SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Original Research Article

Repeated measures discriminant analysis using multivariate generalized estimation equations

Statistical Methods in Medical Research
2022, Vol. 31(4) 646–657

DOI: 10.1177/09622802211032705
journals.sagepub.com/home/smm
SAGE

Anita Brobbey[1], Samuel Wiebe[1,2], Alberto Nettel-Aguirre[3], Colin Bruce Josephson[1,2], Tyler Williamson[1], Lisa M Lix[4] and Tolulope T. Sajobi[1,2]

Abstract

Discriminant analysis procedures that assume parsimonious covariance and/or means structures have been proposed for distinguishing between two or more populations in multivariate repeated measures designs. However, these procedures rely on the assumptions of multivariate normality which is not tenable in multivariate repeated measures designs which are characterized by binary, ordinal, or mixed types of response distributions. This study investigates the accuracy of repeated measures discriminant analysis (RMDA) based on the multivariate generalized estimating equations (GEE) framework for classification in multivariate repeated measures designs with the same or different types of responses repeatedly measured over time. Monte Carlo methods were used to compare the accuracy of RMDA procedures based on GEE, and RMDA based on maximum likelihood estimators (MLE) under diverse simulation conditions, which included number of repeated measure occasions, number of responses, sample size, correlation structures, and type of response distribution. RMDA based on GEE exhibited higher average classification accuracy than RMDA based on MLE especially in multivariate non-normal distributions. Three repeatedly measured responses namely severity of epilepsy, current number of anti-epileptic drugs, and parent-reported quality of life in children with epilepsy were used to demonstrate the application of these procedures.

Keywords

Discriminant analysis, multivariate repeated measures data, generalized estimating equation, multivariate non-normal distribution, classification

1 Introduction

Discriminant analysis (DA) is commonly used to classify an individual into one of two (or more) populations on the basis of correlated response measures. In more recent years, relevant work has been done in capturing the longitudinal nature of clinical data and using it for classification via discriminant analysis.[1–8] These research studies include DA extensions to repeated measures data with multiple response. Utilizing the correlation structure across responses with a multivariate model could increase the classification accuracy.[9] Classical DA does not model the covariance structure, and thus the information regarding the possible structure in the covariance for

[1]Department of Community Health Sciences, University of Calgary, Calgary, Canada
[2]Department of Clinical Neurosciences, University of Calgary, Calgary, Canada
[3]Centre for Health and Social Analytics, National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, Australia
[4]Department of Community Health Sciences, University of Manitoba, Winnipeg, Canada

Corresponding author:
Tolulope Sajobi, Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, 3280 Hospital Drive, NW Calgary, AB T2N 4Z6, Canada.
Email: ttsajobi@ucalgary.ca

repeated measurements taken on the same individual and between responses is lost.[10–13] Moreover, classical DA is based on multivariate normality assumption to guarantee an optimal solution. Equal covariance structures are assumed in the groups[10] for linear discriminant analysis (LDA), while quadratic discriminant analysis (QDA) allows for unequal covariance structures between the groups.[11–13]

Most DA methodologies in multivariate repeated measures data are based on mixed effects model. Multivariate linear and non-linear mixed effects models that assume unstructured[1,14] and parsimonious structure[7,9,15,16] for the variance-covariance matrix have been introduced. For instance, several continuous markers and a multivariate linear mixed effects model were used to evaluate a prognosis of primary biliary cirrhosis patients[14] and non-linear mixed effects model to distinguish between women with and without pregnancy abnormalities.[15] Similarly, three continuous markers were used to classify patients suffering from prostate cancer.[7] Generalized linear mixed effects models have been extended in multivariate repeated measures studies for different type of responses (continuous, counts, and binary).[1,3,17] Most mixed effects model DA assume that the random effects follow a multivariate normal distribution. Moreover, the dimension of random-effects quickly increases as more responses and more measurements occasions are added to the model, increasing the computational burden and instability.[1,8,14] In addition, it is difficult to evaluate the marginal likelihood of jointly generalized linear mixed effects models when the response is non-normal.

Contrary to mixed effects models approaches, some researchers have utilized the generalized estimating equations (GEE) based on multiple marginal models of multiple responses. To avoid the specification of the full likelihood function especially for discrete data, GEE[18] is a suitable approach for parameter estimation for repeated measures data without full specification of the likelihood. Specifically, GEEs directly specify a marginal mean model for each response and induce the correlation between measurements of responses through a working correlation matrix. GEEs offer a computationally non-intensive parameter estimation algorithm and the resulting parameter estimates have population-averaged interpretation. A joint modeling of multiple response variables is based on straightforward extension of univariate GEEs with correlation structure across responses which provides separate set of regression parameters for each response variable.[19,20] GEEs are less sensitive to covariance misspecification compared to mixed effects models.[18,21]

This study examines the accuracy of discriminant analysis based on multivariate GEE framework for classification in multivariate repeated measures designs with same/different types of responses. The article is organized as follows. In section 2, we describe the GEEs framework for multivariate repeated measures data. The proposed approach, the extension of the GEE framework to discriminant analysis, is presented in section 3. In section 4, we summarize the results of a Monte Carlo simulation study to assess the accuracy of the proposed repeated measures discriminant analysis based on GEE RMDA approach under diverse simulation scenarios. Data from a multivariate longitudinal study of children with epilepsy were used to demonstrate the application of these procedures in section 5. Finally, a discussion of the key findings from the study and its implications are described in section 6.

2 GEE for multivariate repeated measures data

Suppose we have a random sample of N individuals. For each individual $i = 1, \ldots, N$, let $\mathbf{y}_i = (\mathbf{y}'_{i1}, \mathbf{y}'_{i2}, \ldots, \mathbf{y}'_{iQ})'$ be a PQ x 1 vector of Q correlated responses that are each repeatedly measured at P occasions, and $\mathbf{X}_i = \mathbf{I}_Q \otimes \mathbf{X}_{i*}$ is a corresponding $KQ \times PQ$ block diagonal covariate matrix, where $\mathbf{X}_{i*} = (\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots \mathbf{X}_{ik}, \ldots, \mathbf{X}_{iK})$ is a K x P matrix of covariates, $\mathbf{I}_Q$ is an Q x Q identity matrix, and $\otimes$ is the Kronecker product symbol. For the analysis of multivariate correlated data, the marginal mean vector is associated with the covariates through a generalized linear model (GLM) as follows

$$\boldsymbol{\mu}_{ipq} = \mathrm{E}\left(\mathbf{y}_{ipq} | \mathbf{X}_{ip}\right) = \mathrm{f}(\mathbf{X}'_{ip}\boldsymbol{\beta}_q), \tag{1}$$

where, $\mathrm{f}(\cdot)$ is the inverse response-specific link function, $\boldsymbol{\beta}_q = (\boldsymbol{\beta}_{q1}, \ \boldsymbol{\beta}_{q2}, \ \ldots \boldsymbol{\beta}_{qk}, \ldots, \ \boldsymbol{\beta}_{qK})'$ is the $K \times 1$ dimensional vector of the qth response regression coefficients, and $\mathbf{X}_{ip}$ is the corresponding covariate at time p for the ith individual. The KQ-dimensional parameter vector of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \ldots, \boldsymbol{\beta}_Q)$ and the marginal model in equation (1) is represented by $\boldsymbol{\mu}_i = \mathrm{E}\left(\mathbf{y}_i | \mathbf{X}_i\right) = \mathrm{f}\left(\mathbf{X}'_i \boldsymbol{\beta}\right)$. In the quasi-likelihood framework with repeated measures responses, the regression coefficients in $\boldsymbol{\beta}$ can be estimated by solving the generalized estimating equations (GEEs)

$$\mathrm{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{D}'_i \boldsymbol{\Omega}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0 \tag{2}$$

where $\mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$ is the block diagonal matrix of derivatives mean with respect to the regression parameters, $\boldsymbol{\mu}_i$ is the marginal mean vector, and $\boldsymbol{\Omega}_i$ is the $PQ \times PQ$ working covariance matrix.

The $PQ \times PQ$ marginal covariance matrix is

$$\boldsymbol{\Omega}_i = \phi \boldsymbol{\Sigma}_i, \tag{3}$$

where ϕ is a scale parameter that can be known or estimated and $\boldsymbol{\Sigma}_i$ is an $PQ \times PQ$ working covariance matrix, which results in a total of $PQ\ (PQ+1)/2$ unknown parameters to be estimated[22,23] which may not always be feasible (i.e. when PQ is close to N). To reduce the dimension of the unknown parameters of the covariance matrix, a parsimonious structure is sometimes used, such as a Kronecker product covariance structure

$$\boldsymbol{\Sigma}_i = \mathbf{A}_i^{1/2} \mathbf{R}(\alpha) \otimes \mathbf{R}(\rho))\mathbf{A}_i^{1/2} \tag{4}$$

where $\mathbf{A}_i$ is an $PQ \times PQ$ block diagonal matrix, which contains the marginal variance of responses on the main diagonals, $\mathbf{R}(\alpha)$ is a $Q \times Q$ working correlation matrix of the responses with the parameter vector α, and $\mathbf{R}(\rho)$ is the working correlation matrix for a given response at different time points with the parameter ρ. This structure reduces the number of covariance parameters to be estimated.[23–27] Consequently, $\mathbf{R}(\alpha)$ and $\mathbf{R}(\rho)$ denote between-response correlation matrix and within-response correlation matrix, respectively. Further, assuming a structured working correlation, such as exchangeable (EX), first-order autoregressive (AR1), or unstructured (UN), for $\mathbf{R}(\alpha)$ and exchangeable (EX) or unstructured (UN) structures for $\mathbf{R}(\boldsymbol{\rho})$, can lead to an even more parsimonious model.[22,28,29] The parsimonious structure provides flexible model for covariance, particularly when sample size is small.[22,28,29] Inferences of interest are easily influenced by the correlation structure's assumptions, and unstructured correlation structure might cause convergence problems as the number of parameters to be estimated grows rapidly.[30] Specifically, $\mathrm{U}\boldsymbol{\beta} = 0$ are solved with a Fisher-Scoring algorithm such that

$$\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} + \left(\sum_{i=1}^{N} \tilde{\mathbf{D}}_i' \tilde{\boldsymbol{\Omega}}_i^{-1} \tilde{\mathbf{D}}_i\right)^{-1} \left(\sum_{i=1}^{\mathrm{N}} \tilde{\mathbf{D}}_i' \tilde{\boldsymbol{\Omega}}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)\right) \tag{5}$$

Under mild regularity conditions, the parameter estimates are consistent and asymptotically normally distributed even when the "working" correlation structure of responses is misspecified, and the variance-covariance matrix can be estimated using a robust "sandwich" variance estimator.[31] The asymptotic covariance matrix of the non-vanishing (non-zero) component of $\hat{\boldsymbol{\beta}}$ via the sandwich estimator formula is[31,32]

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \hat{\mathbf{D}}_i' \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1} \hat{\mathbf{M}}_* \left(\sum_{i=1}^{N} \hat{\mathbf{D}}_i' \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1}, \tag{6}$$

with

$$\hat{\mathbf{M}}_* = \sum_{i=1}^{N} \hat{\mathbf{D}}_i' \hat{\boldsymbol{\Omega}}_i^{-1} \widehat{\mathrm{cov}}(\mathbf{y}_i) \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{D}}_i \tag{7}$$

and $\widehat{\mathrm{cov}}(\mathbf{y}_i) = (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)'$ is an estimator of the true variance-covariance matrix of $\mathbf{y}_i$.[18,31] Note that if $\boldsymbol{\Omega}_i$ is correctly specified, $\boldsymbol{\Omega}_i = \mathrm{cov}(\mathbf{y}_i)$.[33,34] Moreover, GEE requires the correct specification of marginal mean and variance as well as the link function, which connects the covariates of interest and the marginal means.

3 GEE extension to multivariate repeated measures discriminant analysis

Following the GEE notation, we assume that the ith individual in the jth population (j =1,2) with multivariate repeated responses $\mathbf{y}_{ji}$, has a marginal mean $\boldsymbol{\mu}_{\mathrm{ji}}$, and variance covariance matrix $\boldsymbol{\Omega}_{ji}$ assumed to be $PQ \times PQ$ positive definite. Analogously, with estimations of $\hat{\boldsymbol{\mu}}_{ji} = \mathrm{f}(\boldsymbol{X}_{ji}\hat{\boldsymbol{\beta}}_j)$ and the variance covariance matrix $\hat{\boldsymbol{\Omega}}_{ji}$ from the GEE model in population j using a pre-defined structure, the homoscedastic model is obtained when the variance components are homogeneous, that is, $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2 = \boldsymbol{\Omega}$, the pooled covariance matrix. Based on LDA, a randomly selected ith individual with multiple response vector $\mathbf{y}_i$ is classified in the first group, if

$$\left(\mathbf{y}_i - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}\right)' \hat{\boldsymbol{\Omega}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) > \log \frac{\hat{\pi}_2}{\hat{\pi}_1} \tag{8}$$

where $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Omega}}_j^{-1}$are the GEE estimates from equations (1) and (2), $\hat{\pi}_1$ and $\hat{\pi}_2$ are the *a priori* probabilities that observations belong to populations 1 and 2. Otherwise, the individual is classified into the second group. For QDA (i.e. $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$), the ith subject with multiple response vector $\mathbf{y}_i$ is classified in the first group, if

$$(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Omega}}_2^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_2) - (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\Omega}}_1^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_1) \ > \log \frac{\hat{\boldsymbol{\Omega}}_1}{\hat{\boldsymbol{\Omega}}_2} + 2\log \frac{\hat{\pi}_2}{\hat{\pi}_1}, \tag{9}$$

otherwise, it is classified into the second group.

4 Simulation study

A Monte Carlo simulation study was conducted to examine the accuracy of linear and quadratic GEE discriminant analysis procedures that assume Kronecker product structured covariances compared to MLE repeated measures discriminant analysis based on structured covariances.[22,29,35] The following conditions were investigated: (a) number of repeated measurements (P), (b) total sample size (N), (c) group sizes (n_1, n_2), (d) pattern and magnitude of correlation among the repeated measurements (ρ), (e) mean configuration, (f) covariance heterogeneity, and (g) population distribution. All procedures were investigated for two independent groups. The number of repeated occasions/time points was set at $P =$ 3 and 5, and number of responses was set at $Q =$ 3 and 5. Previous studies about DA procedures for multivariate repeated measures data have considered P ranging from 3 to 10, an increase in classification accuracies were quite significant when P increased from three to five.[36,37] Total sample sizes of $N = 80$, 140 and 200 were investigated. This is consistent with previous simulation studies that examined the accuracy of DA for multivariate repeated measures data between 60 and 500. Moreover, consistent with previous studies that examined the impact of equal and unequal group sizes,[36–39] we investigated conditions of $N = 80$, $(n_1,\ n_2) = (40,\ 40)$, and (32, 48), which represent a group size ratio of 1:1 and 2:3, respectively. Similar equal and unequal group size ratios were investigated when $N = 140$ and $N = 200$. Furthermore, the accuracy of DA procedures is known to be influenced by both the magnitude and pattern of within- and multivariate-response correlations.[40] Therefore, we investigated the following within-response correlation structures: (a) Compound Symmetry with $\rho = 0.3$ and $\rho = 0.7$, (b) autoregressive order 1 with $\rho = 0.3$ and $\rho = 0.7$[36,37] for $\mathbf{R}(\boldsymbol{\rho})$, and the between-responses correlation, $\mathbf{R}(\boldsymbol{\alpha})$ was assumed to be unstructured (See Table 1 for more details).

Hence, we assumed two Kronecker correlation structures; UNAR = Unstructured between-responses and Autoregressive order-1 within-response correlation matrix, and UNCS = Unstructured between-responses and Compound symmetry within-response correlation matrix. For covariance heterogeneity, we assumed $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$ and $\boldsymbol{\Omega}_1 = 3\boldsymbol{\Omega}_2$.

In order to assess the performance of the discriminant function, we investigated multivariate correlated continuous response variables, count response variables, and different types of correlated responses, namely Case 1, Case 2, and Case 3, respectively. Case 1: For the correlated continuous response variables, we assumed three normal variables jointly observed for $\boldsymbol{n}_j$ subjects, where each observed at P time points. The true marginal mean response model $\boldsymbol{\mu}_{ipq}$ was assumed to take the following functional form that uses an identity link function

$$\boldsymbol{\mu}_{ipq} = \ \boldsymbol{\beta}_{q1}x_{ip} + \boldsymbol{\beta}_{q2}t_{ip} \tag{10}$$

The number of covariates, $K = 2$, where x_{ip} was generated from an independent normal random variable $N(0,1)$ as a time-invariant covariate, and t_{ip} denoted the time of observation as a time-varying covariate. Details of the true parameters $\boldsymbol{\beta}$ for population 1 and population 2 can be found in Table 2.

On the other hand, the marginal variance matrix of responses was assumed to have a common variance of 60. Case 2: For the multivariate count response variables, data were generated from a multivariate Poisson

Table 1. Configuration of unstructured between-responses correlation matrix given within-response correlation coefficient for the Monte Carlo Study.

Within-response Correlation Coefficient (ρ)	0.3	0.7
$Q = 3$	$\begin{bmatrix} 1 & 0.15 & 0.30 \\ 0.15 & 1 & 0.45 \\ 0.30 & 0.45 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.65 & 0.66 \\ 0.65 & 1 & 0.70 \\ 0.66 & 0.70 & 1 \end{bmatrix}$
$Q = 5$	$\begin{bmatrix} 1 & 0.28 & 0.25 & 0.28 & 0.28 \\ 0.28 & 1 & 0.30 & 0.40 & 0.23 \\ 0.25 & 0.30 & 1 & 0.24 & 0.24 \\ 0.28 & 0.40 & 0.24 & 1 & 0.37 \\ 0.28 & 0.23 & 0.24 & 0.37 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.70 & 0.79 & 0.64 & 0.70 \\ 0.70 & 1 & 0.73 & 0.65 & 0.74 \\ 0.79 & 0.73 & 1 & 0.63 & 0.62 \\ 0.64 & 0.65 & 0.63 & 1 & 0.62 \\ 0.70 & 0.74 & 0.62 & 0.62 & 1 \end{bmatrix}$

Q: Number of responses.

Table 2. True parameters ($\boldsymbol{\beta}$) for population 1 and population 2 simulated data.

Population distribution	Number of responses	Population 1	Population 2
Normal/mixed-type	3	(0.3,1, 2,0.1,1, 1.5)	(0.6,2, 4,0.2,2, 3)
	5	(0.2,1, 2,1.5,1, 0.4,0.7,3, 1.2,0.8)	(0.4,2, 4,3, 2,0.8,1.4,6, 2.4,1.6)
Poisson	3	(0.3,0.1,0.2,0.1,0.3,0.5)	(0.9, 0.3, 0.6, 0.3 ,0.9, 1.5)
	5	(0.3,0.1,0.4,0.1,0.45,0.6,0.2,0.15,0.3,0.4)	(0.9, 0.3,1.2,0.3,1.35,1.8,0.6,0.45,0.9,1.2)

distribution using the log link function instead of identity link in Case 1 and log transformation of time of observation as a time-varying covariate.

$$\boldsymbol{log}(\boldsymbol{\mu}_{ipq}) = \boldsymbol{\beta}_{q1}x_{ip} + \boldsymbol{\beta}_{q2}\log(t_{ip}) \tag{11}$$

The true parameters $\boldsymbol{\beta}$ for population 1 and population 2 can be found in Table 2. Case 3: For generating different types of correlated responses, one of the responses generated from case 1 (multivariate normal distribution data) was converted to Bernoulli response using the NORmal-To-Anything (NORTA) algorithm[41] with probabilities from the logit function.

Linear and quadratic discriminant analysis rules were developed using the marginal mean and variance-covariance matrix estimated via GEE, and MLE for equal and unequal covariance matrix, respectively. The classification performance of the procedures was evaluated using the overall average classification accuracy and its corresponding standard errors.

$$\text{Overall classification accuracy} = \frac{\text{correct classifications}}{\text{number of classifications } (N)} \tag{12}$$

All combinations of simulation conditions were investigated for each procedure and each method of estimation, resulting in a total of 194 combinations. There were 500 replications for each combination. All analyses were completed in R statistical software version 3.5.3.

4.1 Simulation results

Tables 3 and 4 describe the average classification accuracies and standard errors of repeated measures linear and quadratic discriminant analysis based on GEE, and MLE, respectively, by population distribution, number of repeated occasions, and number of responses. For each type of estimator, there were negligible differences ($<$ 0.04) for linear DA UNCS and UNAR procedures. But RMDA based on GEE procedures were more accurate

Table 3. Overall mean accuracy (standard error) for repeated measures LDA procedures based on GEE, and MLE by population distribution, number of responses, number of measurements occasions, and correlation structure.

Population distribution	Number of responses	Number of measurements occasions	GEE		MLE	
			UNAR	UNCS	UNAR	UNCS
Normal	3	3	0.62 (0.04)	0.64 (0.04)	0.63 (0.04)	0.65 (0.04)
		5	0.73 (0.04)	0.75 (0.04)	0.69 (0.04)	0.70 (0.04)
	5	3	0.68 (0.04)	0.74 (0.04)	0.66 (0.04)	0.66 (0.04)
		5	0.83 (0.03)	0.89 (0.03)	0.82 (0.03)	0.63 (0.03)
Poisson	3	3	0.88 (0.04)	0.90 (0.03)	0.79 (0.04)	0.81 (0.04)
		5	0.97 (0.02)	0.97 (0.03)	0.84 (0.05)	0.85 (0.05)
	5	3	0.99 (0.01)	0.99 (0.01)	0.89 (0.04)	0.90 (0.04)
		5	0.99 (0.01)	0.99 (0.01)	0.95 (0.02)	0.95 (0.02)
Mixed-type	3	3	0.62 (0.04)	0.63 (0.04)	0.55 (0.04)	0.55 (0.04)
		5	0.72 (0.04)	0.74 (0.04)	0.58 (0.04)	0.58 (0.04)
	5	3	0.68 (0.04)	0.72 (0.04)	0.67 (0.04)	0.57 (0.04)
		5	0.81 (0.03)	0.87 (0.03)	0.62 (0.04)	0.62 (0.04)

UNAR: unstructured between-responses and autoregressive order 1 within-response correlation matrix; UNCS: unstructured between-responses and compound symmetry within-response correlation matrix; GEE: generalized estimating equation; MLE: maximum likelihood estimation.

Table 4. Overall mean accuracy (standard error) for repeated measures QDA procedures based on GEE, and MLE by population distribution, Number of responses, number of measurements occasions, and correlation structure.

Population distribution	Number of responses	Number of measurements occasions	GEE		MLE	
			UNAR	UNCS	UNAR	UNCS
Normal	3	3	0.77 (0.04)	0.80 (0.04)	0.65 (0.04)	0.66 (0.04)
		5	0.85 (0.04)	0.88 (0.04)	0.71 (0.04)	0.71 (0.04)
	5	3	0.85 (0.04)	0.89 (0.04)	0.66 (0.04)	0.66 (0.04)
		5	0.90 (0.03)	0.94 (0.03)	0.85 (0.03)	0.90 (0.02)
Poisson	3	3	0.93 (0.03)	0.94 (0.03)	0.78 (0.04)	0.79 (0.04)
		5	0.99 (0.01)	0.98 (0.03)	0.85 (0.05)	0.85 (0.05)
	5	3	0.99 (0.01)	0.99 (0.01)	0.90 (0.04)	0.92 (0.03)
		5	0.99 (0.01)	0.99 (0.01)	0.95 (0.02)	0.95 (0.02)
Mixed-type	3	3	0.74 (0.04)	0.75 (0.04)	0.56 (0.04)	0.55 (0.04)
		5	0.84 (0.04)	0.85 (0.04)	0.58 (0.04)	0.58 (0.06)
	5	3	0.83 (0.04)	0.86 (0.04)	0.58 (0.04)	0.58 (0.04)
		5	0.91 (0.03)	0.94 (0.03)	0.63 (0.04)	0.63 (0.04)

UNAR: unstructured between-responses and autoregressive order 1 within-response correlation matrix; UNCS: unstructured between-responses and compound symmetry within-response correlation matrix; GEE: generalized estimating equation; MLE: maximum likelihood estimation.

than RMDA based on MLE among UNCS procedures. For example: for the UNCS correlation matrix under GEE, the average accuracy for $P = 3$ was 0.74 and $P = 5$, it was 0.89, while for the UNCS correlation matrix under MLE, the average accuracy for $P = 3$ was 0.66 and $Q = 5$, it was 0.63 when the number of responses was five (Table 3).

Moreover, RMDA based on GEE had the highest average classification accuracy compared to RMDA procedures based on MLE when responses were sampled from a multivariate Poisson distribution and mixed type responses. For example, when $P = 3$ and $Q = 5$, the average classification accuracies of RMDA procedures based on GEE and MLE were 0.97 and 0.84 when data were sampled from a multivariate Poisson distribution with response variables. Whereas, the average accuracy of the GEE and MLE procedures were 0.72 and 0.58, respectively, when mixed type responses, under UNAR correlation matrix (Table 3). In the quadratic discriminant analysis procedures, RMDA procedures based on MLE were least accurate regardless of number of repeated occasions, number of responses, estimation method or multivariate distribution of response variables (Table 4).

For example, when Q = 5 and $P = 3$ under UNAR correlation matrix, the average classification accuracies of RMDA procedures based on GEE and MLE were 0.85 and 0.66 when data were sampled from a multivariate normal distribution with response variables.

Furthermore, the average accuracy of each linear and quadratic discriminant analysis procedure increased as the number of repeated occasions and number of responses increased, regardless of the estimation method or multivariate distribution of response variables. For example, for Q = 3, when data were sampled from a multivariate normal distribution, the average increase in classification accuracy of the RMDA procedure based on GEE and MLE were about 0.11 and 0.05, respectively, as p increased from three to five, under UNCS correlation matrix (Table 3). Likewise, the average increase in classification accuracy of the RMDA procedure based on GEE and MLE were about 0.10 and 0.01, respectively, as Q increased from three to five, under UNCS correlation matrix and $P = 3$ (Table 3).

It is worth mentioning that, we observed little or no differences in classification accuracies for linear and quadratic discriminant procedures when RMDA procedures based on MLE were used, whereas the classification accuracies for quadratic discriminant procedures based on GEE increased compared to its corresponding linear discriminant procedures (Tables 3 and 4)

For example: the average accuracy for RMDA procedure based on GEE and MLE were 0.64 and 0.65, respectively, for linear discriminant procedure (Table 3), while for quadratic discriminant procedure, the average accuracies were 0.80 and 0.66, respectively (Table 4) under the UNCS correlation matrix, when data were sampled from a multivariate normal distribution with response variables and $P = 3$.

5 Health-Related Quality of Life in Children with Epilepsy Study (HERQULES)

Multivariate repeated measures data were obtained from the Health-Related Quality of Life (HRQOL) in Children with Epilepsy Study (HERQULES), a two-year prospective cohort study assessing the course and characteristics potentially associated with HRQOL in children with new onset epilepsy across Canada.[12,13] Details of HERQULES have been described elsewhere.[12,13] Data were collected as soon as possible following the diagnosis of epilepsy at baseline (0 month), and approximately 6 months, 12 months, and 24 months later. Standardized questionnaires were used to collect parent-report of their children's HRQOL and a series of child and family characteristics, while a neurologist report form collected information on clinical characteristics of the child's epilepsy.

Using these multivariate repeated measures data, we sought to identify patients who will not achieve remission from seizures within two years from disease onset. This group is referred to as the refractory group. A patient is defined as being in remission if they had a continuous 12-month period without any seizures at any point within two years from diagnosis.[3] Early identification of patients who have refractory epilepsy can allow clinicians to explore alternative treatment options (e.g. surgery) to manage seizures and other aspects of the disease.[3] Data for this numeric example consist of response variables such as severity of epilepsy,[14] current number of anti-epileptic drugs (AEDs), and parent-reported quality of life in children using epilepsy-specific scale which were measured over four measurement occasions and the covariates were time of observation as a time-varying covariate, age at seizure onset, and sex as time invariant covariates. Repeated measures linear and quadratic discriminant analysis classification rules were developed based on multivariate GEE model using these data.

Of the 187 patients included in this analysis, 101 patients were in the remission group and 86 patients were in the refractory group within two years. The sample included children ages 4 to 12 years. The average age (standard deviation) in the remission group was 8.25 (2.46) years and in the refractory group was 8.25 (2.46) years. The patients included 45.54% and 41.86% females in the remission and refractory groups respectively. The QOLCE-55 ratings underwent a linear transformation such that domain scores yield values from 0 (low HRQOL) to 100 (high HRQOL). The ratings were treated as a continuous variable. The GASE scale is a seven-point Likert scale ranging from 1 (not severe at all) to 7 (extremely severe) was recoded as a binary variable, with ≥ 3 coded as severe, thereby using the median severity 3, corresponding to "somewhat severe" as a cut-off.[42]

5.1 Results for HERQULES data

Figure 1 describes the longitudinal changes in the levels of each of the response variables for all patients in each diagnostic group. For patients who achieved remission, severity of seizures appears to decrease over time, whereas seizure severity remained high for the refractory group. The difference between the overall quality of life of the two groups is less noticeable. However, the overall quality of life appears constant over time in the refractory

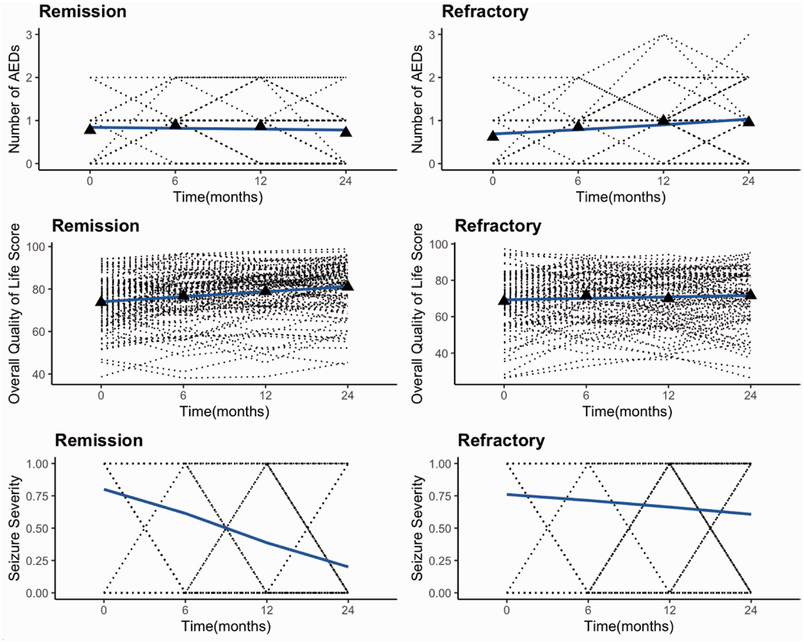

Figure 1. Observed longitudinal profiles of number of anti-epileptic drugs (AEDs), quality of life, and seizure severity from the Remission group (left column) and the Refractory group (right column). Solid lines show LOESS smoothed profiles for Poisson, normal, and binomial models calculated using data from all patients. Baseline (0 month), and 6 months, 12 months, and 24 months.

Table 5. GEE group-specific correlation parameter estimates for HERQULES data by the assumed correlation structure.

	Remission		Refractory	
	UNAR	UNCS	UNAR	UNCS
ρ	0.812	0.749	0.744	0.726
Corr(Y_2Y_1)	–0.025		–0.023	
Corr(Y_3Y_1)	0.003		0.001	
Corr(Y_3Y_2)	–0.042		–0.038	

UNAR: unstructured between-responses and autoregressive order 1 within-response correlation matrix; UNCS: unstructured between-responses and compound symmetry within-response correlation matrix; number of (AEDs) ($\mathbf{Y}_1$), HRQOL ($\mathbf{Y}_2$), Severe Seizure ($\mathbf{Y}_3$).

group, but as time increases, the overall quality of life of the remission patients gradually increases. The number of AEDs increased over time for the refractory patients, while those in the remission group had slightly reduced number of AEDs.

Table 5 gives the group-specific correlation parameter estimates of the joint modeling of the multiple repeated responses using multivariate GEE. We observed that in both remission and refractory groups, HRQOL was negatively associated with severity of seizures and the number of AEDs. However, there was little to no association between severity of seizures and the number of AEDs.

The accuracy of LDA and QDA classifiers based on GEE and maximum likelihood estimators are described in Table 6. Overall, RMDA procedures based on GEE exhibited higher overall classification accuracy than RMDA based on MLE in both LDA and QDA. Moreover, the classification accuracies observed using GEE estimators increased when QDA (accuracy, 0.79) was used for classification compared to its LDA (accuracy, 0.71) approach, while the accuracy using MLE estimators for remains the same for both QDA and LDA (accuracy, 0.67). The classifiers were more accurate in correctly reclassifying patients in the remission group but less accurate for reclassifying those in the refraction group.

Table 6. Classification accuracy for the generalized estimating equation (GEE), and maximum likelihood estimation (MLE) methods for repeated measures LDA and QDA by the assumed correlation structure.

	GEE		MLE	
	UNAR	UNCS	UNAR	UNCS
LDA				
Remission	0.772	0.770	0.762	0.760
Refractory	0.651	0.640	0.570	0.558
Overall	0.711	0.705	0.665	0.660
QDA				
Remission	0.871	0.880	0.752	0.750
Refractory	0.709	0.698	0.581	0.570
Overall	0.790	0.789	0.667	0.660

LDA: linear discriminant analysis; QDA: quadratic discriminant analysis; GEE: generalized estimating equation; MLE: maximum likelihood estimation; UNAR: unstructured between responses and autoregressive order 1 within response correlation matrix; UNCS: unstructured between responses and compound symmetry within response correlation matrix.

6 Discussion

This study investigated discriminant analysis procedures for multivariate repeated measures data using GEE for discriminating between population groups. The proposed approach allows the incorporation of repeated measures responses and covariates to improve the accuracy of the classifier. Our results showed that the RMDA based on GEE model resulted in better classification accuracy than the conventional RMDA based on maximum likelihood estimators especially in multivariate repeated measures data with discrete and/or mixed type of responses.[43,44] This is because the GEE allows for the incorporation of multivariate repeated measures outcomes of different types without the need to fully specify the likelihood.[20,30,44,45] Another advantage of these procedures is their ability to accommodate both time-invariant and time-varying covariate to improve the accuracy of modelclassifiers.

Furthermore, our study revealed the impact of increasing repeated occasions and number of responses on the accuracy of the investigated procedures. The impact of increasing number of repeated occasions is consistent with literature on other RMDA methods[22,37]; however, the studies from literature did not investigate the impact of increasing number of responses. Specifically, the RMDA based on GEE was most accurate with increases in the number of repeated occasions and number of responses compared to RMDA based on MLE. Overall, the quadratic discriminant analysis was able to better classify individuals than the linear discriminant analysis in RMDA based on GEE. QDA provides a less restrictive procedure by allowing different covariance matrixes for each group, which minimizes misclassification. Even though classification rules based on LDA can perform badly if the assumption of a common within-class covariance matrix is violated, classification rules based on QDA require a larger sample size to overcome the singularity problem.[13,46,47] Even though the procedures developed in this study are based on two-group multivariate repeated designs, our conclusions can be extended and generalized to multi-group designs.[48,49]

Despite the unique strengths of this class of repeated measures discriminant analysis models, they are not without their limitations. First, the RMDA based on GEE relies on correctly specified link function and parsimonious covariance structures, which might not be tenable in typical multivariate repeated measures data. It is well known that GEEs yield asymptotically consistent parameter and variance estimates even under incorrect specification of the correlation structure but correctly specified link function.[18,43,45] This means that a crucial step in the GEE approach is to select a correct link function linking the mean response to the covariates.[50] With regard to parsimonious covariance structures, even though several authors have observed many advantages of using Kronecker product structure for analyzing multivariate repeated measures data,[22,24,37,51,52] one could use the usual unstructured variance covariance matrix when there is sufficient data. Moreover, some work has been done on the testing of hypotheses of Kronecker product structure.[22,24,26] It is also not clear whether the misspecification of the working correlation structures for these procedures could influence their classification accuracy.[53] However, one does not know a priori which correlation structure is correct. Future research is needed to examine the impact of misspecification of covariance structure on the accuracy of these classifiers. In addition, to help in choosing a working correlation matrix that is close to the true correlation matrix, a quasi-likelihood under the independence model criterion (QIC) which is a modified Akaike information criterion (AIC) has recommended for GEE model.[54,55] Secondly, the assumption of complete multivariate repeated measures data in which there are no missing data on all responses and at all measurement occasions might not be realistic in multivariate repeated measures data often encountered in applied research. Even in a well-controlled repeated measures study, missing data may frequently occur due to missed visits, withdrawal from the study, or loss to follow-up.[20] Some studies have been done to address drop-out problems in repeated measures studies via weighted generalized estimating equations[56] and imputations. Further research could extend the DA procedures based on GEE by implementing some of the multiple imputation techniques.[20,57–59] Finally, our study focussed on comparing marginal models (GEE and covariance pattern models), which may not be efficient when accounting for individual-specific variations and dealing with missing data. Discriminant analysis based on mixed models constitute an alternative class of longitudinal classifiers that can account for individual-specific variations in longitudinal trajectories and accommodate incomplete longitudinal data, however, rely on the multivariate normality assumption.[7,9,15,16] Future research will investigate the accuracy of discriminant analysis classifiers based on marginal and random-effects conditional models.

In summary, this study proposes a new class of discriminant analysis procedures based on GEE, which can be used for distinguishing between population groups in multivariate repeated measures data characterized by multivariate non-normal distributions with continuous, binary, or mixed types of response variables.

Acknowledgments

Special thanks to Dr. Kathy Speechley (Western University, Canada) for providing access to the HERQULES dataset.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a Natural Sciences & Engineering Research Council Discovery grant to Dr. Sajobi.

ORCID iDs

Anita Brobbey https://orcid.org/0000-0001-9386-7954
Tyler Williamson https://orcid.org/0000-0001-5029-2345
Lisa M Lix https://orcid.org/0000-0001-8685-3212

References

1. Fieuws S, Verbeke G, Maes B, et al. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics* 2007; **9**: 419–431.
2. Galecki AT. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Commun Stat – Theory Meth* 1994; **23**: 3105–3119.
3. Hughes DM, Komárek A, Czanner G, et al. Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Stat Meth Med Res* 2018; **27**: 2060–2080.
4. Inoue LYT, Etzioni R, Morrell C, et al. Modeling disease progression with longitudinal markers. *J Am Stat Assoc* 2008; **103**: 259–270.
5. Li Y, Wang Y, Wu G, et al. Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. *Neurobiol Aging* 2012; **33**: 427. e415–427. e430.
6. Marshall G and Barón AE. Linear discriminant models for unbalanced longitudinal data. *Stat Med* 2000; **19**: 1969–1981.
7. Morrell CH, Brant LJ, Sheng S, et al. Screening for prostate cancer using multivariate mixed-effects models. *J Appl Stat* 2012; **39**: 1151–1175.
8. Verbeke G, Fieuws S, Molenberghs G, et al. The analysis of multivariate longitudinal data: a review. *Stat Meth Med Res* 2014; **23**: 42–59.
9. Marshall G, De la Cruz-Mesía R, Barón AE, et al. Non-linear random effects model for multivariate responses with missing data. *Stat Med* 2006; **25**: 2817–2830.
10. Lachenbruch PA and Goldstein M. Discriminant analysis. *Biometrics* 1979: 69–85.

11. Marks S and Dunn OJ. Discriminant functions when covariance matrices are unequal. *J Am Stat Assoc* 1974; **69**: 555–559.
12. Flury BW and Schmid MJ. Quadratic discriminant functions with constraints on the covariance matrices: some asymptotic results. *J Multivariate Anal* 1992; **40**: 244–261.
13. Wahl PW and Kronmal RA. Discriminant functions when covariances are unequal and sample sizes are moderate. *Biometrics* 1977: 479–484.
14. Komárek A, Hansen BE, Kuiper EM, et al. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Stat Med* 2010; **29**: 3267–3283.
15. Marshall G, De la Cruz-Mesía R, Quintana FA, et al. Discriminant analysis for longitudinal data with multiple continuous responses and possibly missing data. *Biometrics* 2009; **65**: 69–80.
16. Roy A. A new classification rule for incomplete doubly multivariate data using mixed effects model with performance comparisons on the imputed data. *Stat Med* 2006; **25**: 1715–1728.
17. Fieuws S, Verbeke G and Molenberghs G. Random-effects models for multivariate repeated measures. *Stat Meth Med Res* 2007; **16**: 387–397.
18. Liang K-Y and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986: 13–22.
19. Lipsitz SR, Fitzmaurice GM, Ibrahim JG, et al. Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune deficiency syndrome data. *J R Stat Soc: Series A (Statistics in Society)* 2009; **172**: 3–20.
20. Inan G and Yucel R. Joint gees for multivariate correlated data with incomplete binary outcomes. *J Appl Stat* 2017; **44**: 1920–1937.
21. Fong Y, Rue H and Wakefield J. Bayesian inference for generalized linear mixed models. *Biostatistics* 2010; **11**: 397–412.
22. Roy A and Khattree R. On implementation of a test for Kronecker product covariance structure for multivariate repeated measures data. *Stat Methodol* 2005; **2**: 297–306.
23. Srivastava MS, von Rosen T and Von Rosen D. Models with a Kronecker product covariance structure: estimation and testing. *Math Meth Stat* 2008; **17**: 357–370.
24. Lu N and Zimmerman DL. The likelihood ratio test for a separable covariance matrix. *Statistics & probability letters* 2005; **73**: 449–457.
25. Roy A. A note on testing of Kronecker product covariance structures for doubly multivariate data. In: Proceedings of the American Statistical Association, *Statistical Computing Section* 2007, pp.2157–2162. Publisher: American Statistical Association.
26. Roy A and Khattree R. Tests for mean and covariance structures relevant in repeated measures based discriminant analysis. *J Appl Stat Sci* 2003; **12**: 91–104.
27. Werner K, Jansson M and Stoica P. On estimation of covariance matrices with Kronecker product structure. *IEEE Trans Signal Process* 2008; **56**: 478–491.
28. Filipiak K, Klein D and Roy A. Score test for a separable covariance structure with the first component as compound symmetric correlation matrix. *J Multivariate Anal* 2016; **150**: 105–124.
29. Roy A and Khattree R. Testing the hypothesis of A Kronecker product covariance matrix in multivariate repeated measures data. In: Proceedings of the statistics and data analysis section. USA: SAS Users Group International, 2005, pp.199–130.
30. Cho H. The analysis of multivariate longitudinal data using multivariate marginal models. *J Multivariate Anal* 2016; **143**: 481–491.
31. Chao EC. *Generalized estimating equations*. Taylor & Francis, 2003.
32. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; **96**: 1348–1360.
33. Kauermann G and Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc* 2001; **96**: 1387–1396.
34. Wang M. Generalized estimating equations in longitudinal data analysis: a review and recent developments. *Adv Stat* 2014; **2014**: 1–11.
35. Roy A and Khattree R. Classification of multivariate repeated measures data with temporal autocorrelation. *J Appl Stat Sci* 2007; **15**: 283–294.
36. Roy A and Khattree R. Discrimination and classification with repeated measures data under different covariance structures. *Commun Stat – Simul Comput®* 2005; **34**: 167–178.
37. Roy A and Khattree R. On discrimination and classification with multivariate repeated measures data. *J Stat Plann Inference* 2005; **134**: 462–485.
38. Barön AE. Misclassification among methods used for multiple group discrimination – the effects of distributional properties. *Stat Med* 1991; **10**: 757–766.
39. He X and Fung WK. High breakdown estimation for multiple populations with applications to discriminant analysis. *J Multivariate Anal* 2000; **72**: 151–162.
40. Thomas DR and Zumbo BD. Using a measure of variable importance to investigate the standardization of discriminant coefficients. *J Educ Behav Stat* 1996; **21**: 110–130.

41. Cario MC and Nelson BL. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical Report, 1997. http://citeseerx.ist.psu.edu/viewdoc/similar?doi = 10.1.1.48.281&type = cc (accessed 30 November, 2018).
42. Speechley KN, Sang X, Levin S, et al. Assessing severity of epilepsy in children: preliminary evidence of validity and reliability of a single-item scale. *Epilepsy Behav* 2008; **13**: 337–342.
43. Qu A, Lindsay BG and Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika* 2000; **87**: 823–836.
44. Wang X and Qu A. Efficient classification for longitudinal data. *Comput Stat Data Anal* 2014; **78**: 119–134.
45. Asar Ö and İlk Ö. Mmm: an r package for analyzing multivariate longitudinal data with multivariate marginal models. *Computer Meth Programs Biomed* 2013; **112**: 649–654.
46. Lu J, Plataniotis KN and Venetsanopoulos AN. Regularized discriminant analysis for the small sample size problem in face recognition. *Pattern Recogn Lett* 2003; **24**: 3079–3087.
47. Pang H, Tong T and Ng M. Block-diagonal discriminant analysis and its bias-corrected rules. *Stat Appl Genet Mol Biol* 2013; **12**: 347–359.
48. Filzmoser P, Joossens K and Croux C. Multiple group linear discriminant analysis: robustness and error rate. In: A. Rizzi, M and Vichi, (eds). *Compstat 2006 – Proceedings in computational statistics*. Berlin: Springer; 2006: p.521–532.
49. Croux C, Filzmoser P and Joossens K. Classification efficiencies for robust linear discriminant analysis. *Statistica Sinica* 2008; 182(2): 581–599.
50. Molefe AC and Hosmane B. Test for link misspecification in dependent binary regression using generalized estimating equations. *J Stat Comput Simul* 2007; **77**: 95–107.
51. Naik DN and Rao SS. Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix. *J Appl Stat* 2001; **28**: 91–105.
52. Krzyśko M and Skorzybut M. Discriminant analysis of multivariate repeated measures data with a Kronecker product structured covariance matrices. *Statistical Papers* 2009; **50**: 817–835.
53. Zhou J and Qu A. Informative estimation and selection of correlation structure for longitudinal data. *J Am Stat Assoc* 2012; **107**: 701–710.
54. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001; **57**: 120–125.
55. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Parzen E., Tanabe K., Kitagawa G. (eds) *Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics)*. New York, NY: Springer 1998, pp. 199–213.
56. Beunckens C, Sotto C and Molenberghs G. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Comput Stat Data Anal* 2008; **52**: 1533–1548.
57. Satty A, Mwambi H and Molenberghs G. Different methods for handling incomplete longitudinal binary outcome due to missing at random dropout. *Stat Methodol* 2015; **24**: 12–27.
58. Yucel RM, He Y and Zaslavsky AM. Using calibration to improve rounding in imputation. *Am Stat* 2008; **62**: 125–129.
59. Yucel RM, He Y and Zaslavsky AM. Imputation of categorical variables using gaussian-based routines. *Stat Med* 2011; **30**: 3447–3460.