1 Introduction

Longitudinal discriminant analysis (LoDA) is a tool used to classify subjects into groups based on the evolution over time of some longitudinal variables. The basic idea is that information is collected repeatedly over time on some variable(s) (or marker), that are thought to be indicative of the group to which a subject belongs. Our particular motivation comes from a medical setting in which we want to use biomarker information collected over a series of clinic visits to inform classification of patients into prognostic groups based on their anticipated disease progression.

Over recent years methods of LoDA have developed from using a single continuous longitudinal marker in the discriminant analysis (Tomasko, Helms, and Snapinn, 1999, Brant, Sheng, Morrell, Verbeke, Lesaffre, and Carter, 2003, Kohlmann, Held, and Grunert, 2009) to allowing several longitudinal continuous markers (Morrell, Brant, Sheng, and Metter, 2012, Marshall, De la Cruz-Mesía, Quintana, and Barón, 2009, Komárek, Hansen, Kuiper, van Buuren, and Lesaffre, 2010). Further extensions have allowed LoDA using multiple longitudinal markers of different types (Fieuws, Verbeke, Maes, and Van Renterghem, 2008, Hughes, Komárek, Czanner, and Garcia-Fiñana, 2018b).

The basis for each of these approaches to LoDA is a mixed model. Models that only consider continuous markers utilise a (multivariate) linear (non-linear) mixed model to model the longitudinal evolution of the markers over time, whilst non-continuous markers can be incorporated within a multivariate generalized linear mixed model (MGLMM). Mixed models are fit to data from patients for whom we already know the prognostic group to which they belong, with one mixed model per group. The parameters from these mixed models are used within a discriminant analysis in order to predict the group membership of new patients.

A key feature of a mixed model is the inclusion of subject-specific random effects, with a joint distribution specified to incorporate the correlation between repeated observations of a single marker and also between observations of multiple markers for the same individual. A common assumption about the joint distribution of the random effects is that they follow a normal distribution.

To the best of our knowledge most assessment of the impact of misspecification of random effects has focused on parameter estimation. In the case of linear mixed models it has been shown that maximum-likelihood estimates are robust to misspecification of the random effects distribution (Verbeke and Lesaffre, 1997). However, in the case of GLMMs the picture is less clear. A general summary of findings is that parameter estimates are reasonably robust to random effects misspecifictation (Neuhaus, Mcculloch, and Boylan, 2011, Marquart and Haynes, 2019) but that in some cases, with a severe departure from normality, incorrect assumptions about the random effects structure can introduce substantial bias to

parameter estimates (Agresti, Caffo, and Ohman-Strickland, 2004, Litière, Alonso, and Molenberghs, 2008, Hernández and Giampaoli, 2018)

Some tests have been developed to diagnose and assess the suitability of random effects modelling assumptions (Zhang and Davidian, 2001, Abad, Litière, and Molenberghs, 2010, Drikvandi, Verbeke, and Molenberghs, 2017).

In work related to the aim of this paper, Albert (2012) and Liu and Albert (2014) consider shared random effects and pattern mixture models respectively to assess the impact of random effects misspecification on classification accuracy for longitudinal data. Both show that assuming a single multivariate normal distribution gives area under curve (AUC) values very close to the theoretical optimal AUC of the *true* model. Our LoDA procedure is similar to the pattern mixture approach. However, both of these papers consider only a single longitudinal marker, and do not consider whether fitting models with alternative random effects distributions would achieve better classification.

In this paper we explore the effect that misspecifying the random effects distribution has on the classification accuracy when the parameter estimates from mixed models are used for classification (specifically within a discriminant analysis model). Accepting that model parameters from a GLMM may be estimated with bias, we are interested in whether this potential bias affects our ability to classify patients into clinical groups using methods of LoDA. Secondly we investigate factors that may affect our ability to accurately estimate a random effects distribution such as sample size and number of repeated measurements.

Model misspecification may occur in other ways than assuming an incorrect random effects distribution. Kim and Kong (2016) and Kohlmann et al. (2009) investigated the consequences of misspecifying the structure of the random effects by assuming, for example a random intercept model when the true model contains a random intercept and random slope. De la Cruz, Meza, Arribas-Gil, and Carroll (2016) investigate the effect of misspecification of the residual errors and show that this kind of misspecification can noticably decrease the AUC obtained. The kind of model misspecification outlined in this paragraph is not the focus of this paper. We focus on the case where the distribution of the random effects is misspecified.

All the above referenced investigations into the effects of random effects misspecification only consider a single longitudinal response. However, in many clinical settings information about multiple longitudinal markers is collected for each patient. It is often desirable to use more than one of these markers to inform clinical decision making. By using multiple markers, we considerably increase the number of random effects considered (in most cases) and we investigate conditions in which these more complex, and higher-dimensional distributions can be estimated accurately to improve classification.

The rest of this paper is organised as follows. We first give a brief overview

of the MGLMM used to model multiple longitudinal markers in Section 2. We explain how the parameter estimates from the MGLMM are used in a discriminant analysis to allow classification of patients in Section 3. We analyse data from a study of primary biliary cirrhosis, and of hepatocellular carcinoma to explore the effect of the choice of random-effects distribution in Sections 4 and 5 respectively. In Section 6 we present the results of a simulation study investigating the effects of random effects misspecification. We provide some intuition about situations in which misspecifying random effects distributions may be problematic in Section 7 and we conclude with a short summary in Section 8.

2 Multivariate generalized linear mixed models

We consider the collection of data on $R \geq 1$ biomarkers at times $\mathbf{t}_r = \left(t_{r,1}, \ldots, t_{r,n_r}\right)$, $t_{r,1} < \cdots < t_{r,n_r} < T$, $r = 1, \ldots, R$. Note that each biomarker does not need to be measured at the same time, and that patients do not necessarily have identical time schedules. The observations of biomarker r for a particular patient are denoted by $\mathbf{Y}_r = \left(Y_{r,1}, \ldots, Y_{r,n_r}\right), r = 1, \ldots, R$. The value of each biomarker may further depend upon additional covariates (those collected at baseline for example), denoted by $\mathscr{C}$.

Our aim in this paper is to use the biomarker data collected until some time point $t < T$ to predict the status of each patient at time T. To do so we require some training data for whom the status at T is known. Specifically, we know the group, $U \in \left\{0, \ldots, G-1\right\}$ to which the patient belongs at time T. A separate MGLMM is fit to each group, where the expected value of the j'th observation ($j = 1, \ldots, n_r$) of the r'th marker of a patient in group g (denoted $Y_{r,j}$) is given by

$$h_r^{-1}\left\{\mathsf{E}\left(Y_{r,j} \,\middle|\, \mathbf{b}, U = g\right)\right\} = \mathbf{x}_{r,j}^{g\top} \boldsymbol{\alpha}_r^g + \mathbf{z}_{r,j}^{g\top} \mathbf{b}_r, \tag{1}$$

where h_r^{-1} is a chosen link function (for example the logit function for binomial responses, log function for Poisson responses and the identity function for Gaussian response). Covariate information is contained in $\mathbf{x}_{r,j}^g = \mathbf{x}_{r,j}^g(\mathscr{C})$ and $\mathbf{z}_{r,j}^g = \mathbf{z}_{r,j}^g(\mathscr{C})$ for each prognostic group g. The vector of regression coefficients to be estimated are denoted $\boldsymbol{\alpha}_r^g$, $r = 1, \ldots, R$, $g = 0, \ldots, G-1$.

Correlations between repeated measurements of a biomarker, and between values of different biomarkers for a particular patient are modelled using an unobserved random effects vector $\mathbf{b} = \left(\mathbf{b}_1, \ldots, \mathbf{b}_R\right)$. It is typical to assume that the random effects vector jointly follows a normal distribution in each prognostic group. An alternative, that allows greater flexibility specifies a weighted mixture of K normal distributions, with mean $\boldsymbol{\mu}_k$ and covariance matrix $\mathbb{D}_k$ for the random effects joint distribution (see Hughes et al. (2018b), Komárek et al. (2010), Komárek and

Komárková (2013) for full details of this model.) Often, due to the complexity of the multivariate mixed models under consideration, a Markov chain Monte Carlo (MCMC) scheme is used to estimate the model parameters.

Our challenge in this paper is to investigate whether the choice of distribution for the random effects influences the accuracy of the classification achieved. As discussed in the introduction, misspecification of the random effects distribution does not usually have much effect on the accuracy of the parameter estimates. However, Komárek et al. (2010) present an example in which using a mixture distribution improves the classification accuracy achieved in comparison to the standard normal distribution assumption. To investigate how robust classification approaches are to misspecification of the random effects distribution, we first describe in Section 3 how classifications into prognostic groups are obtained using the MGLMM parameters in a longitudinal discriminant analysis (LoDA).

3 Longitudinal discriminant analysis

Our aim is to use the model parameters from each MGLMM to classify new patients given their clinical history. Classification involves first a calculation of the probability that the patient belongs to a particular group g, given their longitudinal data and covariate information. This can be calculated using Bayes' Theorem

$$\mathscr{P}_{g,new} = \frac{\pi_g \hat{f}_{g,new}}{\sum_{\tilde{g}=0}^{G-1} \pi_{\tilde{g}} \hat{f}_{\tilde{g},new}} \qquad g = 0, \ldots, G-1. \tag{2}$$

where $\hat{f}$ is a predictive density which assesses the likelihood of the observed markers given the group and model parameters. The prior probability of belonging to group g are denoted by π_g, and is commonly assumed to be the prevalence of the group in the study population. In a frequentist setting (which covers most of the references discussed in the introduction regarding misspecification of random effects distributions), $f_{g,new}$ is estimated using the maximum likelihood estimates of the relevant model parameters in group g. In more complex models such an approach can be computationally challenging, so in this paper we take a different approach and estimate $f_{g,new}$ by calculating the mean of the posterior predictive density estimated for each of M samples from a MCMC scheme.

Morrell, Brant, and Sheng (2007) propose three different ways of specifying the predictive density $f_{g,new}$, specifically a marginal prediction (which compares the new patient's profile to the group specific average profiles computed using the MGLMMs in each group), a conditional prediction (which estimates a new patient's random effects and compares their longitudinal profiles with patients with similar

random effects in each group) and a random effects prediction (which compares the new patient's estimated random effects with the mean random effects distribution in each prognostic group). These three methods have been compared previously (Morrell, Sheng, and Brant, 2011, Komárek et al., 2010, Hughes, El Saeiti, and García-Fiñana, 2018a) and the random effects and marginal approaches have been seen to be the most promising, so these are the focus of our work in this paper.

The marginal prediction approach calculates the marginal predictive density

$$f_g^{marg}\left(\mathbf{y}_1, \ldots, \mathbf{y}_R; \boldsymbol{\psi}^g, \boldsymbol{\theta}^g, \mathscr{C}\right) = \int \prod_{r=1}^{R} \prod_{j=1}^{n_r} p_r\left(y_{r,j} \,\middle|\, \mathbf{b}; \boldsymbol{\psi}^g, \mathscr{C}\right) f_g^{re}\left(\mathbf{b}; \boldsymbol{\theta}^g\right) \mathrm{d}\mathbf{b},$$

where f_g^{re} is the estimated density of the random effects distribution in group g, and is given by a weighted sum of normal distributions.

$$f_g^{re}\left(\mathbf{b}; \boldsymbol{\theta}^g\right) = \sum_{k=1}^{K^g} w_k^g \, \varphi(\mathbf{b}; \boldsymbol{\mu}_k^g, \mathbb{D}_k^g).$$

Here, $K^g = 1$ corresponds to the typical assumption of a normal distribution for the random effects and $K^g > 1$ corresponds to the mixture distributions. $\boldsymbol{\psi}^g$ denotes all the fixed effects parameters whilst $\boldsymbol{\theta}^g$ denotes all the parameters related to the random effects distribution. Once these parameters have been calculated, the marginal group membership probabilities can be calculated for each sample from the MCMC procedure and then averaged to give the final probability, using Equation 2. The random effects group membership probabilities can be calculated similarly, by replacing f_g^{marg} with f_g^{re} in Equation 2.

A new patient is classified as belonging to a prognostic group g, if $\mathscr{P}_{g,new}$ is greater than a chosen threshold. In the two-group case presented in this paper, the threshold is chosen using the point closest to the top-left corner of a ROC plot, as is standard in many classification procedures. Many alternatives are available (see Hughes, Komárek, Bonnett, Czanner, and García-Fiñana (2017) for comparisons of various options for choosing a threshold).

4 PBC example

To illustrate how choice of distribution for the random effects affects classification accuracy we present an example based on the Mayo clinic Primary Biliary Cirrhosis (PBC) dataset (Dickson, Grambsch, Fleming, Fisher, and Langworthy, 1989). This data is publicly available within the `mixAK` (Komárek and Komárková, 2014) package in `R` (R Core Team, 2016) (The data is also available in Appendix D of Fleming and Harrington (1991) and also electronically at http://lib.stat.cmu.edu/datasets/pbcseq).

The initial study aimed to investigate whether treating patients with D-penicillamine increased the length of patient survival. Data on a large number of clinical variables were recorded for 312 patients over a median of 6.3 years per patient.

Komárek et al. (2010) also explored PBC, although using data from the Dutch Multicenter Primary Biliary Cirrhosis study. They used three continuous markers, bilirubin, albumin and alkaline phosphotase (all available within the Mayo PBC data) and showed that using a mixture distribution for the random effects with $K = 2$, gave a better Area under Curve (AUC) than a single normal distribution.

We present here an application of multivariate LoDA using continuous, binary and Poisson markers to the Mayo PBC data. We use data collected in the first 2.5 years of follow up to predict whether a patient will die or require liver transplant in the following 2.5 years (i.e. within 5 years of their initial recruitment to the trial). There were 253 patients known to be alive two and a half years after recruitment. 51 of these died or required a liver transplant at some point in the following 2.5 years. We considered three longitudinal markers, log(bilirubin) (a continuous marker), platelet count (Poisson, counted per cubic ml/1000) and blood vessel malformation (binary). Log(bilirubin) and platelet count were modelled using a random intercept and slope. There were an average of 3.53 visits per patient with bilirubin measured on every visit, and an average of 3.47 measurements of platelet count, and 3.51 assessments of blood vessel malformations per patient.

The model for the binary blood vessel malformations included a random intercept and a fixed time slope (largely for the sake of numerical stability). We considered four potential LoDA classifiers, a model using the typical assumption of a single normal distribution for the random effects ($K = 1$), and models using 2, 3 and 4 component mixture distributions for the random effects distribution in each prognostic group (those who die or require liver transplant (Group 1) and those who do not (Group 0)).

We compared the fit of the model to the data using penalised expected deviance (PED, Plummer (2008)). Lower PED values indicate better model fit. Table 1 shows that for the patients who do not die or require a liver transplant a 2-component mixture distribution for the random effects gives a slightly better model fit. However, for patients in Group 1, a single normal distribution gives the best PED. We note that we have consistently seen that models trained on small numbers of individuals favor simpler models.

Leave one out cross-validation was used to obtain predictions for each of the 253 patients in our sample. The classification accuracy results using the marginal approach (which was best for the PBC data) are shown in Table 1. It is clear that in a small sample, complex models involving $K = 3, 4$ mixture components are unsuitable. There is not much difference between using a single normal distribution or a two-component mixture with both achieving similar classification results. The

model utilising the typical assumption ($K = 1$) has a slightly better AUC, indicating better performance. ROC curves for the 4 models are shown in Supplementary Figure 1. At the optimal threshold the $K = 2$ model achieves slightly better sensitivity, but worse specificity and probability of correct classification (PCC). We conclude in this application that using a more flexible distribution for the random effects does not improve classification accuracy.

Table 1: Model performance and prediction accuracy using PBC data for models with $K = 1, 2, 3, 4$ mixture components. Sens=Sensitivity, Spec=Specificity, PCC=Probability of Correct Classification, AUC=Area Under Curve, PPV = Positive Predictive Value, NPV = Negative Predictive Value. In this, and all following tables, the cutoff is the value of the threshold that gave results closest to the top left corner of the ROC plot. All results are reported at this cutoff value.

Model	PED		Classification Accuracy						
	Group 0	Group 1	Cutoff	Sens	Spec	PCC	AUC	PPV	NPV
K = 1	11112.80	**2987.29**	0.20	0.78	0.82	0.81	0.86	0.53	0.94
K = 2	**11021.41**	3469.04	0.07	0.82	0.73	0.75	0.84	0.43	0.94
K = 3	11046.15	4439.53	0.01	0.65	0.81	0.78	0.74	0.46	0.90
K = 4	11160.76	4547.39	0.02	0.65	0.62	0.62	0.64	0.30	0.87

5 Hepatocellular carcinoma example

We further demonstrate the influence that choice of random effects distribution has on classification accuracy in a screening study for hepatocellular carcinoma (HCC). Our dataset comes from the Ogaki municipal hospital in Japan. The dataset under consideration in this paper consists of 3333 patients with longitudinal measurements of alpha-fetoprotein (AFP), Des-gamma-carboxy prothrombin (DCP) (modelled as continuous longitudinal markers) and platelet counts (again modelled as a Poisson longitudinal marker for this application). The measurements were collected at regular screening visits for the early detection of HCC.

Our dataset consists of 395 patients who develop HCC whilst under observation and 2938 who did not. Note that some of these patients may have gone on to develop HCC in the future, but for the purposes of this investigation are considered as non-HCC patients. Patients had an average of 23.43 clinic visits, with AFP and DCP measured at each visit, and an average of 22.52 platelet measurements per patient. Profile plots of these three markers are shown in Supplementary Figure 3

In this analysis, we log transformed AFP and DCP measurements and considered a random intercepts and random slopes model with each marker also having fixed effects for the age at first screening and gender. See Hughes, Berhane, de Groot, Toyoda, Tada, Kumada, Satomura, Nishida, Kudo, Miura, Osaki, Kolamunage-Dona, Amoros Salvador, Bird, García-Fiñana, and Johnson (2020) for further details of this cohort with a model that only considers longitudinal AFP measurements. We removed 117 pregnant patients for this analysis as pregnancy is known to influence DCP levels. We considered a dynamic allocation scheme whereby each clinic visit was considered in turn. If a patient's risk was over the chosen threshold they were allocated to the HCC group and removed for further investigations. If the patient's risk was below the chosen risk their next visit was considered and their predicted group membership probabilities updated using the additional information. This was continued until all visits had been considered or a patient was classified into the HCC group. So a patient was considered an HCC case if *any* of their predicted risks were above the chosen threshold and a non-HCC case if *none* of their risks were above the threshold. This scheme is similar to prostate cancer prediction scheme of Brant et al. (2003).

Table 2 reports the prediction accuracy for each model. The model with a two-component mixture of multivariate normal distributions gave the best sensitivity, specificity, PCC, PPV and NPV. The improvement in classification accuracy is not substantial with more flexible models in this case, although both the $K = 2$ and $K = 4$ models achieved slightly better sensitivity and specificity than the model with the standard single multivariate normal distribution assumption. In the non-HCC group the models with more flexible random effects distributions achieved better PED indicating better model fit. For HCC patients, the model with a two and three-component mixture of normal distributions achieved better PED than the standard multivariate normal distribution. However, despite more flexible models providing better model fit, the improvement in classification accuracy was minimal.

Table 2: Model performance and prediction accuracy using HCC data (AFP+DCP+Platelet count) for models with $K = 1, 2, 3, 4$ mixture components.

Model	PED		Classification Accuracy						
	Group 0	Group 1	Cutoff	Sens	Spec	PCC	AUC	PPV	NPV
K = 1	621655.4	75135.5	0.18	0.79	0.74	0.75	0.83	0.29	0.96
K = 2	613887.4	**74589.1**	0.31	0.80	0.76	0.77	0.83	0.31	0.97
K = 3	612935.3	74686.3	0.23	0.79	0.74	0.74	0.82	0.29	0.96
K = 4	**612768.6**	84098.4	0.12	0.80	0.75	0.76	0.83	0.30	0.96

These results are in contrast to the findings of Komárek et al. (2010) who

show an improvement in classification when using mixture distributions. This suggests that it is currently unclear what effect random effects misspecification has on classification accuracy, as with the uncertainty over the impact on parameter estimates in GLMMs. To investigate this problem further we conduct a simulation study in Section 6 aiming to determine how robust classification accuracy is to random effects misspecification.

6 Simulation study

6.1 Simulation Design

We designed a simulation set up similar to that of the PBC data described in Section 4, but adjusted where necessary to investigate scenarios where random effects misspecification might lead to poorer classification results. We considered two overall sample sizes, $n = (250, 2500)$, in each case keeping the prevalence of Group 1 at 20% to reflect the PBC data. We simulated longitudinal profiles for each patient for each of the three biomarkers considered in Section 4, log(bilirubin), platelet count and blood vessel malformations. We simulated four clinic visits per patient (to roughly correspond to the average of 3.53 per patient in the PBC data) and each biomarker was measured at each visit. The first visit occurred at $t = 0$, and then the remaining visit times were generated from uniform distributions in the intervals (170,200), (350,390) and (710,770) days, representing approximately a follow up visit at 6 months, 1 year and 2 years.

We considered a number of simulation scenarios. We first considered the case where the true distribution of the random effects is a single normal distribution (the typical assumption) to explore whether using a mixture had an adverse effect on the classification accuracy. In this case we used a model fit to the PBC data (Supplementary Table 1) to provide parameter values for our simulations

Next we considered 2 and 3-component mixtures as the true random effects distribution. In each case we considered two scenarios. First we considered a scenario with parameter estimates from models fit to the PBC data (See Table 2 and 4 in the supplementary material). In these cases the amount of departure from normality was small and it was not always clear from a visual inspection that the "true" distribution was in fact a 2-component mixture. This setup was designed to investigate the effect of only small departure from normality in the random effects. Secondly we considered a much more severe departure from normality, where the two groups had 2 (or 3) component mixture distributions with identically positioned means but different variances. This meant the only difference between the two groups was the spread around the component means.

Supplementary Figures 4 and 5 show the shape of the assumed random effects distributions for the scenarios where a 2-component mixture was considered the "true" distribution for small and large departures from a single multivariate normal distribution respectively. Note that the plots give no impression of the correlation between different random effects and are simply shown to give an indication of the different shapes of random effects in each group and how severe the departure from normality is. Similar plots for the remaining simulation scenarios are shown in Figures 6-9 of the supplementary material, where details of correlation between random effects is shown in the respective tables of true parameter values.

Finally we considered the case where the true random effects distribution was a T-distribution with 3 or 5 degrees of freedom, again corresponding to larger and smaller departures from normality. This led to 7 different simulation scenarios for each of the two sample sizes. The parameters used for each simulation scenario are presented in Tables 1-5 of the supplementary material.

For the sample size of 250 patients we also considered the effect of changing the number of visits. We hypothesised that with more data, the random effects could be estimated more accurately and this may improve classification accuracy. In particular, for the $K = 1$, $K = 2$ with severe departure and T-distribution with 3 degrees of freedom scenarios, we repeated the simulations but with 9 equally spaced visits instead of 4 (corresponding to follow up approximately every 3 months. The visits were generated from uniform distributions in the intervals (70,110), (160,200), (250,290), (345,385), (430,470), (520,560), (610,650) and (710,750) days.

For each scenario we applied the LoDA approach with each of $K = 1, 2, 3, 4$ models to assess whether more flexible models could better handle random effects misspecification, and how well the $K = 1$ model performed even in situations where the truth was not a single multivariate normal distribution for the random effects.

For the simulation scenarios with 250 patients, leave one out cross-validation was used to assess the classification accuracy for each model, whilst for the scenarios with 2500 patients, each simulated dataset was split into 100 training sets of 70% of the patients in each group, and test sets of the remaining 30%, and averaged the classification accuracy results.

For each scenario we simulated 50 datasets and compared classification accuracy measures, namely AUC, sensitivity, specificity and PCC. In each scenario MCMC was used to estimate the models with 15,000 MCMC samples with a thinning factor of 10 and the first 5000 samples discarded as burn in. We present boxplots showing the spread of AUC, sensitivity, specificity and PCC to show the variability across simulated datasets. These methods are computationally intensive, and 50 simulated datasets for each scenario was thought to be a reasonable balance between reliability of results, and computational efficiency

6.2 Misspecification of Random Effects

6.2.1 Effect on marginal prediction

When the true random effects distribution in each group was a 2-component mixture of Gaussians, small departures from normality show little difference in prediction accuracy between a theoretically correct model with $K = 2$ mixture components and the standard assumption of a single multivariate normal distribution (Table 3 and Supplementary Figure 11). The models with $K = 3, 4$ mixture components were unstable and did not preform as well as the simpler models, possibly due the small sample size in relation to the complexity of the model. Increasing the sample size to 2500 gave an increase in AUC and other accuracy measures.

When the departure from normality was more severe, then Table 3 and Supplementary Figure 12 show a consistent improvement by using the model with $K = 2$ mixture components. This is due to the fact that the two component mixture is able to detect the difference in variability between the two disease groups, and the locations of the two component means, whereas a single normal distribution is unable to capture this detail.

Table 3: Prediction accuracy of marginal prediction under the assumption that the random effects jointly follow a 2-component normal mixture distribution. Prediction accuracies are reported as the mean over all simulated datasets

	Model	Cut	Sens	Spec	PCC	AUC	PPV	NPV
Small Departure from Normality Assumption								
250	1	0.15	0.87	0.89	0.89	0.94	0.68	0.96
	2	0.17	0.88	0.90	0.89	0.95	0.69	0.97
	3	0.71	0.80	0.77	0.78	0.82	0.49	0.94
	4	0.76	0.80	0.82	0.81	0.85	0.53	0.94
2,500	1	0.12	0.94	0.97	0.96	0.98	0.88	0.99
	2	0.17	0.96	0.97	0.97	0.99	0.90	0.99
	3	0.30	0.92	0.93	0.93	0.94	0.82	0.98
	4	0.46	0.88	0.90	0.90	0.92	0.73	0.97
Large Departure from Normality Assumption								
250	1	0.18	0.83	0.87	0.86	0.91	0.62	0.95
	2	0.17	0.89	0.92	0.92	0.95	0.75	0.97
	3	0.40	0.82	0.81	0.82	0.85	0.55	0.95
	4	0.63	0.77	0.72	0.73	0.77	0.42	0.93
2,500	1	0.16	0.87	0.90	0.89	0.94	0.69	0.96
	2	0.17	0.92	0.95	0.94	0.98	0.81	0.98
	3	0.67	0.86	0.82	0.83	0.88	0.58	0.96
	4	0.55	0.81	0.80	0.80	0.84	0.54	0.94

We observed very similar findings when the true random effects distribution

was a 3-component mixture. More flexible models were only a benefit when the departure from normality was more severe, and the sample size was large (See Figure 1, Supplementary Figure 13 and Supplementary Table 7).

When the underlying random effects distribution was truly a normal distribution, as expected, a single multivariate normal distribution gave the most accurate prediction accuracy, both in terms of AUC and in terms of optimal sensitivity and specificity. However, there was very little difference in prediction accuracy or in the ROC curves if we assumed a 2-component mixture of Gaussians for the random effects distribution (See Table 6 and Figure 10 in the supplementary material). More complicated mixture distributions did not aid the prediction accuracy and produced less accurate results, much less stable estimates of sensitivity and specificity (shown by wider boxplots) for 3 and 4 component mixture models. We suspect this is largely due to the fact that such a complicated model is unwarranted in a small group of patients, especially considering that the smaller group only has 50 patients.

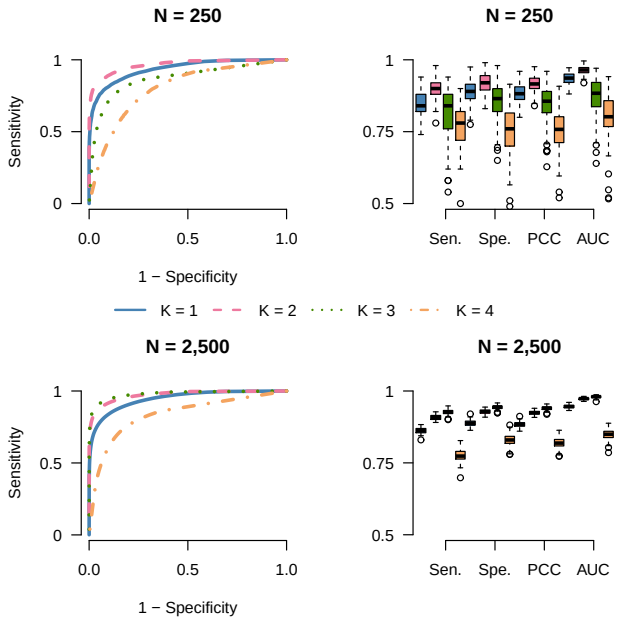

Figure 1: Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a three component mixture of normals with large departure from normality. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures across 100 simulated datasets.

Finally we considered the case where the random effects followed a t-distribution. Table 4 shows that when the degrees of freedom of the t-distribution is small, and hence departure from normality more severe, using a mixture-distribution gave more accurate predictions. This was specially clear in the case with 2500 patients, where all three mixture distributions achieved at least a 5% better AUC than assuming a single normal distribution (See Supplementary Figure 14). The picture was less clear as the degrees of freedom was increased. The $K = 2$ mixture distribution still performed the best in this scenario but the improvement over the single multivariate normal distribution was small (Supplementary Figure 15).

Table 4: Prediction accuracy of marginal prediction under the assumption that the random effects jointly follow a t-distribution with 3 (top two sections) and 5 (bottom two sections) degrees of freedom. Prediction accuracies are reported as the mean over all simulated datasets

Size	K	Cut	Sens	Spec	PCC	AUC	PPV	NPV
3 degrees of freedom								
250	1	0.14	0.73	0.74	0.74	0.77	0.46	0.92
	2	0.16	0.76	0.77	0.77	0.81	0.46	0.93
	3	0.17	0.73	0.75	0.74	0.79	0.44	0.92
	4	0.23	0.70	0.70	0.70	0.74	0.39	0.90
2,500	1	0.13	0.73	0.73	0.73	0.74	0.47	0.93
	2	0.18	0.76	0.75	0.75	0.79	0.48	0.94
	3	0.17	0.77	0.77	0.77	0.83	0.47	0.93
	4	0.17	0.74	0.75	0.75	0.80	0.43	0.92
5 degrees of freedom								
250	1	0.16	0.81	0.81	0.81	0.86	0.53	0.95
	2	0.15	0.81	0.82	0.82	0.87	0.53	0.95
	3	0.22	0.78	0.78	0.78	0.83	0.48	0.93
	4	0.24	0.77	0.76	0.76	0.81	0.46	0.93
2,500	1	0.16	0.81	0.81	0.81	0.86	0.54	0.95
	2	0.19	0.81	0.82	0.81	0.88	0.54	0.95
	3	0.17	0.81	0.81	0.81	0.88	0.52	0.94
	4	0.20	0.79	0.79	0.79	0.86	0.50	0.94

To summarise our findings for the marginal prediction approach, we have found that as long as the departure from normality is not large, there is little loss in assuming a single multivariate normal distribution for the distribution of random effects in each group. However, we also note that assuming a $K = 2$ mixture distribution almost always performed comparably and there was no loss in assuming this slightly more flexible model. We also found this to be the case in simulations not shown here where the departure from normality was more severe but the location of the means of mixture components was different between groups. In these cases a single normal distribution could capture differences between the groups suitably

well to allow accurate classification, even if the estimates of the random effects were not accurate. All that was needed for accurate classification was an indication of which location the new patient was 'closer' to.

When the departure from normality was more severe, and there were differing variabilities between groups then there was a benefit to assuming more flexible distributions. Using a 2-component mixture could improve the AUC and lead to more accurate predictions. However, the improvement was not always substantial.

6.2.2 Effect on Random Effects Prediction

So far, we have focused on the marginal prediction approach since that was found to be the most accurate for the PBC data. However, (Komárek et al., 2010) show an example in which random effects prediction gave more accurate classification. In such a case, where the estimated random effects are actually being used in the prediction (as opposed to the marginal approach where they are integrated out), it may be the case that misspecified random effects are more costly in terms of prediction accuracy, and it would be important to ensure that the random effects were estimated accurately. One way in which this could be achieved would be to include more visits per patient, to get a better idea of individual patient trajectories. We investigated this aspect by considering the $K = 1$, $K = 2$ with severe departure and T-distribution with 3 degrees of freedom simulation scenarios and focused on the random effects prediction results with either 4 or 9 visits per patient. This 9 visit schedule was designed to correspond to approximately visits every three months. The results for the latter two scenarios are shown in Table 5, with the results for the $K = 1$ true scenario shown in Supplementary Table 8.

In each of the scenarios considered, the dataset with 9 visits per patient achieved greater classification accuracy of the random effects prediction approach than the 4 visit dataset. With more visits per patient, a more accurate estimate of the patient specific intercepts and slopes could be obtained. Similarly, having more patients allowed more accurate classification, again due to improved accuracy of random effects estimates. In the example we present here, there is no clear benefit of using mixture distributions instead of a single normal distribution. However, we do note again that in the case where the true distribution is a 2-component mixture and in the case of the larger sample size of 2,500 patients, using the $K = 2$ model does improve the AUC slightly and the specificity, PCC and PPV noticeably. Although the benefit in this case is arguable (and clearly not a lot is lost by incorrectly assuming a single normal distribution for the random effects), this suggests that there will be cases where with large enough sample sizes and sufficient departure from normality, using more complex models will allow more accurate prediction.

7 Factors affecting the impact of misspecification of random effects distributions

We have shown in this paper that in many situations, random effects misspecification only has a very small impact on classification accuracy. However, there are cases where the departure from normality is more severe, and in such cases, failing to acknowledge this in model building can lead to losses in classification accuracy. This is consistent with the literature on the impact of random effects misspecification on parameter estimates referred to in the introduction. It is possible to get sufficiently accurate estimates of equation 2 even when individual fixed effects may be biased due to random effects misspecification. Conversely in some cases more flexible modeling of the random effects distribution can lead to more accurate estimate of $f_{g,new}$ and in turn, improved classification accuracy. The amount of impact that misspecification of random effects distributions has on classification is complex to determine, but we believe is influenced by at least two factors.

1. **The distance between the prognostic groups.** A key factor is how separated the two groups of longitudinal profiles are. If the two groups are well separated then misspecification of random effects distributions may make little difference. For example, even if the true random effects distribution was a 3-component mixture in each group, if the groups are well separated, a single multivariate normal distribution would probably be sufficient to determine which region of the sample space, a patient belonged to. Even though the wrong random effects distribution had been chosen, equation 2 would be sufficiently well estimated to allow accurate classification. We are not aware of an automatic way to determine the separability of the groups a-priori. Plotting of the longitudinal profiles would allow an initial investigation of the amount of overlap between groups, and provide an indication about the likely usefulness of more flexible random effects distributions.
2. **The amount of divergence from normality within a group.** If the amount of divergence from normality within a group is small, then the effect of misspecifiying the random effects distribution will be minimal. Estimates of equation 2 will be approximately equal in both models, at least as sample size increases. Researchers could investigate this by using one of the tests derived for assessing random effects misspecification (Drikvandi et al. (2017) and references therein). A particular benefit of the Drikvandi et al. test, is that the use of the gradient function derived by Verbeke and Molenberghs (2013) could allow the user to assess how severe the misspecification is, by observing how much greater than 1, the gradient function is.

The impact of misspecification on classification accuracy is likely to be a complex interplay between these two factors. If the groups are well separated then even if the within group divergence from normality is large, then accurate classification could be obtain from misspecified (but simpler) models. Equally, if the separation between the two groups is small, but the amount of divergence from normality is large, and especially if the divergence is different in each group, then more accurate classifications could be obtained by more flexible models, because the random effects densities are estimated more accurately.

8 Summary

In this paper we presented extensive simulation results to investigate the effect on classification accuracy of misspecification of the random effects distribution in longitudinal discriminant analysis. We have shown that if there is a *small* departure from normality, then assuming a single normal distribution will not lead to substantially less accurate classifications. This is consistent with the general findings mentioned in the introduction. Although some parameters may be estimated with bias, this does not impact classification much.

In contrast, when the departure from normality is more severe, we have shown (in agreement with Komárek et al. (2010)) that more accurate classifications can be obtained by assuming a more flexible random effects distribution.

If more complex models are to be considered, then we have shown that larger sample sizes must also be available. Since assuming a 3-component mixture distribution substantially increases the number of parameters to be estimated from a model, then the number of patients must be correspondingly large. In large datasets, the user has more freedom to consider more complex, flexible distributions for the random effects in order to guard against misspecifiction. In smaller datasets, even if misspecification is suspected, a single multivariate normal distribution may perform just as well. Similarly, more observations per patient are expected to lead to more accurate classification when using the random-effects prediction approach.

9 Supporting Information

Supplementary material referenced in Section 6, is available online.

References

Abad, A. A., S. Litière, and G. Molenberghs (2010): "Testing for misspecification in generalized linear mixed models," *Biostatistics*, 11, 771–786.

Agresti, A., B. Caffo, and P. Ohman-Strickland (2004): "Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies," *Computational Statistics & Data Analysis*, 47, 639–653.

Albert, P. S. (2012): "A linear mixed model for predicting a binary event from longitudinal data under random effects misspecification," *Statistics in medicine*, 31, 143–154.

Brant, L. J., S. L. Sheng, C. H. Morrell, G. N. Verbeke, E. Lesaffre, and H. B. Carter (2003): "Screening for prostate cancer by using random-effects models," *Journal of the Royal Statistical Society, Series A*, 166, 51–62.

De la Cruz, R., C. Meza, A. Arribas-Gil, and R. J. Carroll (2016): "Bayesian regression analysis of data with random effects covariates from nonlinear longitudinal measurements," *Journal of multivariate analysis*, 143, 94–106.

Dickson, E. R., P. M. Grambsch, T. R. Fleming, L. D. Fisher, and A. Langworthy (1989): "Prognosis in primary biliary cirrhosis: model for decision making," *Hepatology*, 10, 1–7.

Drikvandi, R., G. Verbeke, and G. Molenberghs (2017): "Diagnosing misspecification of the random-effects distribution in mixed models," *Biometrics*, 73, 63–71.

Fieuws, S., G. Verbeke, B. Maes, and Y. Van Renterghem (2008): "Predicting renal graft failure using multivariate longitudinal profiles," *Biostatistics*, 9, 419–431.

Fleming, T. R. and D. P. Harrington (1991): *Counting processes and survival analysis*, volume 169, John Wiley & Sons.

Hernández, F. and V. Giampaoli (2018): "The impact of misspecified random effect distribution in a weibull regression mixed model," *Stats*, 1, 48–76.

Hughes, D. M., S. Berhane, C. E. de Groot, H. Toyoda, T. Tada, T. Kumada, S. Satomura, N. Nishida, M. Kudo, T. Miura, Y. Osaki, R. Kolamunage-Dona, R. Amoros Salvador, T. Bird, M. García-Fiñana, and P. Johnson (2020): "Serum levels of alpha fetoprotein increase more than 10 years before detection of hepatocellular carcinoma," *Clinical Gastroenterology and Hepatology*, In Press.

Hughes, D. M., R. El Saeiti, and M. García-Fiñana (2018a): "A comparison of group prediction approaches in longitudinal discriminant analysis," *Biometrical Journal*, 60, 307–322.

Hughes, D. M., A. Komárek, L. J. Bonnett, G. Czanner, and M. García-Fiñana (2017): "Dynamic classification using credible intervals in longitudinal discriminant analysis," *Statistics in medicine*, 36, 3858–3874.

Hughes, D. M., A. Komárek, G. Czanner, and M. Garcia-Fiñana (2018b): "Dynamic longitudinal discriminant analysis using multiple longitudinal markers of

different types," *Statistical methods in medical research*, 27, 2060–2080.

Kim, Y. and L. Kong (2016): "Classification using longitudinal trajectory of biomarker in the presence of detection limits," *Statistical methods in medical research*, 25, 458–471.

Kohlmann, M., L. Held, and V. P. Grunert (2009): "Classification of therapy resistance based on longitudinal biomarker profiles," *Biometrical Journal*, 51, 610–626.

Komárek, A., B. E. Hansen, E. M. Kuiper, H. R. van Buuren, and E. Lesaffre (2010): "Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution," *Statistics in Medicine*, 29, 3267–3283.

Komárek, A. and L. Komárková (2013): "Clustering for multivariate continuous and discrete longitudinal data," *The Annals of Applied Statistics*, 7, 177–200.

Komárek, A. and L. Komárková (2014): "Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data," *Journal of Statistical Software*, 59, 1–38, URL http://www.jstatsoft.org/v59/i12/.

Litière, S., A. Alonso, and G. Molenberghs (2008): "The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models," *Statistics in Medicine*, 27, 3125–3144.

Liu, D. and P. S. Albert (2014): "Combination of longitudinal biomarkers in predicting binary events," *Biostatistics*, 15, 706–718.

Marquart, L. and M. Haynes (2019): "Misspecification of multimodal random-effect distributions in logistic mixed models for panel survey data," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182, 305–321.

Marshall, G., R. De la Cruz-Mesía, F. A. Quintana, and A. E. Barón (2009): "Discriminant analysis for longitudinal data with multiple continuous responses and possibly missing data," *Biometrics*, 65, 69–80.

Morrell, C. H., L. J. Brant, and S. Sheng (2007): "Comparing approaches for predicting prostate cancer from longitudinal data," in *2007 Proceedings of the American Statistical Association*, Biometrics Section, Alexandria: American Statistical Association, 127–133.

Morrell, C. H., L. J. Brant, S. Sheng, and E. J. Metter (2012): "Screening for prostate cancer using multivariate mixed-effects models," *Journal of Applied Statistics*, 39, 1151–1175.

Morrell, C. H., S. L. Sheng, and L. J. Brant (2011): "A comparative study of approaches for predicting prostate cancer from longitudinal data," *Communications in Statistics-Simulation and Computation*, 40, 1494–1513.

Neuhaus, J., C. Mcculloch, and R. Boylan (2011): "A note on type ii error under

random effects misspecification in generalized linear mixed models," *Biometrics*, 67, 654–656.

Plummer, M. (2008): "Penalized loss functions for bayesian model comparison," *Biostatistics*.

R Core Team (2016): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/.

Tomasko, L., R. W. Helms, and S. M. Snapinn (1999): "A discriminant analysis extension to mixed models," *Statistics in Medicine*, 18, 1249–1260.

Verbeke, G. and E. Lesaffre (1997): "The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data," *Computational Statistics & Data Analysis*, 23, 541–556.

Verbeke, G. and G. Molenberghs (2013): "The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models," *Biostatistics*, 14, 477–490.

Zhang, D. and M. Davidian (2001): "Linear mixed models with flexible distributions of random effects for longitudinal data," *Biometrics*, 57, 795–802.

Table 5: Prediction accuracy of random-effect prediction with differing number of visits and different sample sizes under the assumption that the random effects jointly follow a 2-component multivariate normal distribution with a high degree of departure from normality (top panel) and a t-distribution with 3 degrees of freedom (bottom panel). Prediction accuracies are reported as the mean over all simulated datasets

Scenario	K	Cut	Sens	Spec	PCC	AUC	PPV	NPV
	2-component multivariate normal distribution							
N = 250 4 Visits	1	0.45	0.66	0.74	0.73	0.67	0.46	0.89
	2	0.15	0.46	0.82	0.74	0.62	0.48	0.86
	3	0.07	0.56	0.73	0.70	0.65	0.37	0.87
	4	0.07	0.50	0.70	0.66	0.60	0.33	0.85
N = 250 9 Visits	1	0.24	0.75	0.79	0.78	0.79	0.53	0.92
	2	0.29	0.61	0.82	0.77	0.71	0.50	0.89
	3	0.22	0.60	0.75	0.72	0.67	0.41	0.88
	4	0.19	0.55	0.66	0.64	0.59	0.32	0.85
N = 2,500 4 Visits	1	0.01	0.71	0.87	0.84	0.83	0.61	0.92
	2	0.07	0.69	0.93	0.88	0.84	0.76	0.93
	3	0.03	0.43	0.75	0.69	0.60	0.68	0.85
	4	0.08	0.30	0.86	0.75	0.58	0.55	0.83
	t-distribution with 3 degrees of freedom.							
N = 250 4 Visits	1	0.21	0.75	0.75	0.75	0.80	0.44	0.92
	2	0.30	0.73	0.74	0.74	0.78	0.42	0.92
	3	0.49	0.68	0.69	0.69	0.71	0.37	0.89
	4	0.48	0.63	0.60	0.61	0.59	0.31	0.86
N = 250 9 Visits	1	0.19	0.76	0.76	0.76	0.81	0.44	0.93
	2	0.23	0.75	0.75	0.76	0.81	0.45	0.92
	3	0.48	0.72	0.70	0.70	0.74	0.39	0.91
	4	0.44	0.61	0.64	0.63	0.61	0.32	0.87
N = 2,500 4 Visits	1	0.14	0.77	0.76	0.76	0.83	0.45	0.93
	2	0.17	0.76	0.77	0.76	0.82	0.45	0.93
	3	0.31	0.75	0.75	0.75	0.81	0.43	0.92
	4	0.57	0.70	0.69	0.69	0.73	0.37	0.90