STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:1377–1398 (DOI: 10.1002/sim.1745)

GEE approaches to marginal regression models for medical diagnostic tests

Peter Martus[1,*,†], Andrea Stroux[1], Anselm M. Jünemann[2], Matthias Korth[2], Jost B. Jonas[3], Folkert K. Horn[2] and Andreas Ziegler[4,‡]

[1]*Department of Medical Informatics, Biometry and Epidemiology, Free University Berlin, Germany*
[2]*Department of Ophthalmology, University of Erlangen-Nürnberg, Germany*
[3]*Department of Ophthalmology, Faculty of Clinical Medicine Mannheim, University of Heidelberg, Germany*
[4]*Department of Medical Biometry and Statistics, University at Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany*

SUMMARY

The evaluation of a new medical diagnostic test may focus on two different scientific questions: (1) The new test may replace an existing one because of lower cost or higher validity. A related question would be the selection of the 'best' test(s) from a bundle of new or established measurements. (2) The new test may be used supplementary to other new or established procedures. In a recent publication, Leisenring and co-workers (*Stat Med* 1997; **16**:1263–1281) developed a general marginal regression model for comparisons of diagnostic tests focussing on question (1), i.e. on the selection of the 'best' procedure. They applied the GEE approach of Liang and Zeger (*Biometrika* 1987; **73**:13–22) to adjust for the correlation of data as a nuisance parameter. Using the general framework provided by Leisenring *et al.*, we extend their approach and apply the GEE methodology to question (2), i.e. to the investigation of which of several diagnostic tests should be used supplementary to each other. We analyse data from a longitudinal study concerning pathogenesis, diagnosis and long-term course of the eye disease glaucoma. We find a dependence of the correlation structure of several diagnostic measurements on the severity of the disease. This result may be useful in clinical applications as regards the selection of subsets of diagnostic measurements in individual diagnostic processes but also in investigations concerning the relationship of the pathogenic process and the rationales of the different diagnostic procedures.

KEY WORDS: GEE; marginal model; regression; diagnosis; glaucoma

1. INTRODUCTION

The biometrical evaluation of medical diagnostic tests has received increasing attention in the last decade. New technological advances require an elaborate methodology to determine their clinical value. The decision for the use of one or a subset of diagnostic procedures depends

*Correspondence to: Peter Martus, Department of Medical Informatics, Biometry and Epidemiology, Free University Berlin, Hindenburgdamm 30, 11200 Berlin, Germany.
† E-mail: martus@medizin.fu-berlin.de
‡ E-mail: ziegler@imbs.uni-luebeck.de

Contract/grant sponsor: SFB 539, DFG

Received March 2002
Accepted September 2003

on their diagnostic efficacy, financial cost, and on prior knowledge about the subjects to be diagnosed: In clinical settings a high sensitivity is desired, whereas in a screening situation, due to the lower prevalence of a disease, the main target will be high specificity combined with low cost. In general, it will not be optimal to apply the same bundle of diagnostic procedures in all situations.

For many diagnostic tests, sensitivity depends on the severity of the disease in the subpopulation of truly affected subjects. Therefore, if the diagnostic procedures to be compared are evaluated in distinct samples, an adjustment for disease severity is obviously required. However, in the preferable situation of intraindividual comparisons, this adjustment is reasonable, too: First, it may increase statistical power. Second, the ranking of diagnostic procedures according to efficacy may depend on the true severity of the disease.

In general, two different criteria for the usefulness of a new diagnostic test can be distinguished:

(1) The new test may replace an existing test. A related problem comprises the selection of the 'best' test(s) from a bundle of new or established procedures.
(2) The new test may be used supplementary to other new or established measurements.

In situation (1), decisions will be based on the comparison of single measurements or groups of measurements according to sensitivity and specificity. If the measurement scales are ordinal or continuous, subsets of ROC-curves or entire curves may be compared. In situation (2), the correlation structure of several diagnostic measurements is of interest too.

In the framework of situation (1), an elegant approach has been developed by Leisenring and co-workers [1]. They propose a general marginal regression model for comparisons of diagnostic measurements adjusted for the severity of disease. Standard errors are estimated by using the method of generalized estimating equations (GEE) [2]. In the simple case of two diagnostic tests without adjustment for disease severity, the method reduces to Mc Nemar's test [1].

In our study, we focus on the question of supplementary information comprised by several diagnostic tests. We present an approach that determines the correlation structure of several diagnostic measurements depending on the severity of disease. For this purpose, we extend the model of Leisenring and coworkers and use the type of generalized estimating equations proposed by Prentice [3, 4].

In our clinical example, we compare five sensory procedures used in the diagnosis of glaucoma, one of the major causes of blindness in the world [5]. The gold standard is comprised by a morphometric evaluation of the optic nerve head.

In Section 2 we present the clinical example; in Section 3 we describe the basic logistic regression model of Leisenring and co-workers [1]. In Sections 4 and 5 we develop the GEE models as applied in our study. In Section 6 we present the results of the example data. In Section 7 we discuss our findings and motivate some future work.

2. CLINICAL EXAMPLE

Glaucoma is the name of a group of eye diseases characterized by specific damage of the optic nerve head, often accompanied by elevated intraocular pressure and followed by specific glaucomatous visual field loss. In early stages, glaucomatous damage of the optic nerve is

difficult to detect: Patients are often unaware of the disease and the functional and morphologic characteristics upon which diagnostic procedures are based are subject to a large biological variability. Early detection of glaucoma and monitoring of patients already diagnosed with the disease are essential for therapeutic decision-making since glaucomatous damage is irreversible. The established diagnostic criteria for glaucoma are the inspection of the optic nerve head and the determination of visual field loss [5]. Ophthalmologists generally agree that elevated intraocular pressure is neither a necessary nor a sufficient diagnostic criterion. In contrast, it is unclear whether the sensory criterion 'visual field loss' is indeed required for the diagnosis of glaucoma or whether it is only an indicator of a later stage of the disease [6].

The broadly accepted sensory test for the determination of visual field loss is perimetry [7]. However, depending on its implementation, this method can be time consuming and depends on the concentration of the proband. Furthermore, it is subject to considerable long-term fluctuations [8] and learning effects [9].

In a clinically based prospective study performed at the University Eye Hospital of Erlangen new diagnostic devices for glaucoma were developed. Three groups of subjects were included in the study in accordance with the protocol: (1) patients with primary and secondary open-angle glaucoma characterized by optic nerve damage, (2) patients at risk for glaucoma (subjects with elevated intraocular pressure and absence of other glaucomatous signs), and (3) unaffected controls. In a specialized glaucoma outpatient clinic within the eye hospital, patients and controls were recruited according to the inclusion criteria defined in the study protocol (age between 18 and 70 years, visual acuity of 20/30 or better, clear optic media, absence of retinal diseases and diabetes mellitus). Controls were examined twice; patients were included in a long-term follow up comprising annual clinical examinations. A comprehensive diagnostic examination was performed during the first two visits; in the follow-up only a limited examination was carried out. The study complied with the principles of the Declaration of Helsinki. It was approved by the local ethics committee. All participants signed an informed consent form.

One aim of the study was to investigate both the sensitivity and specificity of new diagnostic measurement techniques developed and implemented within the project. Besides perimetry, two psychophysical and two electrophysiologic procedures served as examples in our study: (1) the measurement of contrast sensitivity (CS) [10] in a small area of the visual field. This area is known as the region where early glaucomatous damage often occurs. In contrast to this extremely localized test, the other procedures under investigation are sensitive to global damage of the visual field: (2) the second psychophysical measure, a full-field flicker test (FLI) [11], and the two electrophysiologic procedures, (3) the amplitude of a pattern reversal electroretinogram with a rapidly alternating black and white stimulus (PERG) [12] and (4) the peak latency of a visually evoked potential with a blue-on-yellow pattern presented in the onset-offset mode (VEP) [13].

The definition of glaucoma is based exclusively on the morphometric criterion 'damage of the optic nerve head'. The sensory 'gold standard', perimetry, is treated equivalent to the new procedures. Perimetry, as implemented in our study, yields differential light threshold measurements at 59 points of the 30° central visual field. The most important parameter obtained from these 59 single measurements is the perimetric mean defect (MD), which is the arithmetic mean of the 59 differences between individual local thresholds and age-adjusted norm values.

The measure for disease severity is given by the quantitative component of the optic nerve head inspection, the neuroretinal rim area (NRRA) [14]. This area represents the number of nerve fibres that are still intact.

3. STATISTICAL MODEL

In a recent publication, Leisenring and co-workers [1] proposed the use of logistic regression analysis for the determination of sensitivity and 1-specificity of diagnostic measurements. In the modelling of sensitivity, they included disease severity as covariate. Their approach was developed for the analysis of dependent data with multiple measurements from different or identical procedures obtained from the same subject. The approach is also applicable to independent data with only one measurement per subject. In this section, we describe the basic model of Leisenring *et al.* without referring to the dependent nature of the data. The treatment of correlations is addressed in the next section.

According to Leisenring *et al.* we distinguish two groups G of affected ($G=A$) and unaffected ($G=U$) individuals. The individuals $i=1,\ldots,n_G$ (from both groups $G=A,U$) are measured at sites $j=1,\ldots,J$. The index G is omitted if it is clear from the context whether n refers to n_A or to n_U. The diseased individuals are characterized further by a severity vector $S_i:=(S_{ij})_{j=1,\ldots,J}$, $i=1,\ldots,n_A$.

For each of both groups a different logistic model will be applied. For the first group, sensitivity, and for the second group specificity will be modelled. However, we shall only further investigate the model for affected individuals in our application. The same battery of K diagnostic tests will be applied for each individual. In our application, there are $J=2$ sites, the left and right eye of an individual, and $K=5$ diagnostic tests. It is not an essential assumption of our model to keep fixed the number of sites and diagnostic tests per individual. These restrictions only simplify notation. The outcome Y of each diagnostic test is dichotomous. Thus, $Y_{ijk}=1$ denotes a positive test result ('diseased') of individual i at site j under procedure k, while $Y_{ijk}=0$ corresponds to negative test results ('not diseased'); $G=A,U$; $i=1,\ldots,n_G$; $j=1,\ldots,J$; $k=1,\ldots,K$.

We set up two different logistic models for the binary outcome Y: The sensitivity, i.e. the probability of a positive test result Y_{ijk} in an affected individual, is considered to depend on the severity S_{ij} of the disease via a logistic regression model. In contrast, the probability of a positive test result in an unaffected individual, i.e. 1-specificity, is assumed to be independent of the severity measure in the control group. This assumption, however, may be crucial for several applications and we shall discuss this issue in Section 7. The diagnostic tests are modelled by $K-1$ dummy variables Z_k ($k=1,\ldots,K-1$) that indicate test type. Set $Z_{ijk}:=Z_k(ij)=1$ if procedure k is applied to individual i at site j and 0 otherwise. Since the severity of the disease may influence the sensitivities of the several tests differently, we include interaction terms $Z_{ijk}\cdot S_{ij}$ between test type and severity in the model of affected individuals ($G=A$). We ignore interaction terms of higher order in S_{ij}. Two groups of parameters are defined for affected individuals (β^A, γ^A): The vector

$$(\beta_0^A,\beta_1^A,\ldots,\beta_{K-1}^A)$$

refers to the sensitivity of diagnostic tests at severity $= 0$. The vector

$$(\gamma_0^A, \gamma_1^A, \ldots, \gamma_{K-1}^A)$$

is related to both the severity S_{ij} and the interaction between severity and diagnostic test. For unaffected individuals the vector β^U is defined analogously to β^A.

With these definitions, we obtain equations (1a) and (1b) for sensitivity in the affected population and for 1-specificity in the unaffected population, respectively:

$$\begin{aligned} &\operatorname{logit} P[Y_{ijk} = 1 \mid G_{ij} = A, S_{ij}] = \beta_0^A + \left(\sum_{l=1}^{K-1} \beta_l^A Z_{ijl} \right) + \gamma_0^A S_{ij} + \left(\sum_{l=1}^{K-1} \gamma_l^A Z_{ijl} S_{ij} \right) \\ &\quad i = 1, \ldots, n_A, \;\; j = 1, \ldots, J, \;\; k = 1, \ldots, K \end{aligned} \tag{1a}$$

and

$$\begin{aligned} &\operatorname{logit} P[Y_{ijk} = 1 \mid G_{ij} = U] = \beta_0^U + \left(\sum_{l=1}^{K-1} \beta_l^U Z_{ijk} \right) \\ &\quad i = 1, \ldots, n_U, \;\; j = 1, \ldots, J, \;\; k = 1, \ldots, K \end{aligned} \tag{1b}$$

Thus, the logit of the sensitivity of a diagnostic test k at a site with severity S is given by $\beta_0^A + \beta_k^A + (\gamma_0^A + \gamma_k^A) \cdot S$ if $k = 1, \ldots, K-1$ and by $\beta_0^A + \gamma_0^A \cdot S$ if $k = K$. The logit of the specificity of test k is given by $\beta_0^U + \beta_k^U$ if $k = 1, \ldots, K-1$ and by β_0^U if $k = K$. Note that the parameters γ_k $(k = 0, 1 \ldots, K-1)$ do not appear in equation (1b) because we assume that specificities are independent of severity S. Furthermore, we do not consider repeated measurements of the same eye using the same procedure. Additionally, we assume a stable disease status and severity of the disease during the diagnostic process. These restrictions, however, do not affect the validity of the approach and may be relaxed.

A more crucial restriction is that we only include probands with an identical status of left and right eye. We have used this strong restriction because it is very likely that the unaffected eye will be diseased in the future in individuals with one affected eye. Therefore, we do not aim to obtain a high 'specificity' in these eyes.

The comparison of specificities between different measurements is based on contrasts of the parameter vector β^U. The comparison of sensitivities, however, depends upon both parameter vectors β^A and γ^A. In the absence of interaction between disease severity and test type, the comparison of sensitivities depends on β^A alone. In this case, equal sensitivities of the several test procedures are equivalent to

$$\beta_1^A = \beta_2^A = \cdots = \beta_{K-1}^A = 0$$

Otherwise, for two procedures l and l' with

$$\gamma_l^A \neq \gamma_{l'}^A$$

the ranking according to sensitivity changes at the severity

$$S = \frac{(\beta_{l'}^A - \beta_l^A)}{(\gamma_l^A - \gamma_{l'}^A)}$$

if this value is in the possible range of S. If $l' = K$ and $\gamma_l^A \neq 0$ we obtain

$$S = \frac{-\beta_l^A}{\gamma_l^A}$$

In our example, test procedures are dichotomized by the definition of cut-off-points on continuous scales. Therefore, we are able to define the specificity equal to 80 per cent for all tests. This leads to $\beta_0^U = \text{logit}(0.2)$ and $\beta_k^U = 0$ for $k = 1, \ldots, K-1$. We slightly underestimate standard errors of β^A due to the fact that cut-off-points for the specificity of 80 per cent are constructed from the control sample and thus are subject to random variation.

The entire approach, however, is immediately transferable to intrinsic dichotomous diagnostic tests. In this situation, the specificity is not fixed but rather estimated from the data.

4. GEE APPROACHES TO MARGINAL REGRESSION MODELS

In this section we describe how to take into account the correlation of measurements obtained from the same individual in the framework of the models described in Section 3. Multiple measurements per individual may arise from (1) intraindividual comparisons of different diagnostic procedures, (2) examinations of multiple sites per individual (e.g. multiple tumour lesions, paired organs), and (3) from repeated measurements using identical procedures at the same site of an individual in order to reduce measurement error or study short-term or long-term variation.

Our data comprise a two-fold correlation structure: First, measurements are obtained using different procedures at the same eye; and second, both eyes of the same subject are examined. We do not include multiple measurements using the same procedure at the same site of an individual on different visits.

The logistic regression models as defined by equations (1a) and (1b) belong to the class of generalized linear models. For these models, Liang and Zeger [2] proposed the generalized estimating equations (GEE) to adjust for correlations. As noted by these authors, GEE may also be thought of as an extension of quasi-likelihood to the case where the second moments are not fully specified in terms of the expectation but rather additional correlation parameters have to be estimated.

The approach of Liang and Zeger has been discussed intensively in the literature and several extensions have been proposed (for an overview, see Ziegler *et al.* [15]). In the following, we sketch the GEE as required in our application.

According to the GEE method, the parameters of the mean structure in equations (1a) and (1b) may be estimated consistently by solving the equations

$$\frac{1}{n}\sum_{i=1}^{n} D_i^t V_i^{-1}(y_i - \mu_i(\beta, \gamma)) = 0 \tag{2}$$

where y_i denotes the $(J \cdot K)$ vector of the dichotomous test results for subject i and μ_i is the $(J \cdot K)$ vector of sensitivities or 1-specificities, respectively. D_i is the matrix of first derivatives of these probabilities with respect to the parameter vectors (β^A, γ^A) and β^U, respectively. Its dimension is $(J \cdot K) \cdot (2K)$ if $G = A$ and $(J \cdot K) \cdot K$ if $G = U$. V_i denotes the $(J \cdot K) \cdot (J \cdot K)$ working covariance matrix of y_i, which has to be specified by the investigator.

Under suitable regularity conditions, the solution of (2) yields consistent estimates of (β, γ) even if the working covariance V_i is misspecified. However, V_i should be a good guess of the true covariance matrix, as misspecifications lead to inefficient weighing of observations in the GEE (2). In Section 5 we discuss specific choices of V_i in our application.

The variances of the estimators may be estimated consistently by the robust estimator of variance [2]. Liang and Zeger proposed to estimate the diagonal elements of V_i from the variance function of the generalized linear model, equalling to $\mu \cdot (1-\mu)$ in our case, and the correlations may be obtained from simple moment estimators using residuals

$$s_{ijk} = \frac{y_{ijk} - \mu_{ijk}}{\sqrt{v_{ijk}}}$$

where v_{ijk} denotes the diagonal element $(V_i)_{jk,jk}$ of V_i. Alternatively, a system of equations similar to (2) can be derived for the association parameters [3]: Let ρ_i denote the vectorized correlation matrix of Y_i, containing $[(J \cdot K) \cdot (J \cdot K - 1)]/2$ non-redundant elements and let z_i be the related $[(J \cdot K) \cdot (J \cdot K - 1)]/2$-dimensional vector of the cross-products $z_{ijk,ij'k'}((jk) \neq (j'k'))$

$$z_{ijk,ij'k'} = s_{ijk} \cdot s_{ij'k'} = \frac{(y_{ijk} - \mu_{ijk}) \cdot (y_{ij'k'} - \mu_{ij'k'})}{\sqrt{v_{ijk}} \cdot \sqrt{v_{ij'k'}}}$$

with $j = 1, \ldots, J, k = 1, \ldots, K$. Then, analogously to V_i, a working covariance matrix W_i of dimension $[(J \cdot K) \cdot (J \cdot K - 1)]/2$ needs to be introduced.

Furthermore, let $\rho_{ijk,ij'k'}$ be the correlation of Y_{ijk} and $Y_{ij'k'}$. This correlation may depend on the disease severity vector of subject i, and be parameterized by a vector α, i.e.

$$\rho_{ijk,ij'k'} = \rho(\alpha, S_i, (j,k), (j',k')) \tag{3}$$

This correlation is be modelled using the artanh association link function of a linear predictor with coefficient vector α for suitably chosen covariates (cf. Section 5) [15–17]. Then the estimation of α may be performed analogous to (2) by the estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \rho_i^t}{\partial \alpha} \cdot W_i^{-1} \cdot (z_i - \rho_i) = 0 \tag{4}$$

A diagonal working covariance matrix W_i is usually used in applications, which is termed matrix for applications [15]. The resulting estimators of the Prentice estimating equations (PEE) (2) and (4) are consistent and jointly asymptotically normally distributed under suitable regularity assumptions and if equations (1a) (alternatively (1b)) and (3) are correctly specified [3].

The robust variance estimator for the PEE has been given by Prentice [3]. The approach has been implemented in the freely available software package MAREG [18] which has been used in our analyses. MAREG can be obtained without charge from the WEB http://www.stat.uni-muenchen.de/~andreas/mareg/winmareg.html. It is important to note that—in contrast to the linear model—for binary data using the logistic regression model the scale parameter is set to 1 and need not be estimated [19, 20].

5. THE WORKING CORRELATION MATRICES

In the second paragraph of Section 4 we have described the two-fold association structure underlying our data, i.e. measurements of different procedures in the same eye and measurements in both eyes of the same subject. In our approach, we are interested only in a specific part of this structure, the 'within eyes' association. Thus, we define a 'partial working independence' for left and right eyes but allow a fully parameterized structure for the correlations of several diagnostic tests on the same eye. In this section, we give the mathematical outline of this approach; in the discussion we justify from the clinical point of view.

The use of this specific structure of the working correlation matrix is motivated by an implicit assumption underlying the GEE method. Pepe and Anderson [21] have claimed that in equation (2) the conditional expectation μ_{ijk} of y_{ijk} given its specific covariates x_{ijk} (in our notation Z_{ijk} and S_{ij}) needs to be independent from covariate values $x_{ij'k'}$ of other observations $y_{ij'k'}$ in the same cluster i if these are not included in the mean structure. This assumption, however, is not necessary if the working correlation matrix is diagonal (independence estimating equations, IEE).

In our notation, the assumption reads as

$$P[Y_{ijk}=1 \mid G_i, S_{ij}, Z_{ijk}] = P[Y_{ijk}=1 \mid (G_i, (S_{ih})_{h=1,\ldots,J}, (Z_{ihl})_{h=1,\ldots,J;l=1,\ldots,K-1})] \tag{5}$$

It can be easily seen that this assumption is relevant only for disease severity S_{ij} ($j=1,\ldots,J$) since the true status G is subject related in our study and Z_{ihl} ($h=1,\ldots,J; l=1,\ldots,K-1$) are test specific dummy variables.

The assumption is trivially fulfilled if cluster-constant covariates are considered exclusively. It is also fulfilled in random effect models of the form

$$\mu_{ijk} = \Phi(\eta_i, S_{ij}, Z_{ijk})$$

with a latent random variable η that is independent of the covariates. The assumption could therefore be ignored in our example if severity was cluster constant.

On the other hand, it would be fulfilled if disease severity of the fellow eye contained no additional information about the sensitivity of a diagnostic test for the selected eye after adjustment for disease severity of the selected eye. Especially, it could be met if a latent variable η could be assumed that caused the association of measurements in both eyes and was independent from disease severity. Unfortunately, none of these conditions can be justified in our example:

First, severity clearly is not cluster (i.e. subject-) constant.

Second, if the association of measurements at different sites of paired organs could be explained by a latent random variable η, then this latent variable should be correlated with disease severity. In our example, the neuroretinal rim area of the fellow eye may well contain information about the sensitivity of a diagnostic procedure for the selected eye, even in case of knowledge of the rim area of the selected eye.

This argument which is based on the possible interpretations of the latent variable η and on ophthalmologic considerations is enlarged upon in the discussion section.

Third, we cannot reduce the model to diagonal covariance matrices V_i ($i=1,\ldots,n$), as we are interested in the association structure of several diagnostic tests. However, we are not

interested in investigating the correlation of left and right eyes of the same subject, but rather the correlation of several measurements in the same eye. In our setting the correlation of left and right eyes is considered to be nuisance. Therefore, we use the working assumption that blocks, i.e. pairs of measurements in different eyes, are independent. However, different tests at the same eye may be related via the cluster specific severity S_i.

Therefore, in the system of equations (2) we use a working correlation matrix in which between eyes correlations have been fixed to zero. Terms in equation (4) corresponding to correlations from measurements between eyes vanish since all derivatives $\partial\rho/\partial\alpha$ are equal to zero for those correlations and the matrix W_i is diagonal.

In the general framework with J sites per subject and K measurements per site, this leads to the block diagonal working correlation matrix

$$V_i = \begin{pmatrix} V_{i1} & 0 & \dots & 0 \\ 0 & V_{i2} & \dots & 0 \\ . & . & . & . \\ 0 & 0 & \dots & V_{iJ} \end{pmatrix}$$

with the dimension of the blocks equal to K. In our application with $J = 2$, we assume an identical parameterization and identical parameter values for left and right eyes.

We show in the appendix that this choice of the covariance matrix leads to consistent parameter estimates for the mean structure and the relevant terms of the association structure even if the assumption of Pepe and Anderson, i.e. equation (5), is not fulfilled. The argument follows the same line as given for IEE by Pepe and Anderson. These authors showed that specific terms in equation (2) violate the consistency of the parameter estimates. These terms vanish if IEE are used, as V_i^{-1} is diagonal for this choice of V_i. However, we argue that it suffices to have only those entries of V_i^{-1} equal to zero, which are related to the crucial terms in equation (2).

In the following, we use the artanh function as association link function to connect the linear predictor δ to the correlation coefficient ρ [16]:

$$\delta = \ln\left[\frac{(1+\rho)}{(1-\rho)}\right]$$

In the patient group, we model the correlation structure using linear terms of the neuroretinal rim area. Therefore, the general equation for the association parameter δ is

$$\delta_{kl} = \alpha_{kl} + \phi_{kl} \cdot S \quad (k, l = 1, \dots, K, k \neq l) \tag{6}$$

Thus, $K \cdot (K-1)$ association parameters need to be estimated. This number equals 20 in our example where 5 tests are employed.

Additionally (model 6 in the results section), we include site-related correlations in the association structure. Therefore, the off-diagonal blocks of V_i also have to be estimated and the correlations of measurements between left and right eyes are modelled for identical and different diagnostic tests. According to Pepe and Anderson, equation (1a) needs to be extended. Thus, the severity S^f of the fellow eye has to be included in the linear predictor of the mean

structure:

$$\begin{aligned}&\operatorname{logit} P[Y_{ijk}=1 \mid G_i=A, S_{ij}]\\&\quad=\beta_0^A+\left(\sum_{l=1}^{K-1}\beta_l^A Z_{ijl}\right)+\gamma_0^A S_{ij}+\left(\sum_{l=1}^{K-1}\gamma_l^A Z_{ijl} S_{ij}\right)+\gamma_0^{f,A} S_{ij}^f\end{aligned} \tag{7}$$

In this model, a more complex association structure is required than the set of equations (6):

First, in the equations concerning measurements in identical eyes, we include as a new covariate the severity of the fellow eye, S^f, with coefficient α_0^f.

Second, new equations concerning measurements in different eyes have to be defined. In these equations, the parameters α_{kk} $(k=1,\ldots,K)$ correspond to the association of identical measurements in different eyes. A constant α_1 is included to adjust for the fact that measurement correlations α_{kl} $(k\neq l)$ should be smaller in different eyes than in identical eyes. Finally, we cannot speak of the selected or the fellow eye in this context, thus coefficients α_2 and α_3 are included that correspond to the mean severity of both eyes and their absolute difference. The severity of the left eye will be denoted by S^l and the severity of the right eye will be denoted by S^r.

To achieve some parsimony, we do not include interaction terms between the pairs of diagnostic procedures and S in this model. This means that the influence of the rim area on the correlation between measurement results in the patient group is assumed to be identical for all pairs of diagnostic procedures. Finally, we end up with model equations

$$\delta_{kl}=\alpha_{kl}+\alpha_0\cdot S+\alpha_0^f\cdot S^f \quad k,l=1,\ldots,K, k\neq l \tag{8a}$$

for the diagnostic tests in the same eye and

$$\delta'_{kl}=\alpha_{kl}-\alpha_1+\alpha_2\cdot(S^l+S^r)/2+\alpha_3\cdot(|S^l-S^r|) \quad k,l=1,\ldots,K \tag{8b}$$

for the correlation of diagnostic tests in different eyes. In our example, 20 parameters have to be estimated. In the control group we use simple moment estimators for the correlations. We assume the correlations to be independent of the severity indicator in this group.

6. RESULTS

In this section we present the results of our clinical example using the mean and association structures as discussed in Sections 3 and 5. In total, we examine six different models of increasing complexity of both structures. Throughout this section, we use the notation as given in equations (1a) and (7) for the mean structure, and in equations (6), (8a), and (8b) for the association structure.

In all analyses, 2878 test results from 297 patients and 1691 test results from 176 control subjects entered the evaluation. As mentioned before, only subjects with identical disease status in left and right eyes were included.

The severity measure 'neuroretinal rim area' was standardized; disease damage has increasing values on the new scale. (Transformation equation: $\text{nrra}_{\text{transformed}}=-(\text{nrra}_{\text{raw}}-1.048)/0.336\,(\text{mm}^2)$).

Table I. Parameter estimates (log ORs) and p-values (in parentheses) for the different mean structure models for sensitivities.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
QIC	3955.13	3603.24	3586.81	3589.40	3589.69	3584.93
Neuroretinal rim area*	—	0.78 (<0.001)	0.67 (<0.001)	0.72 (<0.001)	0.73 (<0.001)	0.50 (<0.001)
Constant†	−0.12 (0.25)	−0.10 (0.35)	−0.11 (0.33)	−0.10 (0.39)	−0.089 (0.42)	−0.057 (0.61)
Perimetric mean defect	0.32 (0.009)	0.35 (0.010)	0.41 (0.003)	0.44 (0.002)	0.44 (0.002)	0.39 (0.009)
Localized contrast sensitivity	0.49 (<0.001)	0.52 (<0.001)	0.55 (<0.001)	0.54 (<0.001)	0.54 (<0.001)	0.53 (<0.001)
Flicker test	0.35 (0.006)	0.37 (0.010)	0.35 (0.011)	0.34 (0.015)	0.33 (0.017)	0.22 (0.13)
Amplitude pattern reversal ERG	−0.20 (0.10)	−0.26 (0.058)	−0.25 (0.061)	−0.25 (0.056)	−0.26 (0.048)	−0.30 (0.022)
Neuroretinal rim area fellow eye	—	—	—	—	—	0.29 (<0.001)
Interaction of neuroretinal rim area with						
Perimetric mean defect	—	—	0.53 (<0.001)	0.56 (<0.001)	0.57 (<0.001)	0.59 (<0.001)
Localized contrast sensitivity	—	—	0.24 (0.10)	0.25 (0.11)	0.24 (0.13)	0.25 (0.055)
Flicker test	—	—	−0.15 (0.29)	−0.17 (0.24)	−0.18 (0.21)	−0.088 (0.48)
Amplitude pattern reversal ERG	—	—	0.042 (0.76)	0.032 (0.82)	−0.004 (0.98)	0.040 (0.76)

*Neuroretinal rim area was standardized (mean 0, standard deviation 1, increasing damage for increasing numerical values) for all models.
†Peak latency of visual evoked potential was chosen as reference category for all models. *Model* 1: IEE ignoring disease severity. *Model* 2: IEE including disease severity (main effect). *Model* 3: IEE including disease severity (main effect and interaction with diagnostic procedures). *Model* 4: PEE with block-diagonal covariance matrix, mean structure identical to model 3, for association structure cf. Table II. *Model* 5: PEE with block-diagonal covariance matrix, mean structure identical to model 3, for association structure cf. Table II. *Model* 6: PEE with full covariance matrix, mean structure includes disease severity of the fellow eye (main effect), for association structure cf. Table II.

Additionally to the parameter estimates and standard errors we present the QIC-criterion as proposed by Pan [22]. This criterion is based on quasi-likelihood and a penalty term. The penalty term is based on the robust estimator of variance [2] and the model based variance estimator of the same mean structure but ignoring correlations within clusters.

In *model* 1 (Table I, second column), the mean structure is defined solely by dummy variables related to the different diagnostic procedures; we do not include severity of disease. Thus, in equation (1a) only the β parameters are estimated. IEE are used, and no parameters of the association structure need to be considered. With respect to the different sensitivities, this model is equivalent to the simple separate estimation of sensitivities for the different

measurements. Inspection of parameters and p-values in Table I reveals a significant superiority of the psychophysical over the electrophysiologic tests: The sensitivity of the three psychophysical tests, 'perimetric mean defect', 'flicker test' and 'contrast sensitivity', is significantly higher than both electrophysiologic procedures, 'amplitude of the electroretinogram' and 'peak latency of the visually evoked potential' (Wald test).

In *model* 2, severity of disease is added to the mean structure. However, interaction between disease severity and the diagnostic procedures is ignored. In equation (1a), parameters β and γ_0^A need to be estimated. As can be seen in column 2 of Table I, sensitivity increases significantly with progression of disease damage. An increase of one unit on the standardized scale yields an odds ratio of 2.2 (95 per cent CI: 1.9–2.5) for the increase in sensitivity. Thus, severity needs to be included in the mean structure. Again, superiority of psychophysical tests over electrophysiologic ones is supported. In this model and in model 3, no parameters for the association structure have been incorporated.

In *model* 3 however, the interaction between severity and type of measurement has been added to the mean structure. This model corresponds to equation (1a), with all parameters β and γ being estimated. As can be seen in Table I, the interaction between perimetric mean defect and disease severity is significant. Figure 1 gives the impression that this procedure is valid especially for moderately and severely progressed disease but not for early cases. The flicker test shows the highest sensitivity for 'early' cases but a less prominent increase in sensitivity with increasing disease damage. This procedure seems to be especially useful in early cases of glaucoma.

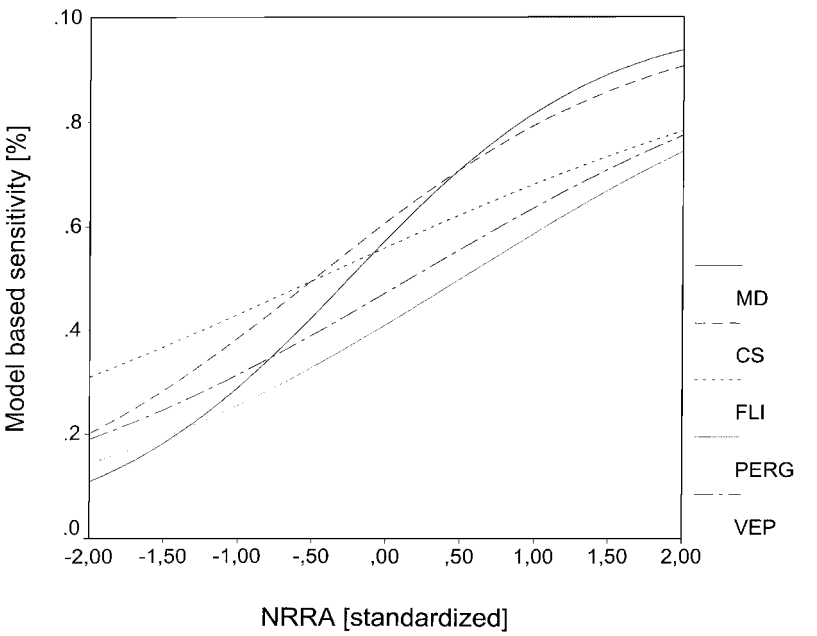

Figure 1. Model 3—Mean structure. MD perimetric mean defect, CS: localized contrast sensitivity, FLI: flicker test, PERG: amplitude pattern reversal electroretinogram, VEP: peak latency of visual evoked potential.

In *models* 4 and 5 the mean structure is identical to the one in *model* 3.

In Table II we present the observed correlations of dichotomized test results within the patient group and within the control group (columns 1 and 2). In the remaining columns we display estimates of the model-based association structures for models 4 and 5. Three observations can be drawn from the first two columns of Table II. Firstly, correlations are always higher in the patient group than in the control group. Secondly, lateral correlations of identical measurements are generally stronger than correlations of different measurements in identical eyes. Except for PERG, all correlations of identical measurements in left and right eyes exceed 0.44 (Table II, bottom). In contrast, all correlations of different measurements in the same eye are <0.40 except for the correlation between MD and CS. Thirdly, in both patients and in controls, correlations between psychophysical tests are generally higher than those between psychophysical and electrophysiologic tests and those between electrophysiologic tests.

In *model* 4 an association structure has been chosen according to equation (6), but without interaction terms between measurements and neuroretinal rim area: For every pair of measurements, a specific correlation has been estimated. Additionally, a main effect of severity (neuroretinal rim area, cf. legend of Table II) has been included. Estimation was performed using PEE according to equation (4). Results are shown in Table II. A decrease in correlations between measurements with increasing disease severity is revealed. However, the most important observation is that correlations within the patient group decrease substantially as compared with column 2. Thus, conditioning on disease severity shows that part of the association between measurements in the same eye is explained by disease severity. Taking into account that the neuroretinal rim area is not a 'gold standard' for disease severity, it may be speculated that some pairs of measurements are indeed independent, conditional on the true disease severity. This idea is explained further in the discussion.

In *model* 5, the association structure contains interaction terms of pairs of measurements with neuroretinal rim area, i.e. the full structure according to equation (6) has been chosen. The most prominent decrease in correlations for increasing disease severity is observed for the correlations of perimetric mean defect with the flicker test, with localized contrast sensitivity, and with the peak latency of the visually evoked potential. Figures 2 and 3 explain the specific pattern of correlations between the psychophysical tests: The high overall correlation of test results is mainly due to the results from patients in the early stage of the disease, whereas for the electrophysiologic procedures, especially PERG, correlations with psychophysical tests are smaller in the whole range of disease severity.

Model 6 is the only model in which correlations between measurements in different eyes from the same subject are considered. The mean structure includes the main effects of diagnostic procedures, the interaction with neuroretinal rim area and the rim area in the fellow eye.

The association structure has been chosen according to equations (8a) and (8b). Main effects are included for each pair of diagnostic measurements. To adjust for the fact that correlations of different measurements from the same eye should be different from those from different eyes, an additional 'penalty' term has been added for such measurements.

For pairs of measurements from the same eye, the neuroretinal rim area in the selected eye and in the fellow eye has been included as covariate. For pairs of measurements from different eyes, the mean and (absolute) difference of both rim areas has been included as covariate. No interaction between pairs of measurements and rim area has been included.

Table II. Parameter estimates, retransformed correlations [in brackets], and *p*-values (in parentheses) for the different association structure models of pairs of measurements.

	Ignoring severity		Model 4	Model 5		Model 6
Test pair	Controls	Patients	Main effects only	Main effect	Inter-action	Main effects only
Correlations of different measurements at the same eye						
MD	—	—	0.72	0.75	−0.33	0.61
CS	[0.22]	[0.47]	[0.35]	[0.36]	—	[0.30]
	—	—	(<0.001)	(<0.001)	(<0.001)	(<0.001)
MD	—	—	0.69	0.71	−0.25	0.61
FLI	[0.23]	0.39	[0.33]	[0.34]	—	[0.29]
	—	—	(<0.001)	(<0.001)	(0.006)	(<0.001)
MD	—	—	0.30	0.29	0.026	0.16
PERG	0.24	0.27	[0.15]	[0.15]	—	[0.08]
	—	—	(0.002)	(0.086)	(0.78)	(0.057)
MD	—	—	0.40	0.40	−0.24	0.26
VEP	0.18	0.30	[0.20]	[0.20]	—	[0.13]
	—	—	(<0.001)	(<0.001)	(0.016)	(0.004)
CS	—	—	0.61	0.60	[−0.063]	0.51
FLI	0.07	0.34	[0.29]	[0.29]	—	[0.25]
	—	—	(<0.001)	(<0.001)	(0.49)	(<0.001)
CS	—	—	0.25	0.25	−0.059	0.21
PERG	0.16	0.23	[0.12]	[0.12]	—	[0.11]
	—	—	(0.005)	(0.005)	(0.48)	(0.010)
CS	—	—	0.25	0.25	−0.21	0.14
VEP	0.05	0.22	[0.12]	[0.12]	—	[0.07]
	—	—	(0.010)	(0.010)	(0.023)	(0.10)
FLI	—	—	0.38	0.38	−0.15	0.30
PERG	0.07	0.25	[0.19]	[0.19]	—	[0.15]
	—	—	(<0.001)	(<0.001)	(0.072)	(<0.001)
FLI	—	—	0.36	0.36	−0.085	0.28
VEP	0.12	0.22	[0.18]	[0.18]	—	[0.14]
	—	—	(<0.001)	(<0.001)	(0.34)	(0.002)
PERG	—	—	0.38	0.38	−0.10	0.31
VEP	0.22	0.27	[0.19]	[0.19]	—	[0.15]
	—	—	(<0.001)	(<0.001)	(0.34)	(<0.001)
Correlations of the same measurement in different eyes						
MD	—	—				0.99
MD	0.47	0.57	—	—	—	[0.48]
	—	—				(<0.001)
CS	—	—				0.64
CS	0.34	0.45	—	—	—	[0.31]
	—	—				(<0.001)

Table II. *Continued.*

FLI	—	—				0.99
FLI	0.62	0.52	—	—	—	[0.46]
	—	—				(<0.001)
PERG	—	—				0.48
PERG	0.17	0.37	—	—	—	[0.23]
	—	—				(<0.001)
VEP	—	—				0.95
VEP	0.57	0.53	—	—	—	[0.44]
	—	—				(<0.001)

NRRA: Neuroretinal rim area; MD: Perimetric mean defect; CS: Localized contrast sensitivity; FLI: Flicker test; PERG: Amplitude of pattern reversal electroretinogram; VEP: Peak latency of blue on yellow VEP.
Column 1 (*Controls*): observed correlations of dychotomized test results without adjustment for covariates. *Column 2* (*Patients*): observed correlations of dychotomized test results without adjustment for covariates, these are identical to moment estimators for correlations of residuals in model 1. *Column 3* (*Model 4*): PEE with block-diagonal matrix according to equation (6) of Section 5. The parameter for the main effect of neuroretinal rim area at the same eye was equal to -0.14 ($p = 0.004$), no interactions between neuroretinal rima area and pairs of measurements were taken into account. All entries correspond to correlations in the same eye. *Columns 4 and 5* (*Model 5*): PEE with blockdiagonal matrix according to equation (6) of Section 5, additionally interactions of neuroretinal rim area with pairs of measurements are included. *Column 6* (*Model 6*): PEE with full covariance matrix according to equations (8a), (8b) of Section 5. For correlations in identical eyes, the parameter of the neuroretinal rim area of the same eye, α_0, was equal to -0.049 ($p = 0.27$), the parameter of the neuroretinal rim area of the fellow eye, α_0^f was equal to -0.051 ($p = 0.22$). For correlations in different eyes, the penalty term for correlations of different measurements in different eyes, α_1, was 0.06 ($p = 0.34$), the parameter of the average neuroretinal rim area, α_2, was equal to -0.15 ($p = 0.001$), the parameter for the absolute difference of the difference of both mean areas was 0.16 ($p = 0.32$). For *models* 4, 5, *and* 6, *p*-values and retransformed correlations are given. These correlations correspond to estimates at the point 'standardized neuroretinal rim area = 0' (models 4 and 5), or to the origin of the subscores in equations (8a) and (8b) pertaining to continuous covariates (model 6).

In contrast to models 1 to 5 which reached the convergence criterion ($< 10^{-10}$) after 10 to 18 iterations in MAREG [18], model 6 required 37 iterations.

The neuroretinal rim area of the fellow eye provides significant information about the expected sensitivity in the selected eye, even after adjustment of the rim area in the selected eye. The sign of the coefficient is positive. Concerning the association structure, the correlations of identical measurements in left and right eyes again decrease with increasing disease severity.

7. DISCUSSION

It has long been recognized that the sensitivity of diagnostic measurements may depend on the severity of disease in individual subjects or populations to be examined [23]. To take into account this possible dependency, Leisenring and co-workers [1] developed a logistic regression approach that incorporates models of sensitivity with disease severity as covariate in the mean structure for sensitivity. Moreover, their model covers repeated measurements at different or identical sites in one subject and intraindividual comparisons of several test

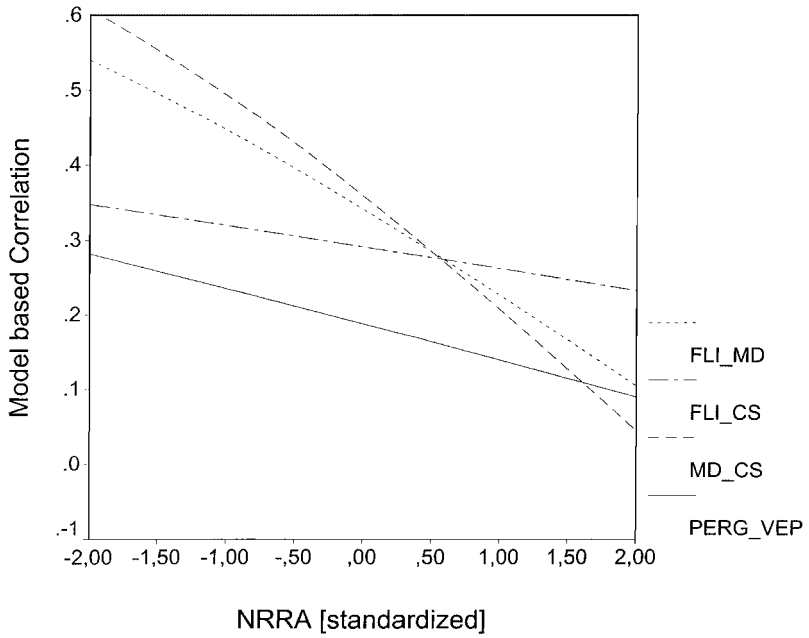

Figure 2. Model 5—Association Structure *within* psychophysical and electrophysiologic measurements. For abbreviations cf. to Figure 1.

procedures. To adjust for correlation between measurements performed in the same subject, the authors applied IEE [2].

An alternative approach for the correlated test results from the same individual at identical or different eyes would be the generalized linear mixed model using random effects. A discussion of random versus marginal models can be found e.g. in the textbook of Diggle and colleagues [24]. The interpretation of parameters for the dummy variables in a random effects model would be subject specific as compared to population averaged in our marginal model. Thus random effect model parameters describe the effect by changing from one procedure to another in a specific subject. This might be of interest if all procedures are available and the clinician selects a specific procedure for each individual separately. If, however, a procedure needs to be selected for routine applications in patients, the population specific parameters would be of primary interest. Furthermore, the comparison of sensitivities obtained in different populations is no longer possible if random effects are used instead of adjustments for disease severity.

Leisenring and co-workers [1] demonstrated the usefulness of the population averaged approach for comparing the validity of different diagnostic tests. In their work, correlations between measurements were treated as nuisance parameters; and no conclusions could be drawn concerning the association structure. However, if the combination of tests in a diagnostic program is to be optimized, not only the validity of the individual tests is of interest but also their association. The stepwise variable selection procedures applied e.g. in logistic regression analysis take this fact implicitly into account: The incremental gain or loss in diagnostic information depends on both the validity and the association structure of the

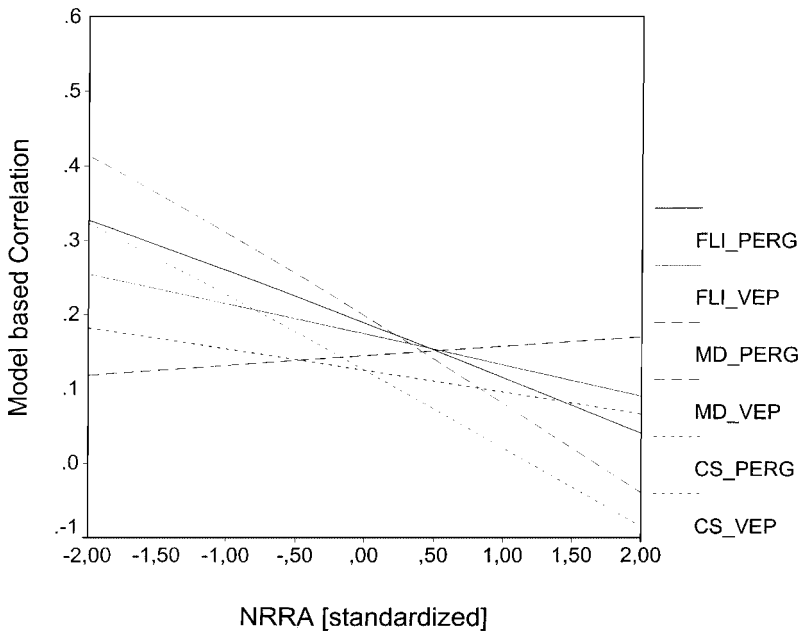

Figure 3. Model 5—Association Structure *between* psychophysical and electrophysiologic measurements. For abbreviations cf. to Figure 1.

measurements under investigation. Nevertheless, these techniques do not permit inference about these associations.

In the present work, we have extended the approach developed by Leisenring and colleagues [1] in order to jointly investigate both the mean and the association structure of sensitivities from different diagnostic tests. Our approach—in contrast to Leisenrings'—permits inference about the correlations of interest for multiple diagnostic measurements in identical subjects. In this context, we substantially weakened a crucial key assumption of GEE that has been described by Pepe and Anderson [21].

In the framework of the Leisenring model for paired organs, the assumption of Pepe and Anderson states that IEE should be used if the sensitivity at the selected organ depends on disease severity of the fellow organ conditional on disease severity at the selected organ. Alternatively, the fellow organ's disease severity should be included as covariate. Otherwise the parameter estimates of the mean structure need not be consistent.

This situation was clearly present in our ophthalmologic example: In model 6 of Section 6, a p-value of less than 0.05 was observed for the fellow eye's disease severity. However, we could show that the assumption of Pepe and Anderson may be relaxed and, instead of IEE, a block diagonal structure of the working covariance matrix suffices to ensure consistency of parameter estimates in the mean structure. In general, structures of this type of working matrices correspond to situations in which specific associations are of interest whereas others are regarded as a nuisance.

In a recent study, Pan and colleagues [25] investigated the case of mutually independent covariates in the AR(1) model. They gave explicit formulae of the bias term for IEE and

general working covariance structures. They showed that the bias tends toward zero with an increasing number of subjects for IEE as it is asymptotically proportional to the covariance of the covariates which is assumed to be zero in the AR(1) model. In contrast, for GEE as opposed to IEE, this bias need not vanish asymptotically. Pan and colleagues [25] also showed that the bias term depends on the sum of non-diagonal elements of the inverted working covariance matrix. This is in accordance with our results as a non-singular block-diagonal matrix has a block-diagonal inverse. However, we did not assume mutually independent covariates.

In our application, we used 297 independent clusters for estimation which may be termed a moderate sample size. However, the fact that our data is nearly balanced with respect to the included dummy variables and that the measure of severity is eye specific support the argument that the sample size is reasonably large for our analysis. Furthermore, parameters for left and right eyes were defined to be identical. Nevertheless, convergence for model 6 was fairly slow. This model thus needs to be interpreted cautiously.

In our example, the tests were derived from dichotomized continuous measurements. Therefore, the evaluation of specificities could be simplified by choosing appropriate cut-off points of equal specificity for all procedures. Furthermore, independence of sensory measurements from the severity measure 'neuroretinal rim area' (NRRA) was assumed within controls. This assumption need not be justified in general: If the severity measure within the control group reflects the biological variability of an underlying trait, diagnostic tests may be associated with this trait in the controls, too. A normal subject for whom this trait lies in the overlap of the normal and the pathologic ranges of values may be likely to exhibit pathologic values in the diagnostic tests as well. However, in our application, the assumption of the independence of sensory measures from the neuroretinal rim area had been shown in a previous study [26] and seems therefore to be well justified.

The comparison of QIC-values suggests the use of model 3 or model 6 (Table I). However, models 4 and 5 revealed only slightly increased QIC-values as compared to model 3. Furthermore, parameter estimates of model 6 should be cautiously interpreted due to the relative slow convergence. As we focus on the interpretation of effects, not only on the identification of the most parsimonious model, examination of models 4 and 5 is of interest, too.

The analysis of the mean structure within the patient group revealed a strong dependency of sensitivities on this severity measure. Additionally, interactions between severity and diagnostic tests were present. Therefore, the suggestion that different tests might be useful at different stages of disease could be confirmed. Especially, the use of perimetry as 'gold standard' seems to be justified only for intermediate or late-stage disease. From the ophthalmologic point of view, the use of the mean defect in the perimetric examination ignores relevant information: the spatial and short-term variance components ('corrected loss variance' and 'short-term fluctuation' in perimetric terminology [7]) and the qualitative, investigator-dependent inspection of the graphical display of measurement results. In summary, the results from the mean structure showed the usefulness of the approach by Leisenring *et al.* [1] for comparing different diagnostic tests.

The associations observed in the control group are in accordance with the specific implementations of diagnostic measurements: In the absence of any dependency of the measurement results on the severity measure, the association should be produced by measurement error or other 'true' factors such as concentration. It seems plausible that psychophysical tests share such common factors, whereas the electrophysiologic procedures should reveal independent

test results in our example: Visual evoked potentials are localized in a central area of the brain whereas the potentials measured by electroretinograms are generated in special layers of the retina. The pattern of correlations found in the control group matches these assumptions exactly.

Within the patient group, stronger associations were found for all pairs of diagnostic measurements. Again, the psychophysical measurements revealed higher correlations than the electrophysiologic ones. However, after adjusting for disease severity as measured by the neuroretinal rim area, these associations decreased substantially. Additionally, associations decreased noticeably with increasing disease severity.

These observations are in accordance with a confirmatory factor analysis applied to data from a sample of glaucoma patients using quantitative measurement values [27, 28]. In this analysis, a one-factor model of glaucoma with dependent error terms of psychophysical measurements was fitted successfully.

In conclusion, analyses like the ones presented may improve the diagnosis of individual subjects considerably. Furthermore, they give insight into the mechanisms of diagnostic procedures and can therefore contribute to the further development of these procedures or even give rise to new measurement techniques.

APPENDIX A

We show that the implicit assumption of GEE (except for IEE) noted by Pepe and Anderson [21]—conditional independence of the expectation μ_{ijk} of y_{ijk} given its specific covariates x_{ijk} from covariate values $x_{ij'k'}$ of other observations $y_{ij'k'}$ in the same cluster i—may be relaxed slightly. Conditional independence is necessary only for those covariate values at fellow sites for which corresponding elements of the inverse of the working correlation matrix are different from zero.

We examine a more general situation than presented in the foregoing part of the paper and change the notation slightly. Let there be $i = 1, \ldots, n$ individuals with an arbitrary number of $j = 1, \ldots, n_i$ observations from individual i. Each observation is related to a covariate row vector X_{ij} of fixed length L. For each individual the covariate matrix X_i has dimension $n_i \cdot L$ with rows corresponding to observations and columns corresponding to covariates. X_{ijl} $(i = 1, \ldots, n;$ $j = 1, \ldots, n_i;$ $l = 1, \ldots, L)$ is the value of covariate l from observation j of individual i. The dependent variable Y is related to observations within individuals: the expectation of Y_{ij} is given by $g(X_{ij} \cdot \beta)$, where g is the link function of the underlying generalized linear model, and β is an L-dimensional parameter vector.

The estimation of equation (2) has the form

$$\frac{1}{n}\sum_{i=1}^{n} S_i(\hat{\beta}) := \frac{1}{n}\sum_{i=1}^{n} \hat{D}_i^t \hat{V}_i^{-1}(y_i - \mu_i(\hat{\beta})) = 0 \tag{A1}$$

D_i is defined as $\partial\mu_i/\partial\beta^t$ and is equal to

$$\frac{\partial \mu_i}{\partial (X_i\beta)^t} \cdot \frac{\partial (X_i\beta)}{\partial \beta^t}$$

The first factor is a $n_i \cdot n_i$ diagonal matrix. It will be denoted by F_i with entries $f_{ijj} = \partial\mu_{ij}/\partial(X_{ij}\beta)$ $(j = 1, \ldots, n_i)$. The second factor $\partial(X_i\beta)/\partial\beta^t$ is equal to X_i. Therefore, equation (A1) may be written as

$$\frac{1}{n}\sum_{i=1}^{n} S_i(\beta) = \frac{1}{n}\sum_{i=1}^{n} X_i^t F_i V_i^{-1}(y_i - \mu_i(\beta)) = 0 \tag{A2}$$

Every single term and the sum in equation (A2) is an L-dimensional vector with every component corresponding to one component of β. Following the argument of Pepe and Anderson [21], consistency of β is fulfilled if the expectation of every single term $S_i(\beta)$ in equation (A2) is zero. This expectation equals

$$E\left(\sum_{k=1}^{n_i}\left(\sum_{j=1}^{n_i} f_{ijj} \cdot X_{ijl} \cdot \omega_{ijk} \cdot E[Y_{ik} - \mu_{ik} \mid (X_{ijl}), j = 1, \ldots, n_i]\right)\right) \quad (l = 1, \ldots, L) \tag{A3}$$

with ω_{ijk} being the jkth element of V_i^{-1}. According to the model assumption, the expectation of $Y_{ik} - \mu_{ik}$ conditional on X_{ikl} $(l = 1, \ldots, L)$ is equal to zero. It is unequal to zero if the expectation of Y_{ik} changes after conditioning additionally on the remaining X_{ijl} $(j = 1, \ldots, n_i, j \neq k)$. However, as Pepe and Anderson have noted, if all the working correlations ρ_{ijk} $(j \neq k)$ are set to zero, the only term remaining in the inner sum of equation (A3) is

$$f_{ikk} \cdot X_{ikl} \cdot \omega_{ikk} \cdot E[Y_{ik} - \mu_{ik} \mid (X_{ijl}), j = 1, \ldots, n_i] \quad (l = 1, \ldots, L)$$

for V_i^{-1} is then diagonal. In this case the internal expectation has only to condition on X_{ikl} to bring the terms in f, X, and ω outside the expectation. The proof is based on the general fact that

$$\begin{aligned} &E[f(X_1, X_2, \ldots, X_{k_1}) \cdot E(Y \mid X_1, X_2, \ldots, X_{k_1}, X_{k_1+1}, \ldots, X_{k_2})] \\ &\quad = E[f(X_1, X_2, \ldots, X_{k_1}) \cdot E(Y \mid X_1, X_2, \ldots, X_{k_1})] \end{aligned} \tag{A4}$$

for arbitrary random variables Y, X_k $(k = 1, \ldots, k_2)$ and any measurable function f if all expectations exist.

Now let V_i (and therefore V_i^{-1}) be block-diagonal and, for $k = 1, \ldots, n_i$, define $R_i(k)$ to be the set of indices $j = 1, \ldots, n_i$; $j \neq k$ with $\rho_{ijk} \neq 0$. By using a similar argument as in the case of diagonal working covariance matrices, it can be seen that the inner sum of equation (A3) is equal to

$$\sum_{j \in R_i(k)} f_{ijj} \cdot X_{ijl}^t \cdot \omega_{ijk} \cdot E[Y_{ik} - \mu_{ik} \mid X_{ijl}, j = 1, 2, \ldots, n_i] \quad (l = 1, \ldots, L)$$

and it suffices to condition $Y_{ik} - \mu_{ik}$ on X_{ij} with $j \in R_i(k)$.

Therefore, if the working covariance matrix is block diagonal, assumption (A4) has to hold only for observations within the same block.

Obviously, a general but in applications presumably not yet very helpful criterion, would define $R_i(k)$ to be the set of indices which are not equal to zero in the inverse of the working covariance matrix. This means that $R_i(k)$ has to contain all the indices of observations assumed to be dependent on observation k after conditioning on all the remaining observations of individual i according to the working correlation matrix of this individual.

Finally we note that, if the associations of interest are correctly specified, parameter estimates of the association structure using PEE (equation (4) in Section 4) are consistent since we use estimating equations of observed 'residuals'

$$\left[\frac{(y_{ij}-\hat{\mu}_{ij})\times(y_{ik}-\hat{\mu}_{ik})}{\sqrt{\text{var}(\hat{\mu}_{ij})\times\text{var}(\hat{\mu}_{ik})}}-\rho_{i,jk}\right]$$

with consistent estimates for μ_{ij} and μ_{ik}.

ACKNOWLEDGEMENTS

The authors are grateful to Dr Christian Heumann (LMU Munich) for modifying the MAREG software to print the covariance matrix of all parameter estimates. The authors thank Jean Pietrowicz for proof reading.

REFERENCES

1. Leisenring W, Pepe MS, Longton G. A marginal regression modelling framework for evaluating medical diagnostic tests. *Statistics in Medicine* 1997; **16**(11):1263–1281.
2. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**(1):13–22.
3. Prentice RL. Correlated binary regression with covariates specific to each binary observation. *Biometrics* 1988; **44**(4):1033–1048.
4. Prentice RL, Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* 1991; **47**(3):825–839.
5. Coleman, AL. Glaucoma. *Lancet* 1999; **354**:1803–1810.
6. Wolfs, RC, Borger PH, Ramrattan RS, Klaver CC, Hulsman CA, Hofman A, Vingerling JR, Hitchings RA, de Jong PT. Changing views on open-angle glaucoma: definitions and prevalences—The Rotterdam Study. *Investigative Ophthalmology and Visual Science* 2000; **41**(11):3309–3321.
7. Flammer J. The concept of visual field indices. *Graefes Archives of Clinical and Experimental Ophthalmology* 1985; **224**(5):398–392.
8. Hutchings N, Wild JM, Hussey MK, Flanagan JG,Trope GE. The long-term fluctuation of the visual field in stable glaucoma. *Investigative Ophthalmology and Visual Science* 2000; **41**(11):3429–3436.
9. Heijl A, Bengtsson B. The effect of perimetric experience in patients with glaucoma. *Archives of Ophthalmology* 1996; **114**(1):19–22.
10. Horn F, Martus P, Korth M. Comparison of temporal and spatiotemporal contrast-sensitivity tests in normal subjects and glaucoma patients. *German Journal of Ophthalmology* 1995; **4**(2):97–102.
11. Horn F, Korth M, Martus P. Quick full-field flicker test in glaucoma diagnosis: Correlations with perimetry and papillometry. *Journal of Glaucoma* 1994; **3**(3):206–213.
12. Korth M, Horn F, Storck B, Jonas JB. The pattern-evoked electroretinogram (PERG): Age-related alterations and changes in glaucoma. *Graefes Archive of Clinical and Experimental Ophthalmology* 1989; **227**(2):123–130.
13. Korth M, Nguyen NX, Jünemann A, Martus P, Jonas JB. VEP test of the blue-sensitive pathway in glaucoma. *Investigative Ophthalmology and Visual Science* 1994; **35**(5):2599–2610.
14. Jonas JB, Fernandez MC, Naumann GOH. Glaucomatous parapapillary chorioretinal atrophy: Occurrence and correlations. *Archives of Ophthalmology* 1992; **110**(2):214–222.
15. Ziegler A, Kastner C, Blettner M. The generalised estimating equations: an annotated bibliography. *Biometrical Journal* 1998; **40**(2):115–139.
16. Lipsitz SR, Laird NM, Harrington DP. Generalized estimating equations for correlated binary data: Using the odds ratio as measure of association. *Biometrika* 1991; **78**(1):153–160.
17. Ziegler A, Kastner C, Brunner D, Blettner M. Familial associations of lipid profiles: a generalized estimating equations approach. *Statistics in Medicine* 2000; **19**(24):3345–3357.
18. Kastner C, Fieger A, Heumann C. MAREG and WINMAREG—a tool for marginal regression models. *Statistical Software Newsletter in Computational Statistics and Data Analysis* 1996; **24**(2):237–241.
19. Yan J, Fine J. Estimating equations for associations. *Statistics in Medicine* 2004 (in press).
20. Franke D, Kastner C, Ziegler A. Generalized estimating equations for association structures: familial correlations of lipid profiles. *Statistics in Medicine* 2004 (in press).

21. Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics and Computation* 1994; **23**(4):939–951.
22. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001; **57**(1):120–125.
23. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine* 1978; **299**(17):926–930.
24. Diggle PJ, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press: Oxford, 1994; 137–142.
25. Pan W, Louis TA, Connett JE. A note on marginal linear regression with correlated response data. *The American Statistician* 2000; **54**(3):191–195.
26. Jünemann A, Horn F, Martus P, Jonas JB, Naumann GOH. Sinnes-physiologie und Papillenbiomorphometrie: Korrelationen bei Normalpersonen. *Klinische Monatsblätter für Augenheilkunde* 1996; **209**(5):286–291.
27. Martus P, Jünemann A, Wisse M, Korth M, Budde W, Horn F, Jonas, JB. A multivariate statistical model for quantification of glaucomatous damage. *Investigative Ophthalmology and Visual Sciences* 2000; **41**(5): 1099–1110.
28. Martus P. A measurement model of disease severity in absence of a gold standard. *Methods of Information in Medicine* 2001; **40**(3):265–271.

