STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:3345–3357

Familial associations of lipid profiles: a generalized estimating equations approach[‡]

Andreas Ziegler[1,*,†], Christian Kastner[2], Daniel Brunner[3] and Maria Blettner[4]

[1] *Medical Centre for Methodology and Health Research, Institute of Medical Biometry and Epidemiology, Philipps-University of Marburg, Germany*
[2] *Institute of Statistics, Ludwig-Maximilians University of Munich, Germany*
[3] *Institute of Physical Hygiene, Wolfson Medical Center, Holon and Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel*
[4] *Department of Epidemiology and Medical Statistics, University of Bielefeld, Germany*

SUMMARY

Elevated plasma levels of apolipoproteins A1 (apoA1) and B (apoB) are important protective factors and risk factors, respectively, for atherosclerosis and coronary heart disease. It is well known that both apoA1 and apoB reveal strong familial aggregation. Our goal was to investigate whether exogenous variables influence these associations. We used marginal regression models for the mean and association structure (generalized estimating equations 2; GEE2) to analyse data from 1435 family members within 469 families of different sizes included in the Donolo-Tel Aviv Three-Generation Offspring Study. The usual robust variance matrix was approximated by extensions of jack-knife estimators of variance to GEE2 models. Estimation of standard errors in models with quite complex correlation structures was possible using this approach. All analyses were easily carried out using a menu-driven stand-alone software tool for marginal regression modelling. We demonstrate that a variety of hypotheses can be tested using Wald statistics by modelling regression matrices for the association structure. We show that correlation for apoB between parent-offspring pairs increased with decreasing age difference and that pairs with individuals of the same gender had more similar apoA1 levels than individuals of different gender. Associations between different relative pairs did not all agree with those expected from differences in kinship coefficients. The analysis using GEE2 models revealed structures that would not have been detected by other models and should therefore be used in addition to traditional approaches of analysing family data. GEE2 should be considered a standard method for the investigation of familial aggregation. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Family studies are often considered the key for discrimination between genetic and environmental aetiologies [1]. If familial aggregation is found, the second step is to discriminate

[*] Correspondence to: Andreas Ziegler, Medical Centre for Methodology and Health Research, Institute of Medical Biometry and Epidemiology, Philipps-University of Marburg, Bunsenstr. 3, D-35033 Marburg, Germany.
[†] E-mail: ziegler@mailer.uni-marburg.de

[‡] Presented at the International Society for Clinical Biostatistics Twentieth International Meeting, Heidelberg, Germany, September 1999.

Contract/grant sponsor: Deutsche Forschungsgemeinschaft

Received October 1999
Accepted August 2000

among environmental and/or genetic factors that contribute to this clustering [2, 3]. For this purpose, several approaches have been proposed (for an overview see reference [3]). These include segregation analyses which are designed to test whether or not the data are compatible with Mendelian expectations by estimating parameters of a given genetic model of inheritance with latent genes [4]. With these parameter estimates that can be obtained for example, by maximum likelihood (ML) [4] or generalized estimating equations (GEE) [5], the magnitude of genetic sources of variation in the variable of interest can be determined. Segregation analyses, however, impose rather restrictive assumptions which may not be valid [4, 6].

As an alternative, the sources of aggregation might be analysed by investigating the mean and the association structure via second-order generalized estimating equations (GEE2). They have been applied several times in family studies (for an overview see, for example, reference [7]). This method may give hints as to the presence or absence of a genetic component. Pairs of siblings, for example, are expected to share 50 per cent of the genetic material. If, however, the sibling correlation is low with respect to the variable of interest, there is not much evidence for the presence of a genetic component that might be relevant for the entire population. If the data additionally reveal a high spouse correlation, it seems plausible that environmental factors play a more important role than genetic factors.

A statistical test can be carried out to compare the correlation between any two different relative pairs. For example, parent–offspring pairs and sibling pairs share on the average the same amount of genetic material. Hence, the respective correlation should be identical after adjustment for environmental factors if Mendelian genetic effects play an important role. In a general situation, the expected correlation between any relative pair depends on its kinship coefficient [8]. This corresponds to the probability that any two related individuals have received a gene from a common ancestor. Thus, the kinship coefficient is a genetic distance measure which can be used to compare the correlation between related pairs.

Apolipoprotein A1 (apoA1) is the major component of high-density lipoprotein C (HDL-C), while apolipoprotein B (apoB) is the predominant protein of low-density lipoprotein C (LDL-C) [9, 10]. Several studies have found that both serum apoA1 and apoB levels are strong predictors for atherosclerosis and coronary/cerebral artery disease (CAD) (for example, references [11–14]). Furthermore, it has been suggested that apoA1 and apoB are better predictors for CAD than HDL-C and LDL-C, respectively [15]. While the risk for atherosclerosis and CAD increases with serum apoB concentrations, it decreases with increasing plasma apoA1 levels.

It is well known that both apoA1 and apoB reveal strong familial aggregation. Results using segregation analysis were quite contradictory [16]. We therefore used GEE2 models to investigate whether this approach can be used to analyse the correlation structure in a large population based family data set.

The data set and the available family structures are briefly presented in Section 2. The GEE2 methodology is outlined in Section 3 as applied to our data set. Furthermore, we extend the jack-knife estimator of the robust variance matrix and some approximations to GEE2. In Section 4 the different variance estimators for GEE2 are compared in a simulation study. Section 5 presents the estimation results for the real data set in detail. Some further discussion and extensions of our work appear in Section 6.

Table I. Number of (blood) relative pairs for apolipoprotein A1 and apolipoprotein B.

Familial relationship	Number of pairs
Parent–offspring	802
Sibling	235
Grandparent–grandchild	229
Uncle/aunt–nephew/niece	198
First cousin	104
Couple	287

2. DATA SET

Detailed descriptions of family recruitment and data collection are given, for example, in references [13, 16, 17]. In short, the present study represents a continuation of the Donolo-Tel Aviv prospective artery disease study which started in 1964 in eight kibbutzim (co-operative settlements). The present study was based on a sample of 1466 individuals who were contacted during 1992 and 1993. One should, however, note that the sample mostly consists of members of agricultural co-operative settlements and cannot therefore be considered as a representative sample of the Israeli population. Medical records were available on all individuals, and the history was abstracted from these records [16]. Blood sampling and quantitative determination of plasma concentrations were performed as described previously [16]. Additional variables were gathered using a standardized questionnaire. These included smoking habits and sociodemographic data. Some misclassification in the confounding variables is possible, as only basic information was obtained and very crude categorization was performed in the study [16]. After validation and plausibility checks, data remained from 1435 living individuals in 469 families. ApoA1 and apoB plasma concentrations were assessed for 1239 and 1240 individuals, respectively. A core data set for apoA1 consisting of gender ($1 = \text{female}$, $0 = \text{male}$), age, body mass index (BMI, kg/m^2) smoking habits ($1 = \text{smoking}$, $0 = \text{else}$), marital status ($1 = \text{married}$, $0 = \text{else}$), immigration to Israel ($1 = \text{yes}$, $0 = \text{no}$) and current sport activity ($1 = \text{yes}$, $0 = \text{no}$) was available for 1226 individuals in 401 families. The core data set for apoB included one additional family. The size of pedigrees varied from 1 to 14 individuals, representing complex three-generation pedigrees [18]. The first generation consisted of 369 individuals (52.3 per cent female) comprising either participants who already took part in the 1964 study or their spouses. The largest group was formed by 523 individuals (49.3 per cent females) from the second generation, while 333 (apoA1) and 334 (apoB) individuals were from the third generation. Table I displays the numbers of specific (blood) relative pairs of whom apoA1 and apoB, respectively, were measured.

3. GENERALIZED ESTIMATING EQUATIONS FOR MEAN AND ASSOCIATION STRUCTURES

Marginal models for the mean structure, termed GEE1, are increasing in popularity since they have been implemented in several standard software packages [19]. Our aim, however, is the consistent estimation of both the mean and the association structure. An appropriate measure of the association is the correlation coefficient which should not be subject to range restrictions for continuous response variables. An advantage of this measure of association is its simple and straightforward interpretation. Alternatives for continuous response variables are the covariance and the second ordinary moments. If they are used, the joint estimating equations for the mean and the association structure are more complicated and need to be solved simultaneously. Furthermore, the interpretation is not as simple as if the correlation coefficient is used [7]. Therefore, we decided to apply the GEE2 of Prentice, that is, using the correlation coefficient as the measure of association [20].

For the derivation of these estimating equations and the asymptotic covariance matrix of the parameter estimates we require some notation. Let y_{it} be the response of individual t, $t=1,\dots,T_i$, from family i, $i=1,\dots,n$. For each y_{it} a vector of covariates x_{it} is available, which possibly contains an intercept. The data are summarized to $y_i=(y_{i1},\dots,y_{iT_i})'$ and $X_i=(x'_{i1},\dots,x'_{iT_i})'$. The pairs (y_i,X_i) are assumed to be independent. In our application we focus on continuous responses. Therefore, we use the identity link function to connect the conditional mean of y_{it} given X_i and the $p\times 1$ parameter vector β of the mean structure

$$\mu_{it}=E(y_{it}\,|\,X_i)=E(y_{it}\,|\,x_{it})=x'_{it}\beta \tag{1}$$

We furthermore assume that the correlation coefficient is a function of the $q\times 1$ parameter vector α of the association structure but independent of the mean structure parameter β. We choose the area hyperbolic tangent as association link function [7] so that for $t\neq t'$

$$\rho_{itt'}=\operatorname{corr}(y_{it},y_{it'}\,|\,X_i)=\frac{\exp\{k(\tilde{x}_{it},\tilde{x}_{it'})'\alpha\}-1}{\exp\{k(\tilde{x}_{it},\tilde{x}_{it'})'\alpha\}+1} \tag{2}$$

Equation (2) guarantees that the correlation coefficient does not exceed 1 in absolute values. k is a function that describes the relationship between the explanatory variables for the association structure $\tilde{x}_{it}$ and $\tilde{x}_{it'}$ and the correlation coefficient [21].

The GEE proposed by Prentice [20] are given by

$$u(\hat{\xi})=u\begin{pmatrix}\hat{\beta}\\ \hat{\alpha}\end{pmatrix}=\sum_{i=1}^{n}\begin{pmatrix}D_i & 0\\ 0 & E_i\end{pmatrix}'\begin{pmatrix}V_i & 0\\ 0 & W_i\end{pmatrix}^{-1}\begin{pmatrix}y_i-\mu_i\\ z_i-\rho_i\end{pmatrix}=0 \tag{3}$$

where μ_i is the $T_i\times 1$ vector of the μ_{it} and ρ_i is the $T_i(T_i-1)/2\times 1$ vector of the $\rho_{itt'}$. z_i is the corresponding vector of the product of the standardized residuals $z_{itt'}=(y_{it}-\mu_{it})(y_{it'}-\mu_{it'})/\sigma_{it}\sigma_{it'}$ with $\sigma^2_{it}=\operatorname{var}(y_{it}\,|\,X_i)$. $D_i=\partial\mu_i/\partial\beta'$ and $E_i=\partial\rho_i/\partial\alpha'$ are the first derivatives, while V_i and W_i are the conditional working covariance matrices of y_i and z_i given X_i, respectively. Usually, W_i is chosen as the working matrix for applications [7] so that W_i is a $T_i(T_i-1)/2$-dimensional identity matrix.

The GEE (3) for β and α may be solved separately by an alternating modified Fisher scoring algorithm because they can be partitioned in two independent estimating equations. Equation (3) can be derived from the generalized method of moments [22]. Thus, $\hat{\xi}=(\hat{\beta}',\hat{\alpha}')'$ is a strongly consistent estimator of $\xi=(\beta',\alpha')'$ under suitable regularity conditions [23], if equations (1) and (2) are correctly specified. Furthermore, $\hat{\beta}$ and $\hat{\alpha}$ are jointly asymptotic normal. The robust variance matrix, also termed Huber or sandwich variance matrix, is given, for example, by Prentice [20]. In the framework of GEE1, Paik [24] recommended using jack-knife estimators of variance instead of the robust variance matrix in small samples because the robust variance matrix yielded biased estimates. Lipsitz *et al.* [25, 26] showed for the GEE1 that the unweighted deletion-1 jack-knife estimator of variance

$$\left(\frac{n-p}{n}\right)\sum_{i=1}^{n}(\hat{\beta}_{-i}-\hat{\beta})(\hat{\beta}_{-i}-\hat{\beta})'$$

is asymptotically equivalent to the corresponding robust variance matrix. This property can be easily extended to the GEE2 of equation (3). Here, $\frac{n-p}{n}$ is replaced by $\frac{n-(p+q)}{n}$. Furthermore, the jack-knife now involves both β and α. Deletion-1 jack-knife estimators are usually obtained by a modified Fisher scoring with starting value $\hat{\xi}=(\hat{\beta}',\hat{\alpha}')'$, where each family is successively omitted in a loop. Instead of the fully iterated (FIJ) jack-knife estimator, a 'one-step' approximation (1-SJ) might be used by stopping the algorithm after one Fisher scoring step [25]. For GEE1, the 'one-step' approximation gave better coverage probabilities than the fully iterated jack-knife estimator in Monte Carlo simulations [25].

The jack-knife estimator of variance can also be approximated without successively leaving out each cluster during the calculations as shown by Ziegler [27] for GEE1. This generally increases the computation speed. The approximation of the jack-knife estimator of variance (AJS) for Prentice's GEE2 is derived in the Appendix.

All three proposed jack-knife estimators are implemented in MAREG which is a freely available menu-driven software package for the analysis of marginal regression models [28]. In this program, the FIJ and the 1-SJ are calculated via a modified Fisher scoring algorithm. The classical robust variance estimator of Prentice and the AJS are available as a standard.

4. SIMULATION STUDY

To compare the properties of the three jack-knife estimators with the usual robust estimator of variance we performed a simulation study using a continuous response variable, the identity link function and the area hyperbolic tangent association link function.

In previous studies, the jack-knife was shown to be superior to the classic robust variance for small sample sizes [24, 25]. Thus, 50 clusters (families) of size 3 were simulated with 1000 replicates for each model. The simulation proceeded as follows. First, the design matrix X was generated for each cluster. Second, the response vector y was simulated for each cluster using a multivariate normal distribution. The pseudo-random numbers were generated using DRAND48, which is supplied by SUNOS 5.5 (man Pages(3C)) as a C-library function [29]. The estimation is done by MAREG [28]. Details on the generation process are described in Kastner and Ziegler [30].

Mancl and Leroux [31] have shown that the efficiency of GEE estimates is quite sensitive to the between- and within-cluster variation of the explanatory variables. Thus, we chose eight different models that specifically focused on this aspect in our simulations. They all included one non-random binary and one non-random continuous explanatory variable for the mean structure. The latter was generated from the frequency distribution of a grouped variable. The covariates were subject to variation as they were chosen to be either cluster-constant or non-mean-balanced cluster-specific. This resulted in four different configurations for the parameters

Table II. Simulation results with mean-unbalanced within-cluster varying binary and continuous explanatory variables for an exchangeable and an unspecified correlation structure.

Association model	Parameter	Theoretical value	Mean parameter estimate	Standard error of the mean	Standard error			
					Prentice	AJS[*]	1-SJ[†]	FIJ[‡]
Exchangeable	θ_0	1	1.000	0.245	0.243	0.249	0.248	0.250
	θ_1	3	3.000	0.163	0.162	0.164	0.162	0.165
	θ_2	−0.2	−0.200	0.040	0.040	0.041	0.041	0.041
	θ_3	0.7	0.646	0.204	0.333	0.241	0.192	0.202
Unspecified	θ_0	1	0.998	0.244	0.239	0.241	0.240	0.246
	θ_1	3	2.997	0.152	0.147	0.147	0.146	0.150
	θ_2	−0.2	−0.200	0.040	0.039	0.040	0.039	0.041
	θ_3	0.5	0.460	0.280	0.376	0.286	0.265	0.274
	θ_4	1	0.974	0.319	0.494	0.376	0.301	0.309
	θ_5	0.7	0.681	0.305	0.383	0.322	0.283	0.292

[*] Approximation of the jack-knife estimator of variance.

[†] One-step approximation of the jack-knife estimator of variance.

[‡] Fully iterated jack-knife estimator of variance.

of the mean structure. In any case the theoretical parameters for the explanatory variables were $\theta_0 = 1$ (regression constant), $\theta_1 = 3$ (binary variable) and $\theta_2 = -0.2$ (continuous variable).

Our aim was to estimate quite complex association structures using the GEE2 for the apolipoprotein data. Therefore, we used two different association structures – exchangeable and unstructured – for the Monte Carlo simulations. Parameters for the unspecified correlation structure were chosen to be $\theta_3 = 0.5$, $\theta_4 = 1$ and $\theta_5 = 0.7$ resulting in correlation coefficients of 0.245, 0.462 and 0.336 for the pairs 1–2, 1–3 and 2–3, respectively. The parameter value for the exchangeable correlation structure was set to $\theta_3 = 0.7$.

Table II summarizes the relevant results from the Monte Carlo simulations with time-varying covariates using the four different approaches for estimating or approximating the robust variance matrix. The upper part of Table II presents the results for the exchangeable correlation structure, whereas its lower part shows the unspecified correlation structure. Table II displays the mean parameter estimate and the standard error of the mean from the 1000 replicates in addition to the theoretical parameter values and the different estimates of the standard error.

Obviously, with either of the four approaches for estimating or approximating the robust variance matrix, the standard error of the mean was well approximated for the parameter estimates of the mean structure, that is, θ_0, θ_1 and θ_2. This is in line with the findings of Lipsitz and colleagues [25, 26] and Ziegler [27]. However, the results differed substantially with respect to the association structure, that is, θ_3, θ_4 and θ_5. The standard errors using the usual robust variance matrix according to Prentice [20] were far too large, resulting in conservative tests. On the other side, the 1-SJ was too liberal for all eight models. The AJS generally was conservative for the simulated models. The best approximation to the true standard error of the mean was obtained with the FIJ. Therefore, we decided to apply more complex Wald tests to the apolipoprotein data using the FIJ.

Table III. Generalized estimating equations results (parameter estimates and p-values) for apolipoprotein A1 (1226 individuals in 401 families).

		P-value			
Variable	Parameter estimate	Prentice	AJS*	1-SJ†	FIJ‡
Mean structure					
Intercept	106.750	<0.001	<0.001	<0.001	<0.001
Sex	11.418	<0.001	<0.001	<0.001	<0.001
Age	0.076	0.014	0.009	0.013	0.012
Smoke	5.297	0.002	0.002	0.003	0.002
Age × smoke	0.116	0.021	0.020	0.022	0.021
Association structure					
Parent–offspring	0.348	0.105	<0.001	0.002	0.001
Sibling	0.695	0.228	0.001	0.009	0.009
Grandparent–grandchild	0.619	0.164	0.016	0.017	0.018
Uncle/aunt–nephew/niece	0.278	0.255	0.076	0.091	0.076
First cousin	0.230	0.512	0.453	0.450	0.449
Couple	0.368	0.079	0.005	0.002	0.002
Gender similarity	0.237	0.141	0.018	0.011	0.010

* Approximation of the jack-knife estimator of variance.
† One-step approximation of the jack-knife estimator of variance.
‡ Fully iterated jack-knife estimator of variance.

5. RESULTS

For the analysis of the apolipoprotein data, we first applied both forward and backward model selection procedures for the mean structure separately for apoA1 and apoB by ignoring the intra-familial correlation. Second, a fine-modelling of the mean structure was performed with the independence estimating equations (IEE) [7] using the union of the parameters from the model selection procedures. We additionally investigated pairwise interactions for the variables remaining in the model of the mean structure. Third, the association structure was systematically evaluated using the six familial relationships described in Section 2. Furthermore, the following explanatory variables were used to model the association structure: absolute age difference of a pair; mean age of a pair; difference in generations of a pair (0, 1, 2); gender similarity (same gender = 1, different gender = 0), and standard deviation of the BMI within a family. The association structure was obtained using a forward selection procedure for variables that yielded p-values < 0.05 when investigated in addition to the six familial relationships.

The final models for apoA1 and apoB are displayed in Tables III and IV, respectively. The upper part of these tables show the explanatory variables of the mean structure with a nominal p-value < 0.05 obtained after model selection. The lower part of Table III and Table IV presents the regression coefficients of the association structure. The final model included explanatory variables for the association structure that revealed nominal p-values < 0.05 as well as coefficients of all relative pairs. The final models for the mean structure were different for apoA1 and apoB; while the model for apoA1 included age, gender, smoking and the interaction between age and smoking, p-values < 0.05 emerged in the apoB model for

Table IV. Generalized estimating equations results (parameter estimates and p-values) for apolipoprotein B (1227 individuals in 402 families).

		P-value			
Variable	Parameter estimate	Prentice	AJS*	1-SJ†	FIJ‡
Mean structure					
Intercept	31.052	<0.001	<0.001	<0.001	<0.001
Sex	19.861	<0.001	<0.001	<0.001	<0.001
Age	1.705	<0.001	<0.001	<0.001	<0.001
Age square	−1.218	<0.001	<0.001	<0.001	<0.001
Age × sex	−0.957	<0.001	<0.001	<0.001	<0.001
Age square × sex	0.991	<0.001	<0.001	<0.001	<0.001
Association structure					
Parent–offspring	1.210	0.034	0.005	0.003	0.003
Sibling	0.574	0.012	<0.001	<0.001	<0.001
Grandparent–grandchild	0.026	0.813	0.803	0.807	0.816
Uncle/aunt–nephew/niece	0.266	0.086	0.026	0.035	0.034
First cousin	−0.195	0.507	0.424	0.448	0.471
Couple	0.216	0.099	0.067	0.057	0.059
Age diff. parent–offspring	−0.027	0.101	0.053	0.046	0.050§

* Approximation of the jack-knife estimator of variance.

† One-step approximation of the jack-knife estimator of variance.

‡ Fully iterated jack-knife estimator of variance.

§ More precise p-value 0.04998.

age and age squared, gender and their respective interactions. Smoking was not significant at the 5 per cent test level for apoB. No other variables of the core data set yielded nominal p-values <0.05 using a single variable Wald test. These findings might be due to the very basic information that was obtained in the questionnaire. The coefficient of determination was 16.3 per cent and 30.8 per cent for the apoA1 and the apoB model, respectively.

The association structure for apoA1 showed moderate parent–offspring and high sibling associations. The final model also included an association parameter for equal gender, that is, individuals of the same gender had more similar apoA1 levels than those of different gender ($p=0.010$). For example, the correlation of either female or male sibling pairs was 0.43, whereas that of siblings of different gender was 0.33. The association structure also contained a significant couple association ($\rho=0.18$, $p=0.002$). The only pairs that revealed no significant association at the 5 per cent test level were uncle/aunt–nephew/niece pairs ($\rho=0.14$ for same gender, $\rho=0.25$ for different gender, $p=0.076$) and first cousin pairs ($\rho=0.11$ for same gender, $\rho=0.23$ for different gender, $p=0.449$) most likely due to the relative low number of pairs. No other explanatory variable for the association structure revealed nominal p-values $<$ 0.05 for any of the four estimates of the robust variance matrix.

In order to test whether associations of different pairs of relatives decrease with increasing genetic distance, we also performed more complex Wald tests. According to their genetic distance, parent–offspring and sibling pairs should share approximately the same amount of genetic material. Thus, if familial aggregation of apoA1 arose on a genetic basis, the correlation ρ between these pairs should be similar. In addition, the correlation of grandparent–grandchild pairs and uncle/aunt–nephew/niece pairs should be approximately $1/2\rho$. Furthermore, first cousins share approximately 12.5 per cent of their genetic material so that the association should be $1/4\rho$. Finally, couples should share no genetic material so that $\rho = 0$, if no environmental effect is present. Therefore, we carried out three different tests using the correlation coefficients. Association parameters α were transformed to correlation coefficients ρ by the multivariate delta method [32]. The first test included all six relations jointly. Second, we performed a Wald test for the five relative pairs that share genetic material. Third, we surmised that the shared and non-shared environment was similar for parent–offspring and sibling pairs. Thus, we tested the equality of the respective correlation parameters. The first hypothesis was rejected at the 5 per cent test level ($p = 0.009$). The second and third hypotheses were not rejected at the 5 per cent test level ($p = 0.291$ and $p = 0.166$, respectively). Thus, the rejection of the first hypothesis was mainly due to the relatively high correlation between spouses. We conclude that the results cannot be explained by genetic effects only. Shared environment influenced familial aggregation.

In contrast to the Monte Carlo simulations presented in the last section, the p-values of the 1-SJ and the FIJ were very similar for the apoA1 models (Table III). The usual robust variance matrix, however, yielded no p-value < 0.05 for any of the association parameters of the final model. The latter result coincides well with the findings of Section 4 where the classic robust variance estimator was too conservative for the simulated models. They can be explained by badly-conditioned [33] covariance matrices; the condition number of both the Fisher information matrix (see Appendix) and the outer product gradient (OPG) [22] was approximately 16. The robust variance matrix, however, which is the sandwich product of these matrices, had a condition number of 99.

Table IV shows that the results of the association structure for apoB are different from those for the apoA1 data. The final model included the coefficients for the relative pairs in addition to an age difference for parent–offspring pairs. The highest correlation was obtained for pairs of siblings ($\rho = 0.279$, $p < 0.001$). Interestingly, both the grandparent–grandchild correlation ($\rho = 0.013$, $p = 0.816$) and the first cousin correlation ($\rho = -0.097$, $p = 0.471$) were low. Instead, the uncle/aunt–nephew/niece correlation revealed some correlation ($\rho = 0.132$, $p = 0.034$).

When interpreting the correlation of parent–offspring pairs, one should bear in mind that this association followed a regression with a simple linear predictor and the inverse of the area hyperbolic tangent as association response function. Thus, the similarity of parent–offspring pairs decreased with their age difference. Nevertheless, the p-values for this continuous variable differed relevantly for the four variance estimators shown in Table IV. While the FIJ and the 1-SJ revealed nominal p-values<0.05, the classical robust variance estimator and the AJS resulted in p-values > 0.05. Since the decision about the final model was made upon use of the FIJ (see Section 4), we retained the variable in the model.

Figure 1 displays this relationship between the age difference and the correlation coefficient bounded by minimum (16 years) and maximum (53 years) age difference. Figure 1 also shows the 95 per cent prediction interval for the correlation coefficient. The relationship seems to follow a straight line. However, it was obtained by transforming the estimated association parameters for given age difference to a correlation coefficient using the area hyperbolic tangent association link function (equation (2)) and the delta method. The correlation between parent–offspring pairs was negative if the age difference was larger than 41 years. For the mean age difference of parent–offspring pairs observed in this sample (about 30 years), the correlation was 0.151. In analogy to the apoA1 data, we tested the three more complex hypotheses

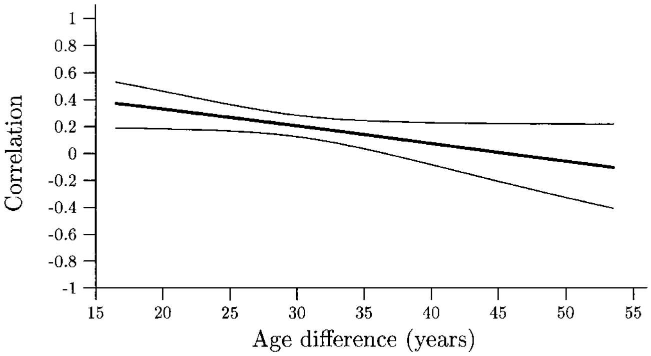

Figure 1. Correlation coefficient and 95 per cent prediction interval of apolipoprotein B for parent–offspring pairs in relationship to their age difference.

about the association structure. For parent–offspring pairs, we used the typical association which was defined by the mean age difference which was 30 years. None of the three tested hypotheses was significant at the 5 per cent test level ($p=0.153$ for hypothesis 1, $p=0.287$ for hypothesis 2, $p=0.231$ for hypothesis 3). Thus, the correlations were consistent with those expected from a genetic model after adjustments for age difference in parent–offspring pairs.

6. DISCUSSION

In this paper we analysed familial aggregation of apolipoproteins by modelling both the mean and the association structure with the GEE2 of Prentice [20]. In the analysis, we used the fully iterated jack-knife estimator, its one-step approximation and an approximation to the jack-knife estimator of variance in addition to the usually applied robust variance estimator. We carried out a Monte Carlo simulation study to investigate the different variance estimators by using cluster-varying explanatory variables that were not mean-balanced. While the standard errors of the different variance estimators were similar for the parameters of the mean structure, they varied substantially for the parameters of the association structure. In accordance with the studies of Paik [24] and Lipsitz *et al.* [25, 26], who investigated jack-knife estimators for the mean structure, we found that the usually applied robust variance matrix was biased for the parameters of the association structure. Furthermore, in all simulated models the robust variance matrix was too conservative. The most appropriate estimator for the variance of the association parameters was the FIJ. This estimator is, however, CPU time consuming. For the analysis of the real data, the estimation process of one model took about 75 min on a Pentium II computer with 266 MHz and 128 MB RAM. The AJS or the 1-SJ seem to be satisfactory alternatives for the model building stage if CPU time is limited. Furthermore, they are readily available.

When the method was applied to the apolipoprotein data set, several new and interesting results were obtained. First, we found that individuals of the same gender had more similar apoA1 levels than individuals of different gender. Second, we demonstrated that the similarity of apoB levels decreased with increasing age difference of parent–offspring pairs. To our knowledge, these associations have not been discussed in the literature before. We furthermore carried out quite complex hypothesis tests about the familial aggregation of apoA1 and apoB with classical Wald tests. They showed that the correlation between family members cannot be explained solely by shared genetics but that there is also a non-negligible environmental contribution. The relatively high correlation between spouses is an indicator for this. Interestingly, for apoB, the correlation decreases with increasing age difference between the pairs which also can be interpreted as an environmental effect. This association had not been found in the previous analysis when mixed models were applied in complex segregation analyses [16].

The advantage of the current data set is that a large number of families was available for the analysis, although for some families not all family members could be included in the study. Also, only some basic information was available on other important cofactors. While an indicator for smoking habits was available and used in the analysis, no information on diet or other risk factors influencing lipid were available.

All analyses were easily carried out by MAREG, which is a menu-driven freely available stand-alone computer program for marginal regression models [28]. Specific association structures can be estimated by setting up an ASCII 'design matrix' Z for the association structure which is similar to the usual X matrix of explanatory variables. The only difference between Z and X is that the families – or more generally clusters – have an additional row that contains the number of pairs that are used for the association structure. Thus, the Z file can generally be created from the X matrix using macros.

For the data analysis, we performed a two-step approach. First the mean structure was modelled and the results of variable selection was then used to model the correlation. The mean structure was modelled without taking into account the correlation between subjects. We used this two-step procedure since no software was available for model building of correlated continuous variables. Furthermore, we believe that the model for the mean structure would not change in our data.

Summing up, our analyses using GEE2 revealed association structures that would not have been seen using other models. Therefore, they may be used in addition to other classical approaches to the analyses of family data, like segregation analysis. These analyses would benefit from the implementation of model selection procedures and standard regression diagnostic tools. Further research may be needed to investigate the stability of the estimators.

APPENDIX: AN APPROXIMATION TO THE JACK-KNIFE ESTIMATOR OF VARIANCE FOR GENERALIZED ESTIMATING EQUATIONS

Ziegler (reference [27], equation (10)) has shown for the GEE1 that an application of the update formula for symmetric matrices can be used to obtain an asymptotically equivalent formulation of the jack-knife estimator of variance

$$\frac{n-p}{n}\hat{A}_{11}^{-1}\left[\sum_{i=1}^{n}\{\hat{F}_i'\hat{K}_i\hat{V}_i^{-1/2}(y_i-\hat{\mu}_i)(y_i-\hat{\mu}_i)'\hat{V}_i^{-1/2'}\hat{K}_i'\hat{F}_i'\}\right]\hat{A}_{11}^{-1} \tag{A1}$$

where $\hat{F}_i=\hat{V}_i^{-1/2}\hat{D}_i \sim T_i \times p$ and $K_i = I_{T_i\times T_i} + (I_{T_i\times T_i} - \hat{F}_i\hat{A}_{11}^{-1}\hat{F}_i')^{-1}\hat{F}_i\hat{A}_{11}^{-1}\hat{F}_i'$. Here, $I_{T_i\times T_i}$ is the T_i dimensional identity matrix, $\hat{V}_i^{1/2}$ is a root of $\hat{V}_i$ satisfying $\hat{V}_i=\hat{V}_i^{1/2}\hat{V}_i^{1/2'}$. In MAREG [28] the Cholesky decomposition is used to obtain $\hat{V}_i^{1/2}$. Furthermore, $\hat{A}_{11}^{-1}$ is the inverse of the upper left block of the Fisher information matrix

$$A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} D_i' V_i^{-1} D_i & 0 \\ -\sum_{i=1}^{n} E_i' W_i^{-1} \dfrac{\partial z_i}{\partial \beta} & \sum_{i=1}^{n} E_i' W_i^{-1} E_i \end{pmatrix}$$

The approximation (A1) may be extended to the GEE2 for $T_i \geqslant 2$

$$\frac{n-(p+q)}{n} \hat{A}^{-1} \left[\sum_{i=1}^{n} \left\{ \tilde{F}_i' \tilde{K}_i \tilde{V}_i^{-1/2} \begin{pmatrix} y_i - \hat{\mu}_i \\ z_i - \hat{\rho}_i \end{pmatrix} \begin{pmatrix} y_i - \hat{\mu}_i \\ z_i - \hat{\rho}_i \end{pmatrix}' \tilde{V}_i^{-1/2'} \tilde{K}_i' \tilde{F}_i \right\} \right] \hat{A}^{-1}$$

Here, $\tilde{F}_i$ is the $(T_i + T_i(T_i - 1)/2) \times (p+q)$ block diagonal matrix of $(\hat{V}_i^{-1/2} \hat{D}_i, \hat{W}_i^{-1/2} \hat{E}_i)$. Analogously, $\tilde{V}_i^{1/2}$ is the block diagonal matrix of $(\hat{V}_i^{1/2}, \hat{W}_i^{1/2})$ and, finally, $K_i = I_{T_i+T_i(T_i-1)/2 \times T_i+T_i(T_i-1)/2} + (I_{T_i+T_i(T_i-1)/2 \times T_i+T_i(T_i-1)/2} - \tilde{F}_i \hat{A}^{-1} \tilde{F}_i')^{-1} \tilde{F}_i \hat{A}^{-1} \tilde{F}_i'$. For $T_i = 1$, the estimation can be performed by letting $z_i - \hat{\rho}_i = 0$ and adding a row and column of 0 to $\hat{F}_i$ so that $\tilde{F}_i = \begin{pmatrix} \hat{F}_i & 0 \\ 0 & 0 \end{pmatrix}$.

ACKNOWLEDGEMENTS

The authors are grateful to Karin Mählmann for her careful analysis of the mean structure. The work of C.K. was supported by the Deutsche Forschungsgemeinschaft.

REFERENCES

1. Dorman JS, Trucco M, LaPorte RE, Kuller LH. Family studies: the key to understanding the genetic and environmental etiology of chronic disease? *Genetic Epidemiology* 1988; **5**:305–310.
2. King MC, Lee GM, Spinner NB, Thomson G, Wrensch MR. Genetic epidemiology. *Annual Review of Public Health* 1984; **5**:1–52.
3. Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of Genetic Epidemiology*. Oxford University Press: New York, 1993.
4. Jarvik GP. Complex segregation analyses: uses and limitations. *American Journal of Human Genetics* 1998; **63**:942–946.
5. Zhao LP. Segregation analysis of human pedigrees using estimating equations. *Biometrika* 1994; **81**:197–209.
6. Ziegler A, Hebebrand J. Sample size calculations for linkage analysis using extreme sib pairs based on segregation analysis with the quantitative phenotype body weight as an example. *Genetic Epidemiology* 1998; **15**:577–593.
7. Ziegler A, Kastner C, Blettner M. The generalised estimating equations: an annotated bibliography. *Biometrical Journal* 1998; **40**:115–139.
8. Vogel F, Motulsky AG. *Human Genetics*: *Problems and Approaches*, 3rd edn. Springer: New York, 1997.
9. Breslow JL. Human apolipoprotein molecular biology and genetic variation. *Annual Revision of Biochemics* 1985; **54**:699–727.
10. Havel RJ, Goldstein JL, Brown MS. Lipoproteins and lipid transport. In *Metabolic Control and Disease*, Bondy CPIC, Rosenberg LE (eds). Saunders, W.B.: Philadelphia, 1980; 393–494.
11. Assmann G, Schulte H, von Eckardstein A, Huang Y. High-density lipoprotein cholesterol as a predictor of coronary heart disease risk. The PROCAM experience and pathophysiological implications for reverse cholesterol transport. *Atherosclerosis* 1996; **124 Supp**:S11–20.
12. Kwiterovich PO Jr, Coresh J, Smith HH, Bachorik PS, Derby CA, Pearson TA. Comparison of the plasma levels of apolipoproteins B and A-1, and other risk factors in men and women with premature coronary artery disease. *American Journal of Cardiology* 1992; **69**:1015–1021.

13. Livshits G, Weisbort J, Meshulam N, Brunner D. Multivariate analysis of the twenty-year follow-up of the Donolo-Tel Aviv Prospective Coronary Artery Disease Study and the usefulness of high density lipoprotein cholesterol percentage. *American Journal of Cardiology* 1989; **63**:676–681.
14. Schmitz G, Lackner KJ. High-density lipoproteins and atherosclerosis. *Current Opinion in Lipidology* 1993; **4**:392–400.
15. Kottke BA, Zinsmeister AR, Holmes DR, Kneller RW, Hallaway BJ, Mao SJ. Apolipoproteins and coronary artery disease. *Mayo Clinic Proceedings* 1986; **61**:313–320.
16. Livshits G, Blettner M, Graff E, Hoting I, Wahrendorf J, Brunner D, Schettler G. Tel Aviv-Heidelberg three-generation offspring study: genetic determinants of apolipoprotein A1 and apolipoprotein B, *American Journal of Medical Genetics* 1995; **57**:410–416.
17. Brunner D, Weisbort J, Meshulam N, Schwartz S, Lin J, Kaplan C, Zhao X, Bisson B, Fitzpatrick V, Dodge H. Relation of total cholesterol and high-density lipoprotein cholesterol percentage to the incidence of definite coronary events. Twenty-year follow-up of the Donolo-Tel-Aviv prospective coronary artery disease study. *American Journal of Cardiology* 1990; **59**:1271–1276.
18. Livshits G, Schettler G, Graff E, Blettner M, Wahrendorf J, Brunner D. Tel Aviv-Heidelberg three-generation offspring study: genetic determinants of plasma fibrinogen level. *American Journal of Medical Genetics* 1996; **63**:509–517.
19. Ziegler A, Grömping U. The generalized estimating equations: a comparison of procedures available in commercial statistical software packages. *Biometrical Journal* 1998; **40**:247–262.
20. Prentice RL. Correlated binary regression with covariates specific to each binary observation. *Biometrics* 1988; **44**:1033–1048.
21. Lipsitz SR, Laird NM, Harrington DP. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* 1991; **78**:153–160.
22. Ziegler A. The different parameterizations of the GEE1 and the GEE2. In *Statistical Modelling Proceedings of the 10th International Workshop on Statistical Modelling*, Seeber GU, Francis BJ, Hatzinger R, Steckel-Berger G (eds). Lecture Notes in Statistics, 104. Springer: Innsbruck, Austria, 1995; 315–324.
23. Hansen L. Large sample properties of generalized method of moment estimators. *Econometrica* 1982; **50**:1029–1055.
24. Paik MC. Repeated measurement analysis for nonnormal data in small samples. *Communications in Statistics – Simulation and Computation* 1988; **17**:1155–1171.
25. Lipsitz SR, Laird NM, Harrington DP. Using the jackknife to estimate the variance of regression estimators from repeated measures studies. *Communications in Statistics – Theory and Methods* 1990; **19**:821–845.
26. Lipsitz SR, Dear KB, Zhao L. Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics* 1994; **50**:842–846.
27. Ziegler A. Practical considerations on the jackknife estimator of variance for generalized estimating equations. *Statistical Papers* 1997; **38**:363–369.
28. Kastner C, Fieger A, Heumann C. MAREG and WinMAREG-a tool for marginal regression models. *Statistical Software Newsletter in Computational Statistics and Data Analysis* 1997; **24**:237–241.
29. SunOS. *SunOS Reference Manual*. Sun Microsystems: Mountain View, CA, 1995.
30. Kastner C, Ziegler A. A comparison of jackknife estimators of variance for GEE2. SFB 386 Discussion Paper # 167, Ludwig-Maximilians University of Munich. Available at http://www.stat.uni-muenchen.de/sfb386/publikation.html, 1999.
31. Mancl LA, Leroux BG. Efficiency of regression estimates for clustered data. *Biometrics* 1996; **52**:500–511.
32. Rao CR. *Linear Statistical Inference and its Applications*, 2nd edn. Wiley: New York, 1973.
33. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics*: *Identifying Influential Data and Sources of Collinearity*. Wiley: New York, 1980.