believe that the age at diagnosis of diabetes may play a role in the lipid level, and the covariate, age of diagnosis, is not time-varying as being measured only once. On the other side, such as the covariate, HbA1c, was measured at each of the same time points as the response lipid measures, and thus is time-dependent. Therefore we need more flexible approaches to catch the patterns of the data.

2.4 Linear Mixed-effects Model

Mixed-effects models provide a flexible and powerful tool for the analysis of longitudinal data. It has been a popular method to model the between-subject and within-subject correlations, to handle both balanced and unbalanced scenarios, and allows the inclusion of covariables.

In the previous section, we described a model (2.2), which is a special example of mixed-effect model. As before $\boldsymbol{\beta}$ is defined as the column vector containing μ and $b_i^{'}$ s, while X_i to be a matrix of $0^{'}$s and $1^{'}$s with n rows each for element of $\mathbf{y_i}$. With these notations, the regression model can be displayed as below

$$\mathbf{y_i} = X_i\boldsymbol{\beta} + \boldsymbol{\varepsilon_i}, \text{ for } i = 1, \ldots, N; \tag{2.5}$$

where $\boldsymbol{\varepsilon_i}$ is the overall error vector with Cov($\boldsymbol{\varepsilon_i}$)=$\Sigma_i$. Note that both the univariate and multivariate ANOVA models could be written in this way also. The alternative definition (2.5) allows for fitting unbalanced data, and can incorporate time-varying and none time-varying covariates.

In mixed-effects models, response variables are assumed to be a function of fixed effect, non-observable random effect, and error term. When both the fixed and the random effects contribute linearly to the response, the model is called linear mixed-effects model, and when some of the fixed and/or random effects occur nonlinearly in the response function, then the model is called nonlinear mixed-effect model [Frees, 2004]. In the thesis, we only discuss the application of linear-mixed effect model in DCCT and GWAS analysis and consider the models in which the error terms and the random effects are normally distributed.

2.4.1 Linear Mixed-Effects Model (LME)

The LME model described by Laird and Ware (1982) can be written as

$$\mathbf{y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i} + \boldsymbol{\varepsilon_i}, i = 1, \ldots, N. \tag{2.6}$$

Where $\mathbf{y_i}$ is a n_i dimensional response vector for the ith subject; $\boldsymbol{\beta}$ is a p-dimensional vector including all the fixed effects; $\mathbf{b_i}$ is the q-dimensional vector of random effects; $\mathbf{X_i}$ $(n_i \times p)$ is the known $(n_i \times q)$ fixed-effects coefficient matrix; $\mathbf{Z_i}$ is the known $(n_i \times q)$ random-effects coefficient matrices; $\boldsymbol{\varepsilon_i}$ is the n_i-dimensional within-subject error vector.

Furthermore, there are some model assumptions:

(1) The random effect $\mathbf{b_i}$ and within-subject error $\boldsymbol{\varepsilon_i}$ are independent for different subjects and independent of each other for the same subject. i.e. Cov($\mathbf{b_i}, \mathbf{b_j}$)=0 if i≠j, Cov($\boldsymbol{\varepsilon_i}, \boldsymbol{\varepsilon_j}$)=0 if $i \neq j$, and Cov($\mathbf{b_i}, \boldsymbol{\varepsilon_i}$)=0.

(2) $b_i \sim MVN_q(0, D)$, where D is a general $(q \times q)$covariance matrix with (i, j) element $\sigma_{ij} = \sigma_{ji}$.

(3) $\boldsymbol{\varepsilon_i} \sim MVN_{n_i}(\mathbf{0}, \boldsymbol{\Sigma_i})$, where Σ_i is an $n_i \times n_i$ covariance matrix.

2.4.2 Likelihood Estimation for LME Models

We focus on two general estimation methods: maximum likelihood (ML) and restricted maximum likelihood (REML) in the thesis [Verbeke and Molenberghs, 1997].

Formula 2.6 can be written as

$$\mathbf{y_i}|\mathbf{b_i} \sim MVN(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i}, \boldsymbol{\Sigma}_i), b_i \sim MVN(0, \mathbf{D})$$

It is therefore called a hierarchical model, in which a conventional density of $\mathbf{y_i}$ follows a multivariate normal. This model can be shown as below [Verbeke and Molenberghs, 2000],

$$\mathbf{y_i} \sim MVN(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z_i}\mathbf{D}\mathbf{Z}_i^{'} + \Sigma_i) \tag{2.7}$$

Let $\boldsymbol{\alpha}$ denote the vector of all variance and covariance parameters (usually called variance components) in $\mathbf{V}_i = \mathbf{Z_i}\mathbf{D}\mathbf{Z}_i^{'} + \Sigma_i$, i.e. $\boldsymbol{\alpha}$ consists of the all different elements in D and all parameters in Σ_i. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{'}, \boldsymbol{\alpha}^{'})^{'}$ be the vector of all the parameters in the model 2.7 for $\mathbf{y_i}$. The classical approach to inference is based on estimators obtained from maximizing the likelihood function 2.7

$$L_{ML}(\boldsymbol{\theta}) = \Pi_{i=1}^{N}\{(2\pi)^{-n_i/2}|V_i(\boldsymbol{\alpha})|^{-\frac{1}{2}} \times \exp(-\frac{1}{2}(\mathbf{y_i} - \mathbf{X}_i\boldsymbol{\beta})^{'}V_i^{-1}(\boldsymbol{\alpha})(\mathbf{y_i} - \mathbf{X}_i\boldsymbol{\beta})\} \tag{2.8}$$

with respect to $\boldsymbol{\theta}$. There are two situations about $\boldsymbol{\alpha}$, known or unknown.

1. **Assume $\boldsymbol{\alpha}$ to be known**

   We directly get $\mathbf{W_i} = \mathbf{V}_i^{-1}(\boldsymbol{\alpha})$, then the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ can be shown as

   $$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} X_i^{'}W_iX_i\right)^{-1}\sum_{i=1}^{N} X_i^{'}W_i\mathbf{y}_i \tag{2.9}$$

   .

2. **Assume $\boldsymbol{\alpha}$ is unknown**

   In most cases, $\boldsymbol{\alpha}$ is not known, and needs to be replaced by an estimate $\hat{\boldsymbol{\alpha}}$. As for obtaining $\hat{\boldsymbol{\alpha}}$, two commonly used methods are ML and REML.

(1) **ML** Maximizing $L_{ML}(\boldsymbol{\alpha}, \hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}))$ with respect to α, to obtain $\hat{\boldsymbol{\alpha}}_{ML}$, then estimate $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\alpha}}_{ML})$. $\hat{\boldsymbol{\alpha}}_{ML}$ and $\hat{\boldsymbol{\beta}}_{ML}$ can also be obtained from maximizing $L_{ML}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, i.e. , with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ simultaneously.

(2) **REML** The REML estimator for the variance components $\boldsymbol{\alpha}$ is obtained from maximizing the likelihood function of a set of error contrasts [Harville, 1974], $U = K^{'}Y$, where K is a $n\times(n\text{-}p)$ full-rank matrix with columns orthogonal to the columns of X matrix. Then we combine all models $\mathbf{y_i} \sim N(X_i\boldsymbol{\beta}, V_i)$ into one model $\mathbf{y} \sim N(X\boldsymbol{\beta}, V)$, where

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, V(\boldsymbol{\alpha}) = \begin{pmatrix} V_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_N \end{pmatrix}$$

So we can obtain,

$$U = \begin{pmatrix} \mathbf{y}_1 - \mathbf{y}_2 \\ \mathbf{y}_2 - \mathbf{y}_3 \\ \vdots \\ \mathbf{y}_{N-2} - \mathbf{y}_{N-1} \\ \mathbf{y}_{N-1} - \mathbf{y}_N \end{pmatrix} = K^{'}Y \sim N(0, K^{'}V(\boldsymbol{\alpha})K)$$

Then the MLE of $\boldsymbol{\alpha}$, which is based on U is called the REML estimate, and denoted by $\hat{\boldsymbol{\alpha}}_{REML}$. Similarly, resulting estimate $\beta(\hat{\boldsymbol{\alpha}}_{REML})$ for $\boldsymbol{\beta}$ will be denoted by $\hat{\boldsymbol{\beta}}_{REML}$.

$\hat{\boldsymbol{\alpha}}_{REML}$ and $\hat{\boldsymbol{\beta}}_{REML}$ can also be obtained from maximizing (2.10)

$$L_{REML}(\boldsymbol{\theta}) = |\sum_{i=1}^{N} X_i' W_i(\boldsymbol{\alpha}) X_i|^{-\frac{1}{2}} L_{ML}(\boldsymbol{\theta}) \tag{2.10}$$

with respect to all parameters simultaneously ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$). Here note that $L_{REML}(\boldsymbol{\theta})$ is not the likelihood for our original data Y.

In general, ML and REML both have the same merits of being based on the likelihood principle which leads to useful properties such as consistency, asymptotic normality, and efficiency. But the REML produces less biased estimators for many special cases [Verbeke & Molenberghs, 1997]. In R software, the default method is REML.

2.4.3 Inference for Marginal Model Parameters

In practice, inference on the parameters in a fitted model is often of a primary interest, due to the generalization of results from a specific sample to general population [Verbeke and Molenberghs, 2000]. Inference for the parameter vector β in the mean structure and variance component $\boldsymbol{\alpha}$ in D and in all Σ_i is described below.

Inference for Fixed-effects Parameters

As discussed in Section 2.4.1, the vector $\boldsymbol{\beta}$ of fixed effects is estimated by

$$\widehat{\beta}(\alpha) = \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} \sum_{i=1}^{N} X_i' W_i \mathbf{Y}_i \tag{2.11}$$

where $W_i = V_i^{-1}(\alpha)$, the unknown $\boldsymbol{\alpha}$ of variance component is replaced by its ML or REML estimate.

Under the marginal model (2.7), and conditionally on $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\beta}}(\alpha)$ follows a multivariate normal distribution with mean vector β and the variance-covariance matrix

$$\begin{aligned}\mathrm{var}(\hat{\beta}) &= \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' W_i \mathrm{Var}(y_i) W_i X_i\right) \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \qquad (2.12)\\ &= \left(\textstyle\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' W_i \mathrm{V} W_i X_i\right) \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1}\\ &= \left(\textstyle\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' W_i X_i\right) \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1}\\ &= \left(\textstyle\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1}\end{aligned}$$

Considering general linear hypotheses of the fixed effects, i.e., testing problems of the form

$$H_0 : H^{'}\boldsymbol{\beta} = 0 \qquad \text{versus} \qquad H_1 : H^{'}\boldsymbol{\beta} \neq 0,$$

with $\boldsymbol{\beta}$ a $(p\times\ 1)$ vector and H a $(p \times q)$ matrix $(q \leq p)$ of full rank $(rank(H)) = q)$ Then the statistic

$$\frac{(H^{'}\boldsymbol{\beta})^{'}(H^{'}(X^{'}WX)^{-1}H)(H^{'}\boldsymbol{\beta})}{rank(H)}$$

approximately follows a F-distribution with $rank(H)$ as degrees of freedom for the numerator [Helms, 1992]. The denominator degrees of freedom is the total numbers of observations minus 1 and minus q.

Inference for Variance Component

With respect to the estimates of the variance components $\boldsymbol{\alpha}$, it is necessary to construct statistical tests for the significance of the random effects in the model [Verbeke and Molenberghs, 2000].

When testing the variance components against $\mathbf{0}$, i.e. the null hypothesis:

$$H_0 : \boldsymbol{\alpha} \in \boldsymbol{\Theta}_{\boldsymbol{\alpha},\mathbf{0}}$$

$\boldsymbol{\Theta}_{\boldsymbol{\alpha},0}$ is a subspace of parameter space $\boldsymbol{\Theta}_{\boldsymbol{\alpha}}$ of the variance components $\boldsymbol{\alpha}$, the value $\mathbf{0}$ of the parameters under hypothesis H_0 are on the boundary of the parameter space. Therefore, the well-known asymptotic normality for estimators and the asymptotic chi-squared null distribution for the likelihood ratio tests (LRT), are not necessarily valid any longer.

We used $\boldsymbol{\alpha}$ to represent the covariance component in the model, which is estimated with ML method or REML method. REML is used in estimating covariance components by maximizing the likelihood function of a set of error contrasts rather than maximizing the likelihood function of the data [Stram and Lee, 1994]. So, when comparing two nested models with equal mean structure (fixed effects), but different covariance structure, the REML likelihood are comparable because the same mean structure leads to the same error contrasts [Verbeke and Molenberghs, 2000].

Stram and Lee (1994) give details of the more general models, and the asymptotic results of likelihood tests would be adjusted for the boundary conditions. Consequently, LR test statistics under the null often follow a mixture of chi-squared distributions instead of one single chi-squared distribution [Stram and Lee, 1994]. Pinheiro and Bates (2000) also used simulations to demonstrate the effect of these adjustments.

This LRT was derived under the assumption of conditional independence, i.e., assuming that all residual covariance $\boldsymbol{\Sigma}_i$ are of the form $\sigma^2 I_{n_i}$ [Pinheiro & Bates, 2000]. However it is likely conservative, as the p-values are normally larger than they should be [Verbeke and Molenberghs, 2000]. Therefore, the test is still useful if p-values show the significant results. Let L_{ML} denote the likelihood function (2.9) and let $-2ln\lambda_N$ be the likelihood ratio test statistic

defined as

$$-2ln\lambda_N = -2ln\left[\frac{L_{ML}(\hat{\boldsymbol{\theta}}_{ML,0})}{L_{ML}(\hat{\boldsymbol{\theta}}_{ML,0})}\right]$$

where $\hat{\boldsymbol{\theta}}_{ML,0}$ and $\hat{\boldsymbol{\theta}}_{ML,1}$ are the maximum likelihood estimates under the null-hypothesis and under the alternative hypothesis respectively.

2.4.4 Residual Correlation Structures for Modeling Dependence

In longitudinal data analysis, when subjects are followed over time, there is a natural ordering of the data for each subject. Correlation structure are used to model dependence among observations, in mixed-effect model, it is used to model dependency among the within-group errors [Pinheiro and Bates, 2000]. There are up to ten observations per subject in our DCCT data set, the time is yearly integer variable. The correlation between two within-group errors $\varepsilon_{ij}, \varepsilon_{ij'}$ is assumed to depend on some distance between them [Cressie, 1993], and ρ is a vector of correlation parameters. Jones (1993) described the serial correlation structures in detail of the linear mixed-effects models, The general serial correlation model is defined as $cor(\varepsilon_{ij}, \varepsilon_{ij'}) = h(\rho)$, where $h(\cdot)$indexes autocorrelation function. Some of the most common serial correlation structures used in practice are shown below, and all of which are implemented in the R *nlme* library [Pinheiro, 2011].

Compound Symmetry

Compound symmetry is the simplest serial correlation structure, which assumes equal correlation among all within-group errors of same subject. The corresponding correlation model is

$$\text{cor}(\varepsilon_{ij}, \varepsilon_{ij'}) = \rho, \forall j \neq j', h(k, \rho) = \rho, k = 1, 2, ...$$

While the compound symmetry correlation model tends to be too simplistic for practice applications.

General (Unstructured)

The general correlation structure represents the other extreme in complexity to the compound symmetry structure. Each correlation is shown by a different parameter, the correlation function is $h(\rho) = \rho_k, k = 1, 2, ...$ While the general correlation model tends to over parameterized model. It is useful for few observations per subject, that leads to precise correlation with observations.

Autoregressive (AR)

Box et al.(1994) described the family of correlation structure which includes different classes of linear stationary models: autoregressive models, moving average models, and mixture of autoregressive-moving average models. We use ϵ_t indexes an observation taken at time t, μ_t indexes a noise term with $E[\mu_t] = 0$, and assumed independent of the previous observations.

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \cdots + \phi_p \epsilon_{t-p} + \mu_t, \ |\phi| < 1$$

p is called the order of the autoregressive model, which denoted by AR(p). There are p correlation parameters in an AR(p) model, given by $\phi = (\phi_1, \cdots, \phi_p)$. The AR(1) model is the simplest and one of most useful autoregressive model. Its correlation function is $h(k, \phi) = \phi^k, k = 0, 1, ...$

For autoregressive models of order greater than 1, the correlation function was defined as below equation (Box et al., 1994).

$$h(k, \phi) = \phi_1 h(|k-1|, \phi) + \cdots + \phi_p h(|k-p|, \phi), k = 1, 2, ...$$

Moving Average Correlation (MA)

Moving average correlation models assume that the current observation is a linear function of independent and identically distributed noise terms. The MA(q) model is called the first order moving average correlation model, in which the current observation is modeled as a linear function.

$$\epsilon_t = \theta_1 \mu_{t-1} + \cdots + \theta_q \mu_{t-q} + \mu_t$$

q is called the order of the moving average model, which is denoted by MA(q). There are q correlation parameters in an MA(q) models, given by $\theta = (\theta_1, \cdots, \theta_q)$. The correlation function for an MA(q) model is

$$h(k,\theta) = \begin{cases} \frac{\theta_k + \theta_1\theta_{k-1} + \cdots + \theta_{k-q}\theta_q}{1+\theta_1^2+\cdots+\theta_q^2}, & \text{k=1,...,q;} \\ 0, & \text{k=q+1,q+2,....} \end{cases}$$

Mixed Autoregressive-moving Average Models (ARMA)

Mixed autoregressive-moving average models are combined together an autoregressive model and a moving average model. In ARMA(1, 1) model, for instance,

$$\epsilon_t = \sum_{i=1}^{p} \phi_t \epsilon_{t-i} + \sum_{j=1}^{n} \theta_j \mu_{t-j} + \mu_t,$$

By convention, ARMA(p,0)=AR(p), and ARMA(0,q)=MA(q), so that both autoregressive and moving average models are particular examples of the general ARMA model. Information criteria can used to evaluate two models based on their maximized log likelihood values. The model with the smaller information criterion is usually preferred. Two commonly used information

criteria are Akaike Information Criterion (AIC) [Sakamoto, 1986] and the Bayesian Information Criterion (BIC) [Schwarz, 1978]. These are model comparison criteria evaluated as

$$\text{AIC=-2log Likelihood} + 2n_{par},$$

$$\text{BIC=-2log Likelihood} + n_{par}\log(N),$$

Where n_{par} denotes the total number of parameters in the model and N is the total number of observations used to fit the model. If we use AIC to compare models for the same data, we prefer the model with the smaller AIC. Similarly, when using BIC we prefer the model with the lowest BIC.

2.5 Model Diagnostics

After selecting the proper covariance structure, evaluation of the final model is necessary. In the linear model, the collinearity of predictors may affect the residual distribution, so collinearity among covariates needs to be checked [Cheng et al, 2011]. And, we need to check the distributional assumptions of the LME model with random effects and residual terms. The *nlme* library library provides several methods for assessing the validity of these assumptions.

2.5.1 Assess Collinearity among Covariates

When the degree of multicollinearity arises, the estimates of the coefficients become unstable and the standard errors of the coefficients can be inflated [UCLA,web book]. Stinnett (1993) gave a detailed discussion of diagnostics and collinearity in mixed model to avoid inflated variances covariates' coefficients in fixed effects. We say that there exist multicollinearity among

the predictors if there exist a linear combination of the regressors which is almost zero:

$$\exists c_j : c_0 + \sum_{j=1}^{p-1} c_j X_j \approx 0, \quad j = 1, ..., p-1.$$

A formal method for diagnosing the multicollinearity is by means of variance inflation factors (VIF) or tolerance which is defined as 1/VIF.

$$\text{VIF}_j = \frac{1}{1 - R_j^2}, \quad j = 1, ..., q.$$

where R_j^2 is the squared correlation of predictor j with the remaining $q-1$ predictors ($q-2$ if the design matrix X includes an intercept). The value of (1-R_j^2) is often called the tolerance. There is strong multicollinearity if the largest VIF is larger than 10, or tolerance value lower than 0.1. Often centering and scaling can substantially reduce collinearity [Cheng et al, 2011].

2.5.2 Assess LME Assumption

(1) Assumption of Random Effect

In practice, histograms of empirical Bayesian estimates are often used to check the normality assumption for the random effects [Pinheiro and Bates, 2000]. In *nlme* library, *qqnorm* can be used to get normal plot of estimated random effects for checking marginal normality and identifying outliers.

(2) Assumption of Residual Terms

There are two kinds of residuals in the mixed model, marginal residual and conditional residual. Marginal residual is a deviation of a subject from group mean, and the conditional residual is a deviation of one measurement within a subject from the mean of that subject over time.

Gurka et al. (2006) claim only normality of marginal residuals is needed. The studentized residual follows a t distribution [Kleinbaum et al, 1988] and its distribution corresponds to the distribution of a predicted future observation [Atkinson, 1985]. Cheng et al, (2011) recommend using jackknifed studentized residual histograms, box plots, and scatter plots of jackknifed studentized residuals versus predicted values over time to help the assessment.

2.6 Fitting the LME Model with R

There are several packages in R for fitting LME model. In this thesis, analysis were performed using R version 2.13.0 with *nlme* package version 3.1-102 [Pinheiro et al., 2011]. Some codes we used in the thesis are listed in appendix.

Chapter 3

The Analysis of the Diabetes Control and Complications Trial (DCCT) Data

In this chapter, we used a real example to illustrate the procedure and strategies discussed in this thesis. 1303 Caucasian probands with T1D in DCCT were followed from 1983 to 1993. LME model will be constructed with function of fixed effects, random effects and correlation structure.

3.1 Introduction of the DCCT Data Set

The DCCT was a randomized clinical trial which was designed to determine if intensive diabetes management, with the goal of normalizing glycaemia levels, would prevent or delay the progression of long-term diabetic complications [DCCT, 1995]. Based on scientific considerations, many factors have a potential impact on the four lipid measures (CHL, HDL, LDL, TRG)(Table 3.2). Some of the factors are non-time varying covariates including cohort, treat-

ment, gender, duration of diabetes prior to diabetes, age of diagnosis, and baseline C-peptide which is a protein that is produced in the body along with insulin (Table 3.1), and time-varying covariates consisting of BMI, HbA1c, insulin dose, and exercise (Table 3.3).

There were two cohorts in the study: the primary prevention cohort included 650 subjects, whose diabetes durations prior to entry into the trial were in the range of 1 to 5 years, with no retinopathy and Albumin Excretion Rate (AER) $< 40mg/day$, which is an important predictor of diabetes nephropathy; the secondary intervention cohort included 653 subjects, whose prior diabetes durations were between 1 and 15 years, with mild retinopathy and AER $< 200mg/day$. Patients in each cohort were randomly assigned to one of two treatment groups: conventional treatment and intensive treatment. The intensive therapy aimed at maintaining glycaemia levels measured by Haemoglobin A1c (HbA1c) as close as possible to the normal value of 6% or less, while the conventional therapy aimed at maintaining clinical well-being with no specific glucose targets [Singer, et al., 1992].

3.2 Data Analysis

In DCCT data set, some observations need to be cleaned or removed [Rahm and Do, 2000]. For example, one patient had amputation at DCCT year 6 because of diabetes complication, and then his body mass index (BMI) was 60. Therefore we considered his BMI values for the following years as missing values. The lipid profile did not measure LDL cholesterol directly but instead estimated them using the Friedewald equation by subtracting the amount of cholesterol associated with other particles LDL = CHL − HDL − 0.2 × TRG [Friedewald et al., 1972]. However there are some limitations for this method: LDL can not be calculated if plasma

Table 3.1: *Mean and standard deviation of diabetes duration prior to diabetes and centered age of diagnosis. Numbers of patients were grouped by Male and Female in different cohort, treatment and baseline C-peptide.*

Covariate	Primary cohort		Secondary Cohort	
Cohort (number)	650		653	
Treatment	Conventional	Intensive	Conventional	Intensive
Gender (M:F)	188 :185	152 :155	174 :149	180 :150
Diabetes duration (month)	31.8(16.6)	32.4(16.8)	104.7(44.7)	105.1(45.2)
Age of diagnose	23.5(7.8)	24.2(7.6)	18.4(7.5)	18.7(7.6)
C-peptide(ng/ml) (number)				
(0,0.03]	84	72	199	221
(0.03,0.2]	142	126	103	94
(0.2,1]	117	106	21	15

Table 3.2: *Mean and standard deviation of four lipid measures for DCCT Data*

Lipids(mg/dl)	Primary cohort		Secondary Cohort	
Treatment	Conventional	Intensive	Conventional	Intensive
CHL	179.25(17.03)	183.84(18.16)	178.17(17.44)	179.54(17.48)
HDL	52.08(5.71)	49.70(5.63)	52.68(5.76)	50.63(5.66)
LDL	111.92(13.9)	116.81(5.19)	110.65(13.97)	112.78(14.62)
TRG	80.66(25.2)	86.57(25.54)	73.84(22.63)	80.54(24.35)

Table 3.3: *Mean and standard deviation of time-varying covariates for DCCT Data*

Covariates	Primary cohort		Secondary Cohort	
Treatment	Conventional	Intensive	Conventional	Intensive
BMI(kg/m^2)	24.17(1.09)	24.48(1.04)	25.07(1.53)	25.29(1.43)
Insulin(u/kg/day)	0.66(0.09)	0.67(0.08)	0.70(0.13)	0.72(0.12)
HbA1c(%)	9.14(0.96)	8.89(0.86)	7.45(0.97)	7.51(0.94)
Exercise(1-4)	2.96(0.55)	2.99(0.58)	2.98(0.54)	3.09(0.55)

triglyceride is $> 400mg/dl$, so in such a case, LDL cholesterol should be set as missing values.

The four lipid measures (CHL, HDL, LDL, TRG) are the response outcomes of interest, and are treated as continuous variables. The distributions illustrated in Figure 3.1 clearly show the four lipid measures all have long tails skewed to the right. To minimize the impact of these extreme observations, we winsorized the four lipid values at 99.5 percentile [Sheskin D, 2003]. In this thesis, we only provide the detailed analysis for CHL measure.

3.2.1 Linear Mixed-effects Model

Figure 3.2 shows that CHL values of 50 random patients change over DCCT years. There was a lot of variation of CHL values among patients, and also a lot of variation for each patient over time. Since some patients participated the DCCT project for 4 years, some for 10 years, the data are unbalanced. Therefore we can not analyze the data set with either repeated-measure ANOVA or MANOVA due to the limitations we have discussed in the previous chapter.

We used linear mixed-effects model to analyze data with random intercept and time since

Figure 3.1: *Distribution of four lipid measures*

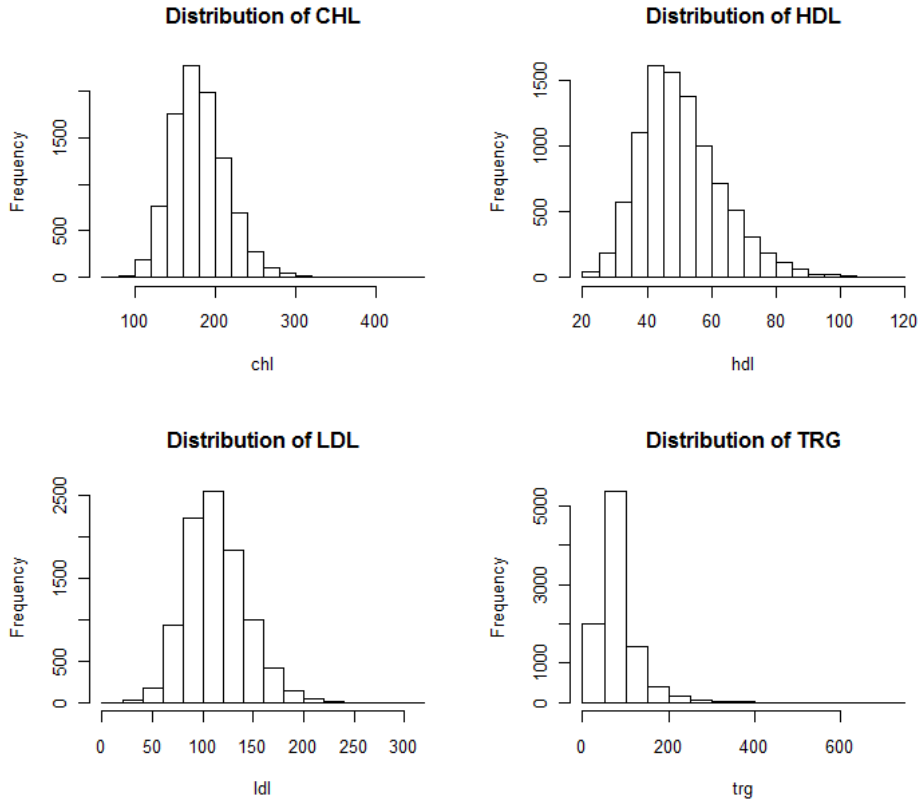
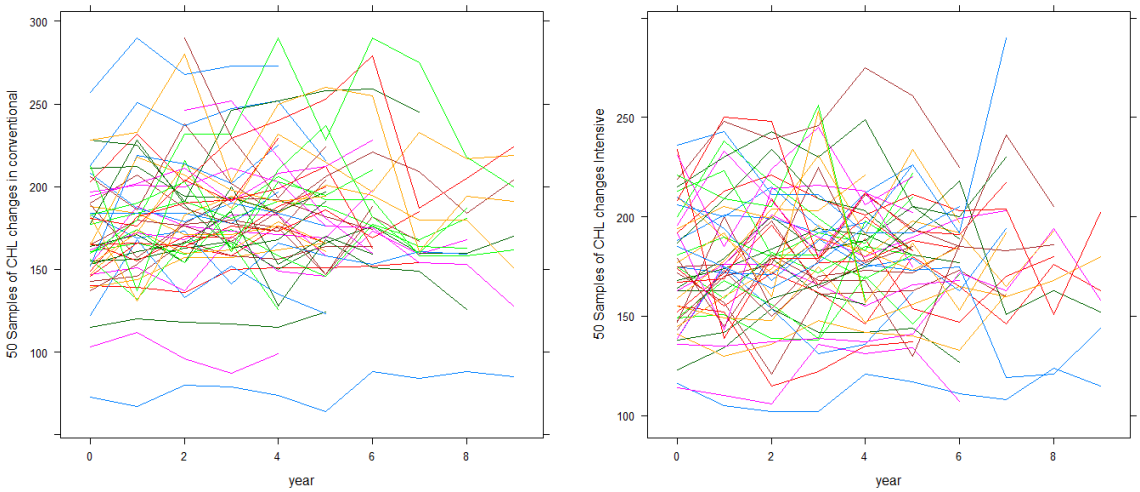
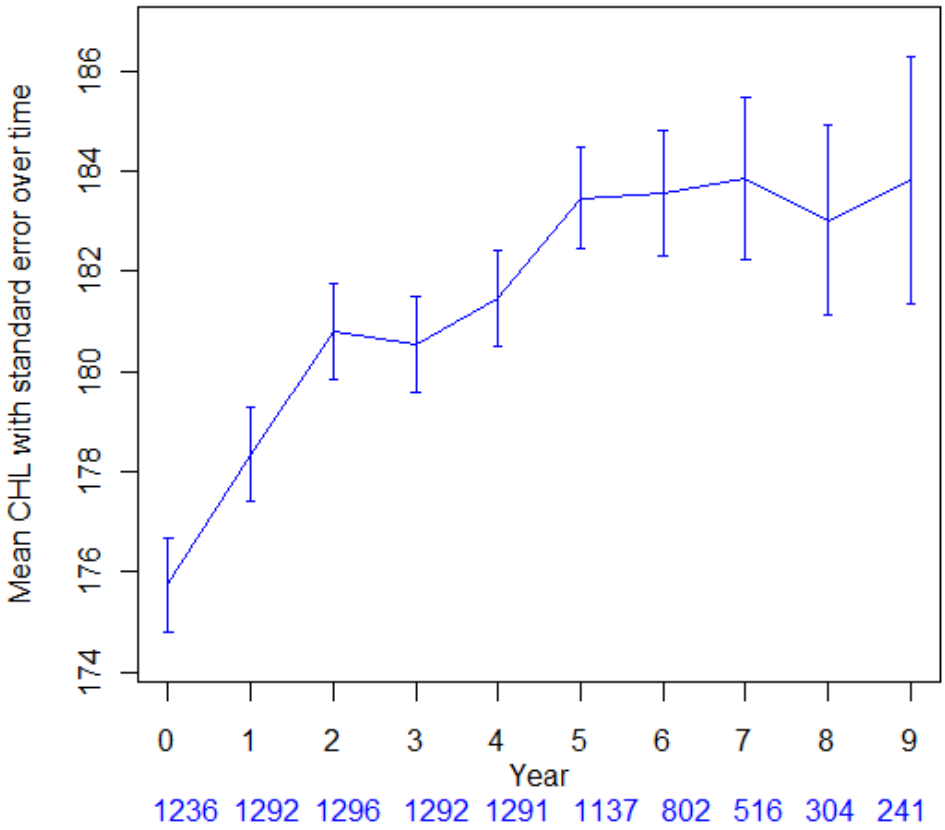


Figure 3.2: *50 Samples of CHL changes over time in two treatment groups*



patients have different baseline values and also values increase or decrease randomly over time based on Figure 3.2. Figure 3.3 displays change of the mean CHL value (the solid line) over time

Figure 3.3: *Mean CHL changes over time with standard error in DCCT years*



with corresponding standard error (the vertical bars). The numbers of x-label are the number of patient visit each year. It shows the increasing tendency with minor fluctuations (from year 2 to 5 and from year 7 to 9), so a quadratic time may be required. Then we considered the fitted model with the quadratic time as a random effect.

3.2.2 Selection of Fixed Effects

The criteria of selecting covariates are based on both statistical consideration and scientific supportance in a specific area [Cheng et al., 2010].

From Section 3.1, two cohorts were distinguished by diabetes duration, these two variables were highly correlated. However, the cohort was a design variable, patients were randomly assigned to two different treatments in each cohort, so cohort as a covariate should be included in the model. Diabetes complications are highly associated with diabetes duration, and the purpose of DCCT was to prevent or stop the diabetes complication, so diabetes duration also should be modeled.

Figure 3.3 indicates that the time and the square of time could be included in the model, although they are highly correlated (VIF=9.57). To avoid collinearity, we centered the time by subtracting the mean so that the centered time and the squared centered-time (VIF=1.27, later is called the time square) could be covariates in the model. A likelihood ratio test confirmed that the squared centered-time should be in the model as a fixed effect (p-value=0.00017), and the coefficient $\hat{\boldsymbol{\beta}}$ for the time square equals -0.17 with p-value=0.0002 [Pinheiro and Bates, 2000].

We then centered the age of diagnosis by subtracting the mean for the purpose of avoiding collinearity with diabetes duration prior to DCCT (VIF=1.30). Therefore the variable "centered age of diagnosis" is included as a covariate in the model.

Furthermore we also considered some non-time varying covariances as fixed effects in the model, they are gender, treatment. The baseline indicator was 1 at baseline year, then changed to 0 in other DCCT years. We studied treatment and baseline indicator interaction, because every patient was treated with conventional treatment at baseline, and then was assigned to two different treatment in other DCCT years, therefore the effect of two different treatments should start from DCCT year 1. That is the reason of baseline indicator and its interaction

with treatment should be included in the model. In conclusion, we used time, time square, gender, cohort, treatment, diabetes duration, age of diagnose, and baseline indicator as fixed effects in the model.

3.2.3 Selection of Random Effects and Correlation Structure

After determining the fixed effects, we need to select a set of random effects which can help defining a model. Verbeke and Molenberghs (2000) discussed that the random effects for time-independent covariates can be interpreted as subject-specific correlation to the overall mean structure, which makes them hard to distinguish from random intercepts. Therefore, one often includes random intercepts, and random effects only for time-varying covariates. In this case, the only time-varying variable are linear time and squared time themselves. We decided to use random intercept and random linear time and squared time as random effects in the fitted model.

Then the model is shown by

$$y_{i,j} = \beta_0 + b_{0,i} + (\beta_1 + b_{1,i})t_{i,j} + (\beta_2 + b_{2,i})t_{i,j}^2 + \sum_{k=3}^{p} \beta_k X_{k,i,j} + \varepsilon_{i,j}. \tag{3.1}$$

where $y_{i,j}$'s are winsorized CHL values.

i indexes the subjects $i = 1, 2, \ldots, 1303$.

j indexes the time visit for subject i, $j = 1, 2, \ldots, n_i$. n_i represents the overall visits of subject i.

$t_{i,j}$ indexes the centered visit time yearly from baseline, while $t_{i,j}^2$ indexes the squared centered-time.

$\beta_0, \ldots, \beta_p$ are the fixed effect coefficient parameters.

$X_{k,i,j}$ are other covariates.

We also assume that $(b_{0,i}, b_{1,i}, b_{2,i}) \overset{\text{iid}}{\sim} MVN_3(0, \mathrm{D})$, where $\mathbf{D}$ is the random effect variance-covariance matrix with 3×3 dimensions that can be written as below

$$D = \mathrm{Var} \begin{pmatrix} b_{0,i} \\ b_{1,i} \\ b_{2,i} \end{pmatrix} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

The error terms $\varepsilon_{i,j}$'s are assumed independent among individuals but dependent within each patient, i.e. $cov(\varepsilon_{i,1}, \varepsilon_{j,1}) = 0$ of $i \neq j$ and $(\varepsilon_{i,1}, \varepsilon_{i,2}, ...\varepsilon_{i,n_i}) \sim MVN_{n_i}(\mathbf{0}, \Sigma_i)$ where Σ_i is the covariance matrix within each patient.

Then, the next important job is selecting an appropriate covariance or correlation structure for the model, which can explain the correlation among the error terms [Cheng et al.,2010]. In model 3.1, the within-subject random vector $\boldsymbol{\varepsilon_i} \sim MVN(\mathbf{0}, \Sigma_i)$. So both the covariance matrix of random effects (D) and residual errors (Σ_i) were used to describe the covariance structure overall for the data.

$$\mathrm{V}{=}\mathrm{Var}(\mathbf{y_i}) = Z_i D Z_i^{'} + \Sigma_i$$

where Z_i is known matrix for D, and D is often treated as an unstructured covariance matrix since only 6 parameters need to be estimated. However, based on the data set, it is not practical to use unstructured residual error covariance, because each subject has highly correlated observations.

Pinheiro (2000) wrote that the random effects and residual correlation structure are competition pairs, and different random effects affect residual correlation structure and also the

model, vice versa. For instance, if time square is included as a fixed effect in the model, then we need to consider whether time square should also be a random effect included in the model.

We first assumed the conditional independence of the residuals, i.e., given the random effects (the intercept, time, and time square), the covariance of the residuals are a constant matrix. We considered 3 models with the same fixed effects but different random effects:

Model 1: random intercept.

$$y_{i,j} = \beta_0 + b_{0,i} + \sum_{k=1}^{p} \beta_k X_{k,i,j} + \varepsilon_{i,j}.$$

Model 2: random intercept and time.

$$y_{i,j} = \beta_0 + b_{0,i} + (\beta_1 + b_{1,i}) t_{i,j} + \sum_{k=2}^{p} \beta_k X_{k,i,j} + \varepsilon_{i,j}.$$

Model 3: random intercept, time, and time square.

$$y_{i,j} = \beta_0 + b_{0,i} + (\beta_1 + b_{1,i}) t_{i,j} + (\beta_2 + b_{2,i}) t_{i,j}^2 + \sum_{k=3}^{p} \beta_k X_{k,i,j} + \varepsilon_{i,j}.$$

Table 3.4 shows the results of LRT test for the three models with different random effects. Comparing Models 1 and 2, Models 2 and 3, we can see that Model 3 is the best fit for the model. After including the fixed and random effects of Model 3, the correlation structures can be compared and shown in Table 3.5. Through comparing with AIC and BIC, AR(1) correlation structure should be more appropriate than others.

3.2.4 Inference for Fixed Effects

Table 3.6 displays the mean structure of the fixed effects in CHL primary model. Time is significantly associated with CHL in the model ($p_{\text{time}} = 6.5 \times 10^{-5}$, $p_{\text{quadratic time}} = 1.4 \times 10^{-2}$).

Table 3.4: LRT for Random Effects with Assumption of Constant Residual Correlation Structure using REML

Model	DF	AIC	BIC	Loglik	Test	L.ratio	p-value
Model 1	12	86830	86916	-43403			
Model 2	14	86610	86710	-43291	1 vs 2	224	$2.28\times^{-49}$
Model 2	14	86610	86710	-43291			
Model 3	17	86538	86660	-43252	2 vs 3	78.1	7.74×10^{-17}

Table 3.5: *the Comparison of Covariance Structure for CHL in Primary Model.*

Residual Covariance Structure	AIC	BIC	Log-Likelihood
CS	86540.67	86669.46	-43252.33
AR(1)	86537.61	86666.41	-43250.81
AR(2)	86539.17	86675.12	-43250.58
MA(1)	86537.78	86666.56	-43250.86
MA(2)	86539.17	86675.12	-43250.58
ARMA(1,1)	86539.34	86675.29	-43250.67

Table 3.6: *Total cholesterol (CHL) analysis in Primary Model*

	$\hat{\beta}$	Stand.Error	DF	t-value	p-value
(Intercept)	175.2	1.7	8167	99.4	0.1E-30
Centered year	0.5	0.1	8167	3.9	6.5E-05
(Centered year)^2	-0.1	0.1	8167	-2.4	1.4E-02
Intensive Treatment	-4.6	1.5	1297	-2.9	2.7E-03
Baseline Indicator	-3.7	1.1	8167	-3.5	4.6E-04
Secondary Cohort	2.4	2.2	1297	1.1	2.8E-01
Female	6.8	1.5	1297	4.4	9.1E-06
Centered age of diag	0.8	0.1	1297	7.7	2.1E-14
DURATION	0.1	0.1	1297	2.8	5.0E-03
Intensive trx:baseline	5.8	1.2	8167	4.8	1.6E-06

When diabetes duration is included in the model, the cohort is not significantly associated with CHL in the model, while when the cohort is in the model, diabetes duration still has a significant impact on CHL. When all other covariates were fixed in the model, the intensive treatment was significantly associated with CHL after baseline year.

3.2.5 Inference for Variance Components

It may be of interest to test whether random time and random time square effects are both needed after we fixed the correlation structure with AR(1), even though we tested them when correlation structure with constant matrix in Section 3.2.4. The corresponding hypothesis is to test whether the variance component is zero, which is also clearly on the boundary of the parameter space $\boldsymbol{\Theta}_{\boldsymbol{\alpha}}$, and results the classical likelihood-based inference cannot be applied (see the discussion in Section 2.4.3).

We considered the below 3 models which based on same fixed effects and AR(1) correlation structure. The results from LRT are listed in Table 3.7, which further confirmed that random intercept, random time and random time square effects are all necessarily needed in the model.

Model 1: random intercept.

Model 2: random intercept and random time effect.

Model 3: random intercept, random time, and random time square effect.

Table 3.7: LRT for Variance Component with AR(1) Residual Correlation Structure using REML

Model	DF	AIC	BIC	Loglik	Test	L.ratio	p-value
Model 1	13	86681	86774	-43327			
Model 2	15	86587	86695	-43278	1 vs 2	97.81	9.5×10^{-9}
Model 2	15	86587	86695	-43278			
Model 3	18	86537	86666	-43250	2 vs 3	56.13	0.0008

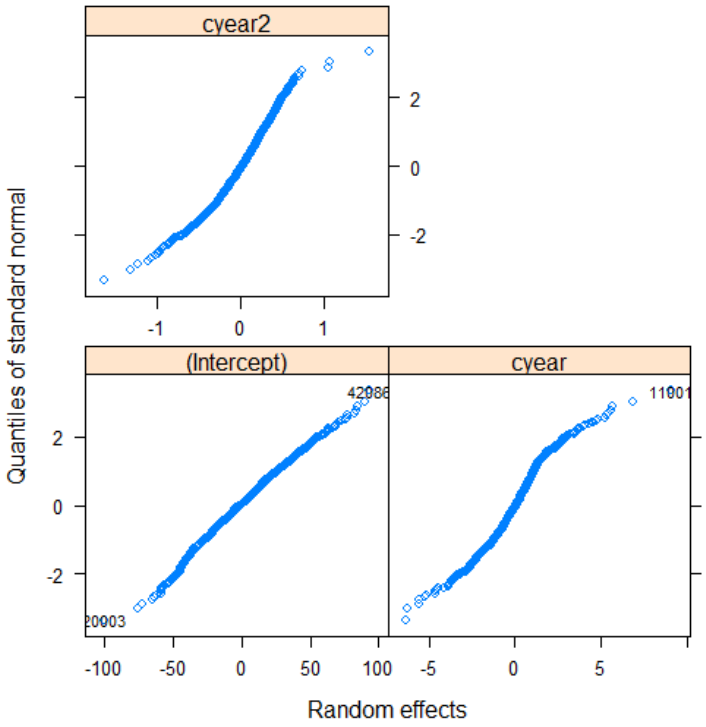
3.3 Model Assumption Diagnostics

For linear mixed effects models, the assumption of normality needs to be assessed by looking at residual errors. In the study, marginal residual which is a deviation of a subject from the group mean needs to be checked [Gurka et al, 2006]. Cheng et al. (2000) recommended using jackknifed studentized residual histograms, and scatter plots of jackknifed studentized residuals versus predicted values over time to help the assessment of normality.

3.3.1 Diagnostics for Random Effects Assumption

Figure 3.4 displays the normal plots of the estimated random effects (random intercept, random time, and random squared time effects). It is plausible to say that these random effects follow the assumption of normal distribution, although some outliers were also identified such as patients 11001, 20003 and 42086. CHL values of patient 20003 in the ten DCCT years were between 64 to 85, which are at the lower bound of overall CHL values, while for patient 42086 which were measured for four years, its CHL values are at the upper bound of overall CHL

Figure 3.4: *Random effect assumption assessment*



values. For patient 11001 which were measured ten years, its CHL value started from 165 in first year, then finished at 268 in the last DCCT year, which increased 62% in ten years.

3.3.2 Diagnostics for Residual Errors Assumption

The studentized residual histogram for each year were plotted in order to see whether the model meets the normal distribution assumption. Figure 3.5 contains the jackknife studentized residual histogram for only DCCT year 0 and 1, which indicate the normality assumption is valid. Figure 3.6 displays the predicted values versus jackknife studentized residuals for DCCT year 0 and 1, which also indicate the normality assumption is valid. The other years residual plots are attached in Appendix.

In conclusion, we proposed a linear mixed-effects model with fixed effects (time, squared

Figure 3.5: *Studentized residual histograms for year 0 and year 1*

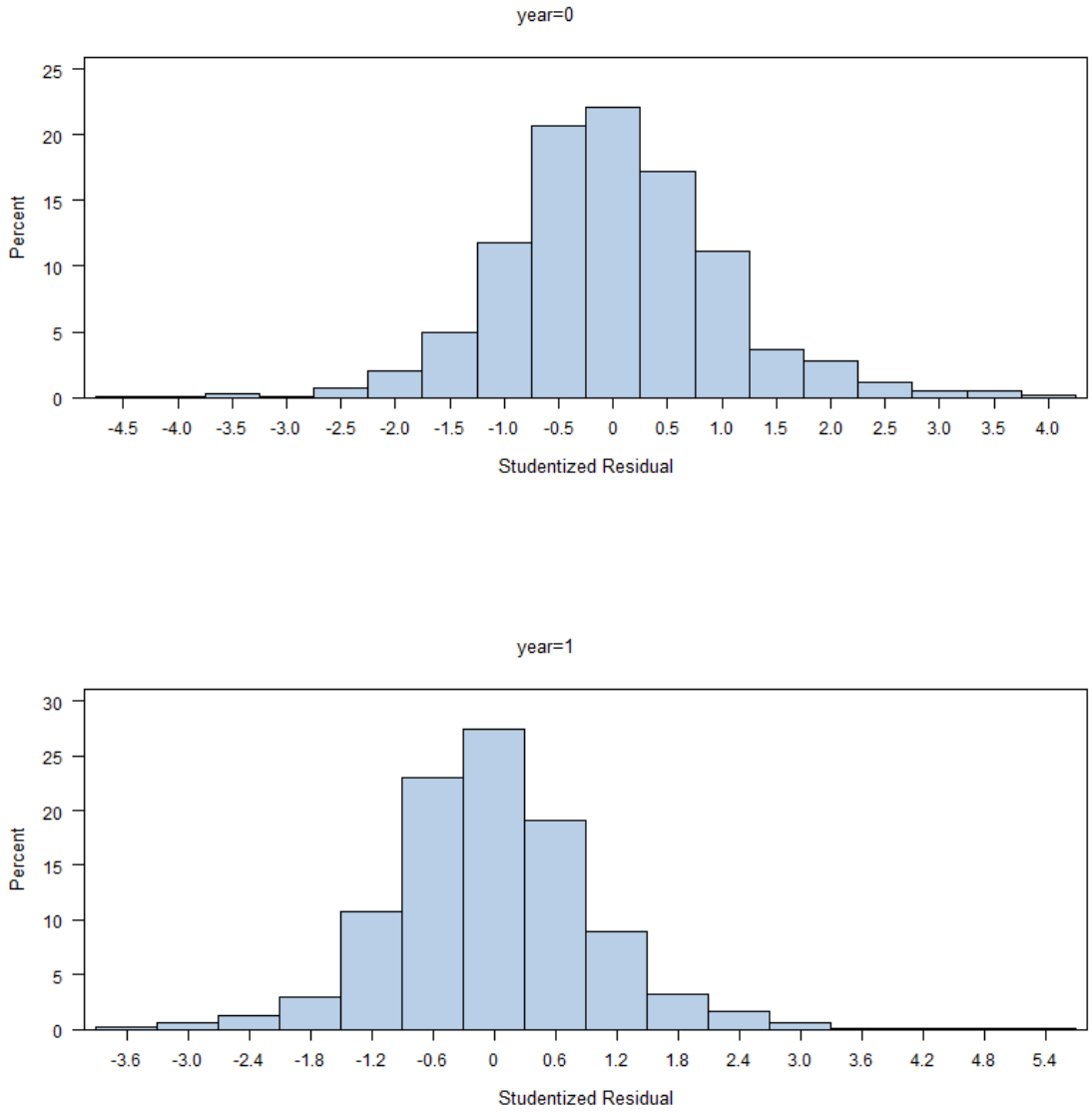
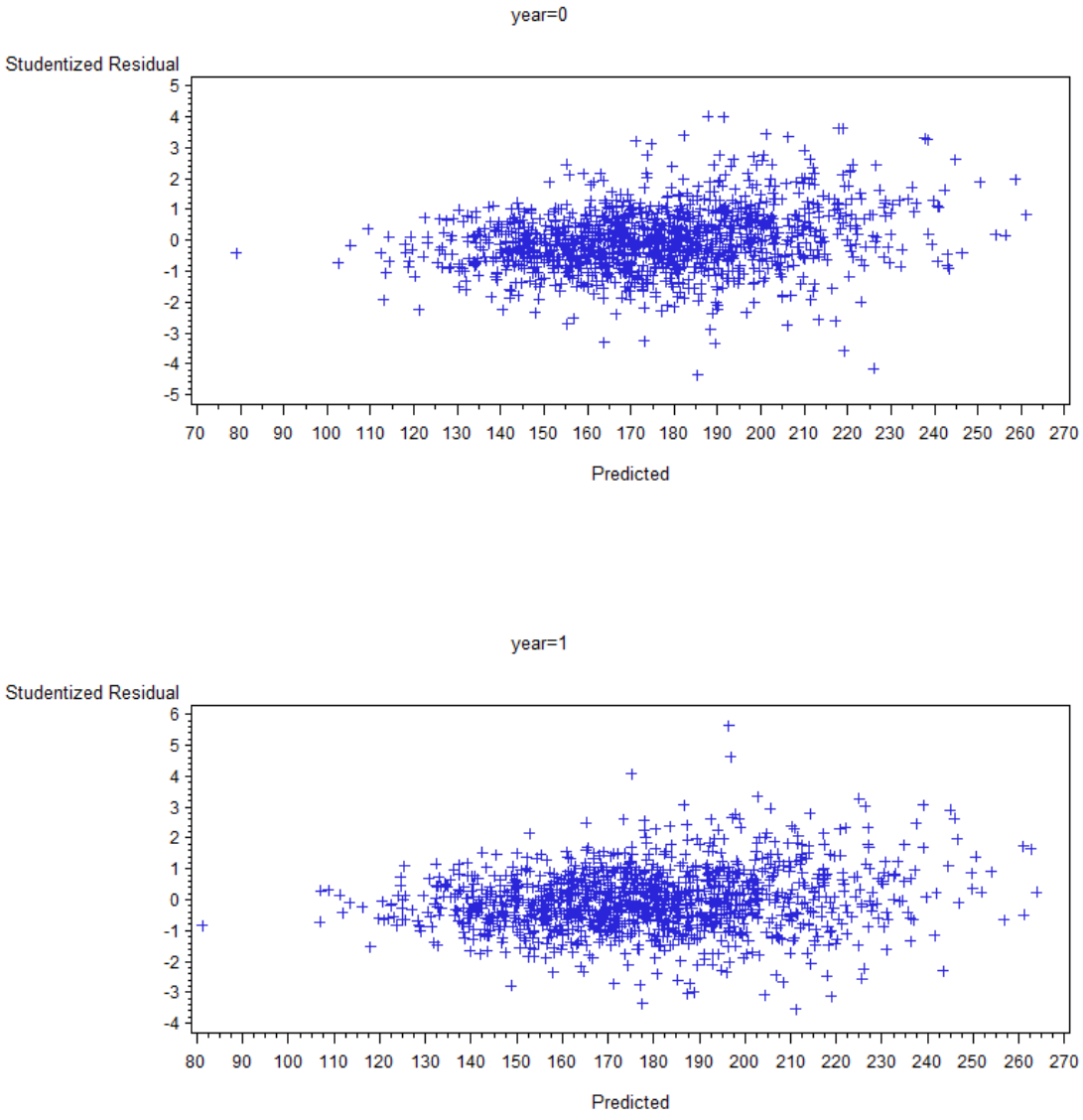


Figure 3.6: *Predicted value versus studentized residual plot for year 0 and year 1*



time, treatment, cohort, gender, baseline indicator, duration prior to DCCT, age of diagnosis), and random effects (random intercept, random time, random squared time) along with the specific residual correlation structure (AR(1)) to sufficiently catch the pattern and characteristics of the lipid measures of DCCT.

Chapter 4

Lipid Genome Wide Association Study (GWAS)

Abnormal lipid levels are the important risk factors of heart disease and nephropathy which are long-term complications of T1D. T1D Sibling patients cross-sectional studies showed genetic factors may be contribute to specific complications. Recently, specific genetic loci which are associated with differences in lipoprotein have been identified, and they are also associated with complications [Freeman (1994), Zannis (1981)]. In this Chapter, LME model with genetics data will be used to identify novel loci which are highly associated with lipid measures.

4.1 Some Preparation for GWAS Analysis

Genotyping was performed using the Illumina 1M beadchip assay. Filtering genotype data should be undertaken before GWAS analysis. We removed SNPs with a minor allele frequency (MAF) <1% and removed SNPs that failed a test of Hardy-Weinberg Equilibrium (HWE).

HWE theorem states both allele and genotype frequencies in a population remain constant, so testing for HWE is maybe the most common quality control procedure in all human genetics [Cutler & Abecasis, 2010]. After filtering, 841,000 autosomal SNPs were analyzed in the present study. In GWAS association study, we added each SNP individually as a fixed effect covariate to the model from Chapter 3 for CHL and LDL (LDL Primary GWAS results are shown in Appendix C). The other two lipid measures GWAS will be done in the near future.

Because of the huge dataset, advanced information technology support, the limitation of the system compatibility and memory size of personal computer, we used the high performance computer cluster at the Hospital for Sick Children in Toronto. Since the number of SNPs analyzed is quite large, we divided it into pieces so that we can submit each subset to the cluster. Specifically, we cut data set into 434 batches which contain approximately 2000 SNPs each. We used PLINK version 1.07 , an open-source C/C++ GWAS tool set [Purcell et al, 2007 a,b] with an R plug in (details of R code in appendix B), in order to obtain the analysis results of our LME with four lipid measures.

At the conventional P$<$ 0.05 level of significance, an association study of 1 million SNPs will show 50 000 SNPs to be "associated" with the phenotype, almost all of which will be false-positive. The most common method of dealing with this problem is to reduce the false-positive rate by applying the Bonferroni correction, in which the conventional p-value is divided by the number of tests performed [Pearson & Manolio (2008, Yang et al., (2005)]. 1 million SNP survey would use a threshold of $P < 0.05/10^6$, or 5×10^{-8}, which is the genome-wide significant threshold.

4.2 Total Cholesterol (CHL) GWAS

In a genome-wide association study, while we hope for some true associations, nearly all the SNPs will not have any association with the outcome, so almost all the p-values should come from a uniform distribution.

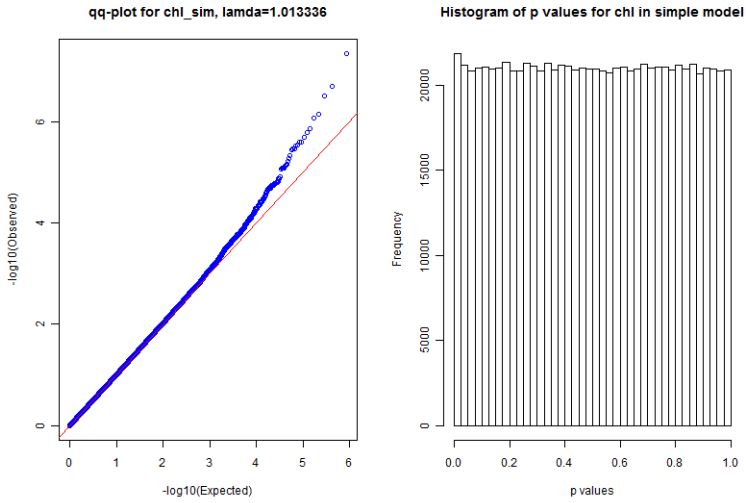
4.2.1 QQ Plot and Histogram

Once GWAS analysis is complete, we need to look at a histogram of p-values to assess the distribution, and a quantile-quantile (QQ) plot resulting from each lipid measure. QQ plot is a graphical diagnostic tool that is used to compare the observed probability distributions of the data to the assumed to assess if the assumptions are valid [Wilk & Gnanadesikan, 1968]. In this study, the null hypothesis is set up as "there is no association between the SNP and lipid measures", i.e. $\boldsymbol{\beta}_{SNP} = 0$. We plotted the observed $-\log_{10}$ p-values along the Y-axis for every SNP tested versus the expected $-\log_{10}$ p-values under the null hypothesis on the X-axis, which makes it easier to focus on the very low p-values [Pearson & Manolio, 2008]. So, if the null hypothesis is true for every single case, the result of a QQ plot will be close to the line $y = x$. Deviations from this line in the upper tail indicate SNPs are smaller than expected p-values by chance.

A numerical summary of the departure from the uniform distribution is so-called "genomic control coefficient" λ [Devin and Roeder, 1999]. Let $\hat{\boldsymbol{\beta}}$ and SE represent the coefficient estimate and the corresponding standard error, then

$$\hat{\lambda} = \frac{\text{Median}[(\frac{\hat{\beta}}{SE})^2]}{0.4549} \tag{4.1}$$

Figure 4.1: *QQ plot and Histogram of CHL primary GWAS*



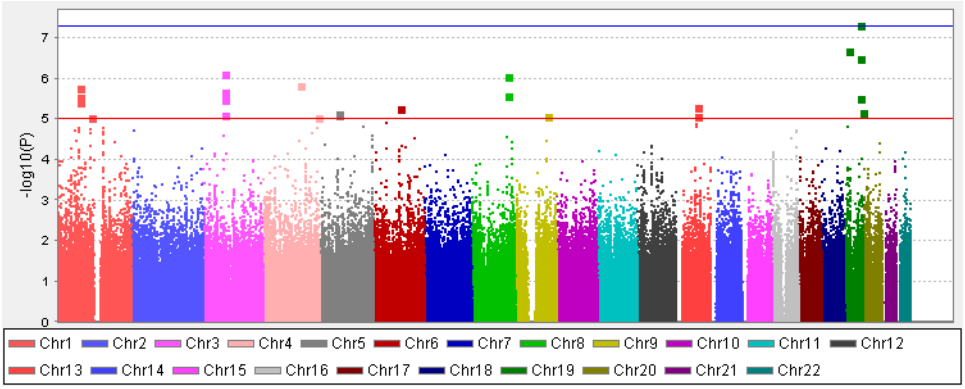
assesses whether the square of the test statistic follow a chi-square distribution with the degree of freedom 1, which has the median equal 0.4549.

Figure 4.1 show the histogram of p-values of SNPs and the QQ plot as well. The right plot of Figure 4.1 clearly shows that the distribution of p-values are very close to the specified uniform distribution, which indicates most of the SNPs have no deviation from the null hypothesis. The QQ plot (the left of Figure 4.1) shows the observed p-values resulting from CHL primary GWAS model versus the expected uniform p-values with $\lambda = 1.013336$. We can see that SNPs p-values do follow a uniform distribution.

4.2.2 Manhattan Plot

We used Haploview version 4.2, which is a Java based tool for the Manhattan plot [Barrett, 2007]. The Manhattan plot represents the significance of the association between a SNP and the trait being measured. The different colors from left to right display the different chromosomes within the range of 1 to 22. The X-axis shows the SNPs ordered by the physical position within

Figure 4.2: *Manhattan Plot of CHL primary GWAS*



chromosome while the Y-axis displays $-\log_{10}$ transformed p-values, which represent the degree of association.

In our Manhattan plot, there are two important lines in the plot, the red line with Y-value is approximately 7.3, which was calculated with $-\log 10(p = 5 \times 10^{-8}) = 7.3$, and the blue line with Y-value is 5 shows the suggestive significant threshold SNPs, and the reason we used the blue line is to find more potential suggestive SNPs which could be associated with lipid measures.

Figure 4.2 illustrates the Manhattan plot of total cholesterol (CHL) GWAS model. At chromosome 19, the highest dot (rs7412) represents the SNP whose p-value is the smallest, and it reaches to the genome-wide significant line ($p = 4.55 \times 10^{-8}$).

4.2.3 Top SNPs of CHL Primary GWAS

Table 4.1 describes the top 20 SNPs whose p-values are smaller than the suggestive cut off value 10^{-5}. Among the SNPs in the table, only SNP rs7412 is genome-wide significant associated with CHL , which is near APOE gene at 19q13.2.

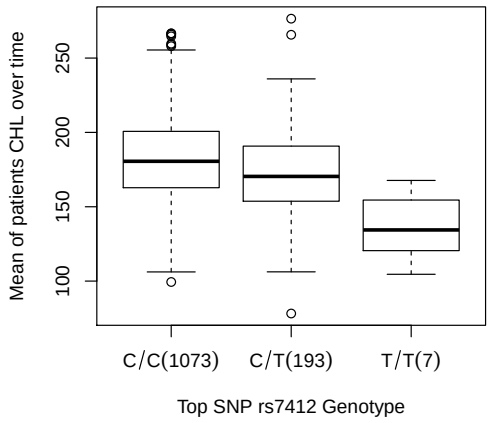
Table 4.1: *Top SNPs of CHL Primary GWAS*

SNP	CHR	BP	A1	A2	MAF	$\hat{\beta}$	Se	STAT	P
rs7412	19	50103919	T	C	0.0813	-11.0509	2.00863	-5.50173	4.55E-08
rs6511720	19	11063306	T	G	0.1018	-9.42115	1.80197	-5.22824	1.99E-07
rs11083751	19	50105073	T	G	0.0791	-10.3337	2.00774	-5.14694	3.06E-07
rs2164883	3	69094955	C	A	0.1222	7.95934	1.59736	4.98282	7.11E-07
rs1995222	8	118298943	A	G	0.4646	5.35298	1.08234	4.94576	8.59E-07
rs11098654	4	123128865	T	C	0.0695	10.4887	2.16116	4.85328	1.36E-06
rs2801328	1	72188872	G	A	0.4349	5.24827	1.08927	4.81814	1.62E-06
rs6787753	3	69078627	A	C	0.1141	7.77633	1.63047	4.76938	2.06E-06
rs7005140	8	118303328	G	A	0.4693	5.12008	1.08243	4.73015	2.49E-06
rs7544722	1	72179997	T	C	0.434	5.15637	1.0915	4.7241	2.57E-06
rs4420638	19	50114786	G	A	0.1669	7.02865	1.49626	4.69747	2.92E-06
rs6785239	3	69084092	G	A	0.116	7.63624	1.62712	4.6931	2.98E-06
rs2768395	1	72174618	A	G	0.4351	5.07898	1.08925	4.66282	3.44E-06
rs1016126	1	72208991	G	A	0.4351	5.07898	1.08925	4.66282	3.44E-06
rs1426173	1	72157184	A	G	0.4343	5.07519	1.09019	4.65534	3.57E-06
rs9571417	13	64992897	C	T	0.1256	7.52243	1.63548	4.59952	4.65E-06
rs994732	6	87234971	T	C	0.3132	5.40259	1.18159	4.57231	5.29E-06

4.2.4 Genotype of Top SNP in CHL Primary GWAS

Figure 4.3 illustrates the Box and Whisker plots of the top SNP rs7412 in CHL primary GWAS. The Y-axis shows the mean of winsorized CHL over time for each patient, while the X-axis shows the minor and major allele and the count numbers of them. The SNP rs7412 is in APOE gene, located at 19q13.32. The minor allele is noted by T, and major allele is represented by C. The minor allele frequency is 0.0813. 1073 patients have major homozygote C/C, 193 patients have heterozygote C/T, and 7 patients have minor homozygote T/T. There are 30 patients with missing genotypes. The rare homozygote T/T genotype is associated with lower mean CHL level.

Figure 4.3: *Box and Whisker Plot of Top SNPs rs7412 in CHL primary GWAS*



4.2.5 Region Plot of Top SNP in CHL Primary GWAS

Genome-wide associations studies frequently identify associations with many highly correlated SNPs in a chromosomal region, due in part to linkage disequilibrium (LD) among SNPs. LD

occurs when genotypes at the two loci are not independent of each other.

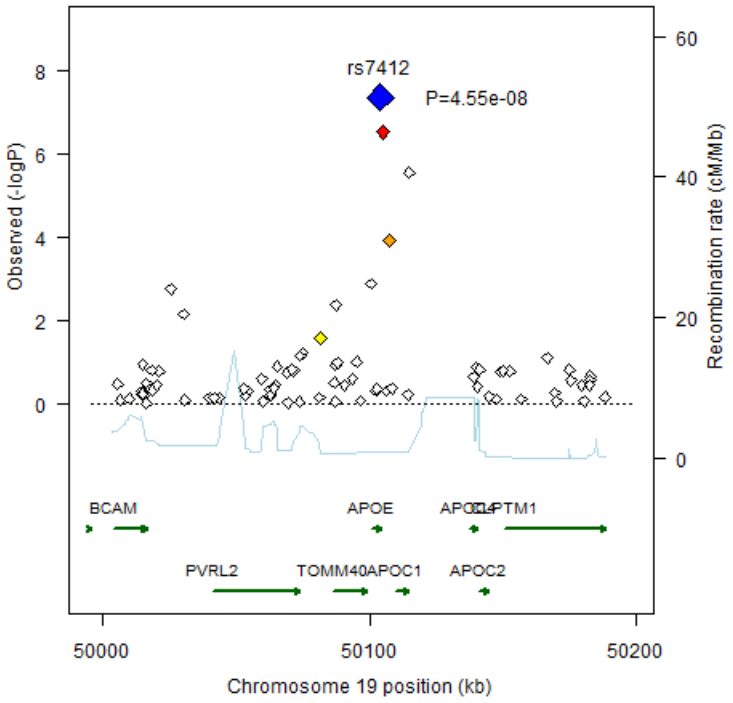
In a region plot [Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, 2007], there are two Y-axis, the left side axis shows the observed $-\log 10$ transformed p-value resulting from the lipid GWAS, while the right side axis displays the recombination rate (cM/Mb), estimated using HapMap 2 Utah Resident with Northern and Western European Ancestry (CEU) samples. Genetic recombination is the breaking and rejoining of DNA strands to form new molecules, and the recombination rate is the number of observed recombination events divided by the total number of events.

The X-axis in the plot illustrates the chromosome physical positions which are based on build 35 (released on May 2004) of National Center for Biotechnology Information (NCBI) of the human genome. Pairwise LD estimates between SNPs (measured as r^2) are calculated with our data set using PLINK. Estimated recombination rates are plotted to reflect the LD structure around the most significant SNP. In the plot, the bright red indicates the SNPs that are highly correlated, while white indicating weakly correlated. Also the dark blue diamond-shaped points represent the index of SNPs, as well as white color indicates $r^2 < 0.2$, gray blue indicates $0.2 < r^2 < 0.5$, orange color indicates $0.5 < r^2 < 0.8$, and red color indicates $r^2 > 0.8$. The bottom panel of the region plot displays the name and location of genes in the University of California Santa Cruz Genome Browser (UCSC), which can be found at http://genome.ucsc.edu [Kent et al., 2002].

Figure 4.4 is a region plot in the APOE locus group at a 100-kb region surrounding the SNP rs7412 on chromosome 19. The dark blue diamond-shaped data points represent $-\log_{10} p$-

value of the top SNP in CHL simple GWAS, rs7412 ($p = 4.55 \times 10^{-8}$). One SNP rs11083751 ($p = 3.06 \times 10^{-7}$, BP: 50105073) which is highly correlated with index SNP rs7412.

Figure 4.4: *Region plot of rs7412 in CHL simple GWAS.*



Chapter 5

Discussion and Future Work

5.1 Discussion of the Thesis

In this thesis, we proposed both non-genetic and genetic analysis of lipid measure (CHL) for 1303 T1D patients in the study of DCCT. We considered a LME model with AR(1) covariance structure to fit each lipid measure, then added SNPs as covariates to fit each lipid GWAS in three specified models. The purpose of the study was to obtain the SNPs which are highly associated with four lipid levels in patients with T1D.

5.1.1 Non-genetics Results

In the non-genetic analysis, we performed primary model, in which linear time, quadratic time, gender, cohort, treatment, duration of diabetes, age of diagnosis, baseline indicator and baseline-treatment interaction as the fixed effects, intercept, linear time, quadratic time as random effects, and AR(1) residual correlation structure. We added some time-varying covariates

(BMI, insulin dose, HbA1c, and exercise) and one non-varying covariate (C-peptide) to primary model as our complex model. After testing the variance component and comparing correlation structure, then we still used intercept, time, time square as random effects, and with AR(1) correlation structure. We emphasized to see whether and which fixed effects are significant associated with CHL when more information (BMI, Insulin, Cpep, exercise) added in (Table 5.1). When other covariates were fixed in the model, time is not significant for CHL. BMI, Insulin dose, and HbA1c are highly significant associated with CHL when other covariates are in the model. When we fixed other fixed effects, the interaction of treatment and baseline indicator is not significant for CHL, i.e. there is no significant difference of intensive treatment with baseline year and after. C-peptide was treated as categorical and three levels factor. It was not significantly associated with response variable, but C-peptide was kept in the complex model because it measures insulin production and is correlated with HbA1c.

5.1.2 Genetics Results

In the genetic analysis, we performed a genome-wide association study to identify the SNPs which are significantly associated with CHL and LDL. We compared the three different linear mixed effect models with SNPs added as a fixed effect. We note that rs7412 is not only the top significant SNP in CHL primary GWAS and complex GWAS, but also the top significant SNP of LDL simple GWAS and complex GWAS. Rs7412 is located at 19q13.2, in APOE gene which is in a cluster with APOC1 and APOC2. APOE, APOC1, and APOC2 were reported to be highly associated with CHL and LDL by Zannis VL et al, (1981) and highly associated with LDL cholesterol by Burkhardt et al, (2008), Kathiresan et al,(2008) and Willer et al, (2008).

Table 5.1: *Total cholesterol (CHL) analysis in Complex Model*

	$\hat{\beta}$	Stand.Error	DF	t-value	p-value
(Intercept)	60.58	5.07	7693	11.96	1.16E-32
Centered year	-0.24	0.15	7693	-1.53	1.25E-01
(Centered year)^2	-0.06	0.05	7693	-1.15	2.50E-01
Intensive Treatment	0.17	1.57	1295	0.11	9.13E-01
Baseline	-3.27	1.05	7693	-3.10	1.94E-03
Second Cohort	0.88	2.20	1295	0.40	6.89E-01
Female	7.85	1.50	1295	5.25	1.80E-07
Centered age of diag	0.93	0.11	1295	8.65	1.47E-17
Diabetes duration	0.11	0.03	1295	4.17	3.32E-05
BMI	2.67	0.15	7693	18.00	5.18E-71
Insulin	6.98	1.60	7693	4.38	1.23E-05
C-peptide(0.2, 1]	0.99	2.16	1295	0.46	6.47E-01
C-peptide(0, 0.03]	-1.34	1.86	1295	-0.72	4.71E-01
Exercise	0.27	0.32	7693	0.84	4.04E-01
HbA1c	4.47	0.23	7693	19.86	1.29E-85
Intensive trx:baseline	0.37	1.25	7693	0.29	7.69E-01

Rs2164883 is a suggestive SNP in CHL Primary ($P = 7.11 \times 10^{-7}$)GWAS. It is located at 3p14.1 in chromosome 3, and is upstream of C3orf64 gene and downstream of FAM19A4. This region was reported by Cho et al. (2011) to be involved in fasting glucose levels for T2D east Asian populations. This SNP seems to be a novel locus for association with CHL of patients with T1D.

Rs10866235 is a suggestive SNP in LDL primary ($P = 1.15 \times 10^{-7}$) GWAS, the P value is very close to geno-wide significant threshold. It shows highly significant with LDL level, but according to our knowledge, no literature has ever proved the association with this SNP and LDL, so rs10866235 seems a promising and novel SNP.

5.2 Future Work

I will shortly complete the primary model GWAS with HDL and TRG, and complex model GWAS study with four lipid measures.

In GWAS analysis, we discussed the SNPs from chromosome 1 to chromosome 22. There are four more chromosome in our dataset, they are X chromosome, Y chromosome, XY chromosome which is pseudo-autosomal region of chromosome X, and chromosome MT (Mitochondrial). We will analyze SNPs on the X chromosome separately by the gender and on the Y chromosome for males only. We already discussed the association with gender and lipid measures, then we are going to identify top SNPs in sex chromosome that are genome-wide associated with lipid measures.

When DCCT closeout in 1993, all subjects were encouraged to adopt intensive treatment and 93% of them participated in the Epidemiology of Diabetes Interventions and Complications

(EDIC) study which is still ongoing. The top SNPs which we found in DCCT will be examined for association with lipid measures during the study of EDIC. To avoid false positive results of our lipid GWAS in DCCT, we are going to replicate our results in two other cohorts of T1D (Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR) and Genetics of Kidneys in Diabetes study (GoKinD)). We will test our top SNPs in simple GWAS of each lipid measures with these two studies, in order to confirm whether the associated SNPs randomly in DCCT study or have the prevalent association with four lipid measures in all population with T1D. WESDR measures CHL and HDL each year except for baseline, while LDL and TRG were measured at visit 5. GoKind study also measures CHL and HDL at the same lab used by DCCT/EDIC, so we want to examine whether the top SNPs in DCCT simple CHL and HDL GWAS are associated with CHL and HDL in GoKind.

Furthermore, we are going to discuss genotype imputation with missing SNPs and ungenotyped SNPs in Lipid GWAS [Jeffrey et al., 1997]. We used Illumina 1M beadchip to obtain the genotype data, and missing SNP data is common in GWAS, sometimes with rates as high as $5 - 10\%$ [Dai et al, 2006]. Imputation uses the correlation between markers present in the reference sample for making predictions of genotypes present in an experimental sample [Jeffrey et al, 2009].

Different dietary habit and usage can affect the lipid levels of T1D patients, such as calories, alcohol or sugar in the body are converted into TRG and stored in fat cells throughout the body. Several dietary measures were collected in our DCCT data set, they are calories intake, alcohol and smoking consumption. We will examine whether these covariates have significant effect on four lipid measures, and we will take GWAS analysis for identify loci which have

association between lipid measures and diet-SNP interactions.

GWAS has successfully identified some genetic risk factors for diseases, such as diabetes, cardiac diseases, Alzheimer's disease etc, but it has several limitations which need to be addressed in the future. The false-positive results, insufficient information on gene function, and small sample sizes, are the major limitations of GWAS [Pearson & Manolio, 2008]. People are still working on the environment exposures and other non-genetic risk factors on GWAS.

Appendix A

Primary LME model with AR(1) correlation structure

```
> attach(lipid.data)

> library(nlme)

> sim.model.chl <- lme(chl.win ~ cyear+cyear2+factor(OBSEX)+ factor(cohort)

+ +factor(trx)*bline

+ + DURATION + cagediag,

+ random = ~ 1+cyear+cyear2|PATIENT,na.action=na.omit,corr=corAR1())
```

Appendix B

Some Plots Code in GWAS

```
# 1. R Plug in with PLINK

cd /adp/home/taowang/dbp_analysis

export PATH=$PATH:/tools/R/2.10.0/bin/:/tools/plink/current/

R CMD /tools/R/2.10.0/lib64/R/bin/Rserve --vanilla

plink --bfile ill1M_filter --remove Removesample.txt --extract /adp/home/taowang/cutfile/


# 2. calculate LD in plink

# calculate LD using the data from all patient

plink --bfile /adp/home/taowang/ill1M_filter --r2 --ld-snp rs7412 --ld-window-kb

1000 --ld-window 99999 --ld-window-r2 0


# 3. Data management in R -- combine the file with LD and the file with P value

# import data
```

```
ldl_7412<-read.table("rs7412.ld",head=T)

# take other snps and r2, so drop CHR_A BP_A SNP_A,

bp_A <- ldl_7412$BP_A[1]

chr_A <- ldl_7412$CHR_A[1]

SNP_A <- as.character(ldl_7412$SNP_A[1])

ldl_7412<-ldl_7412[,c(4:7)]

names(ldl_7412)<-c("CHR","BP","SNP","R2")

# import data p value

ldl_reg_p<-read.table("ldl_sim_p.txt",head=T)

# sort by bp first for two data

merge<-merge(ldl_reg_p,ldl_7412,by.x="SNP",by.y="SNP")

merge.1<-merge[,c(1,2,3,4,7)]

names(merge.1)<-c("SNP","CHR","POS","PVAL","RSQR")

p_A <- merge.1$PVAL[merge.1$SNP==SNP_A]

m<-which(abs(merge.1$POS-bp_A)<=2.5e5)

ldl<-merge.1[m,-2]

names(ldl)

TYPE <- rep("typed", length(m))

ldl <- cbind(ldl, TYPE)

row.names(ldl) <- ldl$SNP

ldl <- ldl[,c("POS","PVAL","TYPE","RSQR")]

#4. in Regionplot.RData to draw region plot
```

```
pdf("assocplot_ldl_rs7412.pdf", width=6, height=4)

make.fancy.locus.plot("rs7412", "", "19", ldl, ceiling(-log10(p_A)+1), p_A)

dev.off()

################################

png(file="region_plot_rs7412_ldl.png")

make.fancy.locus.plot("rs7412", "", "19", ldl, ceiling(-log10(p_A)+1), p_A)

dev.off()
```

Appendix C

LDL Primary GWAS Results

Table C.1: *Top SNPs of LDL Primary GWAS*

SNP	CHR	BP	A1	$\hat{\beta}$	Se	DF	STAT	P
19	rs7412	50103919	T	-14.62	1.78	1267	-8.22	5.06E-16
19	rs11083751	50105073	T	-14.21	1.78	1283	-8.00	2.79E-15
19	rs445925	50107480	T	-9.83	1.53	1297	-6.44	1.69E-10
19	rs6511720	11063306	T	-9.56	1.61	1296	-5.92	4.02E-09
4	rs10866235	181472674	T	-5.29	0.99	1295	-5.33	1.15E-07
19	rs4420638	50114786	G	7.10	1.34	1267	5.31	1.31E-07
1	rs7528419	109618715	G	-6.10	1.19	1297	-5.11	3.70E-07
1	rs2801328	72188872	G	4.96	0.98	1293	5.07	4.63E-07
4	rs13106873	181472150	A	-5.04	1.00	1296	-5.06	4.90E-07
1	rs7544722	72179997	T	4.88	0.98	1281	4.98	7.21E-07
1	rs7527501	72164045	T	6.20	1.25	1297	4.97	7.43E-07
1	rs1426173	72157184	A	4.84	0.98	1295	4.95	8.41E-07
1	rs3856029	72171311	C	6.17	1.25	1296	4.95	8.42E-07
1	rs2768395	72174618	A	4.83	0.98	1296	4.94	8.86E-07
1	rs1016126	72208991	G	4.83	0.98	1296	4.94	8.86E-07
1	rs11209872	72176049	C	6.13	1.25	1297	4.92	9.86E-07

Bibliography

Atkinson AC (1985). *Plots, Transformations, and Regression*. Clarendon Press: Oxford.

Barrett JC, et al. (2005). Haploview: Analysis and Visualization of LD and Haplotype Maps. *Bioinformatics*, **vol 21**, pages 263-265.

Box G, Jenkins GM, and Reinsel G (1994). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Burkhardt R, et al. (2008). Common SNPs in HMGCR in Micronesians and Whites Associated with LDL-cholesterol Levels affect alternative splicing of exon 13. *Arteriosclerosis, Thrombosis, and Vascular Biology*, **vol 18**, pages 2078-2084.

Burton PR, et al. (2007). Genome-Wide Association Study of 14000 Cases of Seven Common Diseases and 3,000 Shared Controls. *Nature*, **vol 447**, pages 661-678.

Canadian Diabetes Association (2012). Canadian Diabetes Association - 2011 Annual Report. http://www.diabetes.ca, Accessed on Date March 10, 2012.

Cho YS, et al. (2011). Meta-Analysis of Genome-Wide Association Studies Identifies Eight New Loci for Type 2 Diabetes in East Asians. *Nature Genetics*, **vol 44**, pages 67-72.

Cressie N (1993). Statistics For Spatial Data. Wiley, New York.

Cheng J, et al.(2010). Real longitudinal Data Analysis for Real People: Building a Good Enough Mixed Model. *Statistics in Medicine*. **vol 29**, pages 504-520.

Culter DJ. and Abecasis GR. (2010) Response to Graffelman: Test of Hardy-Weiberg Equilibrium. *The American Journal of Human Genetics*, **vol 86**, pages 818-819.

Davidian M.,Vock DM., Tsiatis AA., and Muir AJ. (2011). Mixed model analysis of censored longitudinal data with flexible random-effects density. *Biostatistics* **vol13**, pages 61-73.

dbSNP production Team (2011). National Center for Biotechnology Information (NCBI). Released at October 13,2011.

The Diabetes Control and Complications Trial Research Group (1995). Implementation of Treatment Protocols in the Diabetes Control and Complications Trial. *Diabetes Care*, **vol 18**, pages 361-376.

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, **vol 316**, pages 1331-1336.

Devlin B and Roeder K (1999). Genomic Control for Association Studies. *Biometrics*, **vol 55**, pages 997-1004.

Fox J.(2008). Applied Regression Analysis and Generalized Linear Model. SAGE Publications, Inc.

Frees EW (2004). *Longitudinal and Panel Data-Analysis and Applications in the Social Sciences*. Cambridge University Press.

Friedewald WT, Levy RJ, and Fredrickson DS (1972). Estimation of the Concentration of Low Density Lipoprotein Cholesterol in Plasma Without Use of the Prepative Ultra-Centrifuge. *Clinical Chemistry and Laboratory Medicine*, **vol 18**, pages 499-502.

Gurka MJ and Edwards LJ (2008). Mixed Models Handbook of Statistics.*Epidemiology and Medical Statistics*, **vol 27**, pages 253-280.

Harville DA, (1974). Bayesian Inference for Variance Components Using Only Error Contrasts. *Biometrika*, **vol 61**, pages 383-385.

Harville DA, (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of Ameriacan Statistical Association*, **vol 72**, pages 320-338.

Hedeker D and Gibbons RD (2006). *Longitudinal Data Analysis*. Wiley.

Jeffrey BC, et al. (2010). Genotype Imputation for Genome-Wide Association Studies. *Nature Reviews Genetics* **vol 11**, pages 499-511.

Juvenile Diabetes Research Foundation Annual Report. (2011).

Kathiresan S, et al. (2008). Common Variants at 30 Loci Contribute to Polygenic Dyslipidemia.*Nature Genetics*, **vol 41**, pages 56-61.

Kent WJ, et al. (2002). The Human Genome Browser at UCSC. *Genome Research*, **vol 12**, pages 996-1006.

Kleinbaum DG et al. (1998). *Applied Regression Analysis and Multivariable Methods.* Duxbury Press.

Little TD, Schnabel KU, and Baumert J (2000). *Modeling Longitudinal and Multilevel Data.* Lawrence Erlbaum Associates.

Laird NM and Ware JH (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, **vol 38**, pages 963-974.

Lodish H., et al. (2000). *Molecular Cell Biology. 4th edition.* New york: W.H.Freeman.

Pagon R et al. (1993). Gene Reviews. University of Washington, Seattle.

Pearson TA and Manolio TA (2008). How to Interpret a Genome-Wide Association Study. *Journal of American Statistical Association*, **299**, pages 1335-1344.

Pinheiro JC and Bates DM(2000). *Mixed-Effects Models in S and S-PLUS.* Springer Verlag New York.

Pinheiro JC, et al.(2011). Linear and Nonlinear Mixed Effects Models. *R help nlme package.*

Purcell S, et al. (2007 a). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Anlyses. *The American Journal of Human Genetics*, **vol 81**, pages 559-575.

Purcell S, et al.(2007 b). PLINK Online Tutorial: Whole Genome Association Analysis Toolset. Version 1.07. http://pngu.mgh.harvard.edu/ purcell/plink/, Accessed on Data July 10, 2011.

Rahm E. (2000) Data Cleaning: Problems and Current Approaches, *Data Engineering Bulletin*. **vol 23**, pages 3-13.

Sakamoto Y, Ishiguro M, and Kitagawa G (1986). *Akaike Information Criterion Statistics*. D. Reidel Publishing Company.

Schwarz G (1978). Estimating the Dimension of Model. *The Annuals of Statistics*, **vol 6**, pages 461-464.

Schallehn E, Sattler K U, and Saake G (2001). Advanced Grouping and Aggregation For Data Integration. *Tenth International Conference on Information and Knowledge Management*. Pages 547-549.

Sheskin D (2003). *Handbook of parametric and Nonparametric Statistical Procedures*. Chapman and Hall/CRC.

Singer DE, et al. (1992). Association of HbA1c with Prevalent Cardiovascular Disease in the Original Cohort of the Framingham Heart Study. *Diabetes*, **vol 41**, pages 202-208.

Stinnett SS (1993). *Collinearity in Mixed Models*. PhD dissertation. University of North Carolina.

Stram DO and Lee JW (1994). Variance Components Testing in the Longitudinal Mixed Effects Model. *Biometrics*, **vol 50**, pages 1171-1177.

Syvanen AC (2005). Toward Genome-wide SNP genotyping. *Nature Genetics*, **vol S**, pages 5-10.

University of California, Los Angeles. Regression Diagnostics. Stata Web Url Books. Accessed on Date July 13, 2012.

Verbeke G, and Molenberghs G (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach*. Springer-Verlag, New York Berlin Heidelberg.

Verbeke G, and Molenberghs G (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics, New York: Springer.

Wilk MB and Gnanadesikan B (1968). Probability Plotting Methods for the Analysis of Data *Biometrika*, **vol 55**, pages 1-17.

Willer CJ, et al. (2008). Newly identified Loci that Influence Lipid Concentrations and Risk of Coronary Artery Disease. *Nature Genetics*, **vol 40**, pages 161-169.

Yang Q, et al. (2005). Power and Type I Error Rate of False Discovery Rate Approaches in Genome-Wide Association Studies. *BMC Medical Genetics*, **vol 6**, Supplement 134.

Zannis VL, Just PW and Breslow JL.(1981). Human Apoliprotein E Subleases Are Genetically Determined. *American Journal of Human Genetics.* **vol 33**, pages 11-24.