Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Yarkiner et al., J Biomet Biostat 2013, S9
DOI: 10.4172/2155-6180.S9-001

Research Article

Open Access

Applications of Mixed Models for Investigating Progression of Chronic Disease in a Longitudinal Dataset of Patient Records from General Practice

Zalihe Yarkiner[1]*, Gordon Hunter[1], Rosie O'Neil[1] and Simon de Lusignan[2]

[1]School of Mathematics, Faculty of Science, Engineering and Computing, Kingston University, Penrhyn Road, Kingston-Upon Thames, KT1 2EE, United Kingdom

[2]Department of Health Care Management and Policy, Faculty of Business, Economics and Law, University of Surrey, Guildford, GU2 7XH, United Kingdom

Abstract

The field of longitudinal analysis is a rapidly developing and increasingly important area of statistical modelling. This is in response to the increasing availability of longitudinal data across many fields and recognition of the rich resources such data might provide. However, a lag between the development of statistical methodologies and their applications to substantive problems has been identified, but current advances in novel longitudinal methods aim to redress this imbalance. Longitudinal data often presents repeated response measures, but these data are often unbalanced in relation to number of and intervals between measures. Although Linear Mixed Models provide a framework which can accommodate such unsystematic response patterns, such models become unreliable when responses do not approximately follow a normal distribution. Extensions of Linear Mixed Models to Generalized Mixed Models allow the analysis of such non-normal outcomes via appropriate transformations of the response. These models, which are based on a repeated measures structure within a two-level multilevel framework, allow both random and systematic effects to be studied simultaneously. Although these are well-established, they are only recently being applied in the medical and social sciences.

Here, applications of these models are illustrated by analysing the progression of Chronic Kidney Disease (CKD) over time, and in relation to the impact of known co-morbidities. The data are taken from routinely collected patient records from a representative sample of UK General Practices (GPs). The aim is to use the longitudinal aspects of the data to further understanding of the early indications and the nature of the progression of CKD. The methodologies should be applicable to other chronic illnesses, which are primarily managed at the GP level.

The results of our models concur with previous research, in regard to the associations between individual co-morbidities and CKD. Furthermore, our models evaluate the impact of combinations of these co-morbidities on the rate of progression of CKD, as measured by repeated estimated glomerular filtration rate (eGFR) readings. Our results provide evidence that this methodological approach is a useful and appropriate mechanism for investigating dynamic relationships within health-related data, and that such routinely collected data can be useful in epidemiological research.

Keywords: Longitudinal analysis; Generalized linear mixed models; Chronic kidney disease

Introduction

Longitudinal research is described as investigation of longitudinal data, where the term "longitudinal" is used in the context of representing a change of response over a long period of time at the individual level. The longitudinal data under examination in this study is routinely collected medical general practice (GP) data, where each individual has repeated measurements of the response variable over time. This retrospective longitudinal data set is used here to analyse individuals who had experienced the same event (e.g. diagnosis of a particular disease) within the same time interval [1].

This type of data frequently poses some statistical challenges when modelling the response, in order to account for the variance and covariance of the repeated measurements from the same individual. These include the repeated measurements being taken at unequal time intervals, the data structure being unbalanced (i.e. unequal numbers of observations per individual) and response variables, which are often not normally distributed.

The study of change is vital in many disciplines [2], where the main objective is to model the change in the response for an individual over time, and the factors affecting that change [3]. When the response is a continuous variable, traditional techniques used to analyse the change are repeated Analysis of Variance (ANOVA), Multivariate Analysis of Variance (MANOVA) and linear mixed models (LMM). All of these traditional models assume that the response variable is normally distributed.

In this study, we use a longitudinal sample of GP records from the UK to investigate the impact of various common diseases on the progression of Chronic Kidney Disease (CKD). Chronic Kidney Disease is defined as the "gradual and usually permanent loss of kidney function over time" [4]. It is a multi-stage disease, with stage 1 being the mildest, through to stage 5, renal failure (also known as end stage

***Corresponding author:** Zalihe Yarkiner, School of Mathematics, Faculty of Science, Engineering and Computing, Kingston University, Penrhyn Road, Kingston-Upon Thames, KT1 2EE, London, United Kingdom, Tel: +44 (0) 7842 772 172; E-mail: z.yarkiner@kingston.ac.uk

Received June 11, 2013; **Accepted** September 28, 2013; **Published** September 30, 2013

Citation: Yarkiner Z, Hunter G, O'Neil R, de Lusignan S (2013) Applications of Mixed Models for Investigating Progression of Chronic Disease in a Longitudinal Dataset of Patient Records from General Practice. J Biomet Biostat S9: 001. doi:10.4172/2155-6180.S9-001

renal disease, ESRD). Kidney function is normally measured using Glomerular Filtration Rate (GFR), which is a measure of the kidney's efficiency. Since the disease lacks symptoms in the early stages, patients are commonly not diagnosed, until they have reached at least stage 3, when their GFR falls below 60 mL/min/1.73 m^2 for at least 3 months, with or without kidney damage. From study of the relevant literature, the main co-morbidities that are established as being related to CKD, and hence used in this study, are diabetes, anaemia and cardiovascular diseases (including hypertension, ischaemic heart disease and peripheral vascular disease).

The responses being investigated in this study are repeated estimated GFR (eGFR) values, computed using the Modification of Diet in Renal Disease (MDRD) formula [5,6]. Since a portion of the original dataset is removed in order to analyse only the patients who are diagnosed to have CKD between stages 3 to 5 (i.e. those with eGFR values less than 60), the distribution of the sample data presented here is negatively-skewed, so the distribution of the response is not assumed to be normally distributed.

The main aim of this paper is to evaluate how eGFR changes over time, and particularly how previously established associated factors and co-morbidities affect the progression of the disease. A list of acronyms and abbreviations used is given in Table 1.

Materials

Data

The data are taken from routinely-collected patient records from a sample of 129 general practices in England and Wales, these practices provide population based primary health care. Initially, the data contained individual records for approximately 1.1 million patients collected over an 11 years period (2000-2010). The data include information about diagnosed diseases, prescribed medications and therapies, results of relevant laboratory tests, as well as other more general health measurements, for example blood tests, blood pressure values and body mass index values. Basic demographic data, e.g. age, gender, ethnicity, etc., and information about some lifestyle factors is also available at the patient level, although collection of this information is not compulsory, and so is not consistently recorded. As several different computerized data recording systems are in use across general practices in the UK, the data for our study was extracted using a Department of Health approved data extract tool, Morbidity Information and Export Syntax (MIQUEST) [7]. Using MIQUEST, records from different practices and systems were extracted then combined into a single database, from which a "flat" file was created for this analysis. Further manipulation of the data to get it into a useable format and subsequent analyses are carried out using SPSS version 21. The study was approved by the ethics committees of Kingston University and St. George's, University of London, and part of a wider ethically approved project [8,9].

Abbreviation	Meaning	Reference
AIC	Akaiki Information Criterion	[25,27]
ANOVA	Analysis of Variance	[28]
BIC	Bayesian Information Criterion	[25,27]
CKD	Chronic Kidney Disease	[29]
CVD	Cardiovascular Disease	[30]
eGFR	Estimated Glomerular Filtration Rate	[31]
ESRD	End Stage Renal Disease	[10]
GFR	Glomerular Filtration Rate	[31]
GLM	General Linear Model	[32]
GLMM	Generalized Linear Mixed Model	[17,33]
GP	General Practice	[34]
IHD	Ischaemic Heart Disease	[18]
LMM	Linear Mixed Model	[20,35]
MANOVA	Multivariate Analysis of Variance	[28,36]
MDRD	Modification of Diet in Renal Disease	[5,6]
MIQUEST	Morbidity Information and Export Syntax	[7]
NHS	National Health Service of the UK	[7]
PVD	Peripheral Vascular Disease	[37]
SEC	Science, Engineering and Computing	
-2LL	-2 LogLikelihood	[38]

Table 1: List of abbreviations.

Data validation and measures

In order to validate the GP dataset, we compared its basic demographic composition (in terms of age and gender) to that of the UK Census of 2011. We found that the two populations are similar in these respects and hence deduced that the GP data provide a good representation of the population of England and Wales. However, because CKD is primarily a disease of adulthood and the MDRD equation is only considered to be valid for people aged between 18 and 75 [5], records of patients outside that age range at baseline (i.e. t=0) are removed from this study. Furthermore, it has also been suggested that the (MDRD) equation is not valid for obese people [4], and so patients who had BMI greater than 30, and hence were obese, at the time of their baseline measurement were excluded from the sample. Further data cleaning techniques were employed to maximise the quality and integrity of the dataset before analysis, e.g. removal of incomplete and incorrect patient records.

The response (dependent) variable in the following analyses is eGFR. This biomarker measure is calculated using the modification of diet in renal disease (MDRD) formula [10]. The data had been adjusted for differences in laboratory assay, prior to standardisation in 2006 [11]. The baseline measure for each patient is defined as the first eGFR measurement which resulted in a diagnosis of CKD between stages 3 and 5. This was set as time t=0, i.e. the baseline time, for each patient.

The data provide repeated measures of the eGFR for individual patients, with between 1 and 15 eGFR readings per patient. As change in eGFR over time is one of the main themes of this study, it was felt that a clearer pattern of the decline in eGFR, and hence the progression of CKD, over time would be best investigated by considering only those patients who had at least 8 repeated eGFR measurements, and where successive measurements were at least 3 months apart. This requirement on the time gap between measurements was imposed, in order to avoid small errors or uncertainties in eGFR values taken in quick succession leading to large errors in the corresponding rates of change. This action results in all patients included in this study having observations recorded over at least four years.

Diagnoses of specific previously proposed co-morbidities of CKD, namely diabetes, anaemia and cardiovascular diseases-including peripheral vascular disease (PVD), ischaemic heart disease (IHD) and hypertension- at or before t=0 (baseline time) are also noted, i.e. a binary coded indicator is included denoting whether or not the co-morbidity is present in that patient at their time of first CKD diagnosis. These disease indicators were binary coded using the SPSS convention, such that disease absent was coded as 1 and disease present was coded as 0. Additionally, we recorded whether patients had diabetes or were anaemic; the latter is a common and important complication of CKD

[12]. Our definition of a patient having cardiovascular disease was that he/she had been diagnosed to have either family history of CVD, having hypertension, PVD or IHD by the time of their diagnosis with CKD. Age is calculated as the difference in years between the patient's year of birth and the year of the baseline (t=0) eGFR measurement.

Methods

Linear Mixed Models (LMMs)

These are the basic type of models used to analyse Gaussian (normal) longitudinal data, where the response is modelled as a linear function, and is assumed to be normally distributed with constant variance. The model is composed of two levels, where two main types of effects are investigated, namely fixed effects and random effects [13]. Level one represents the repeated eGFR values and level two represents the subjects. Fixed effects account for between-subject variation, are taken into account at level two and model the mean structure. In contrast, the random effects take account of within-subject variation, are considered at level one and model the different types of variations, such as serial correlation, measurement error and random effects [14]. Serial correlation is the variation due to the association between the variable and itself over various time lags. Measurement error is the variation due to an inaccuracy occurring during the measurement of the response variable. A random effect is the variation due to random factors that cannot be measured or controlled.

A LMM is of the form;

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i \quad b_i \sim N_q\left(0,\psi\right) \quad \varepsilon_i \sim N_{ni}\left(0,\sigma^2\Lambda_i\right) \qquad (1)$$

where

Y_i is the response vector, which is the sequence of repeated eGFR measurements in n_i×1 dimensions, where n_i is the total number of observations, for individual *i*.

X_i is the model matrix for the fixed effects, which is in n_i×p dimensions, where *p* is the total number of fixed effects,

β is the vector for fixed-effects coefficients in *p*×1 dimensions,

Z_i is the model matrix for the random effects, which has n_i×q dimensions, where *q* is the total number of random effects,

b_i is the vector for random-effects coefficients in *q*×1 dimensions,

ε_i is the vector of errors in n_i×1 dimension,

ψ is the covariance matrix for random errors in *q*×q dimensions and

$\sigma^2\Lambda_i$ is the covariance matrix for errors in $n_i \times n_i$ dimensions [15].

Initially, the response is assumed to be normally distributed and the best LMM model containing the most significant co-morbidities is found. However, the histogram of the response variable shown in Figure 1 illustrates that the eGFR values are not normally distributed for our data, and tests for normality (Table 2) confirmed this. It can be seen that the distribution is slightly negatively skewed. Therefore, the normality assumption is no longer valid for the analysis, and so alternative models for non-normal data are considered.

Generalized Linear Mixed Models (GLMM)

Since the distribution of our data is skewed, the normality assumption that is assumed in the LMM is violated. Hence, a function of the mean response is modelled instead of the mean response itself.

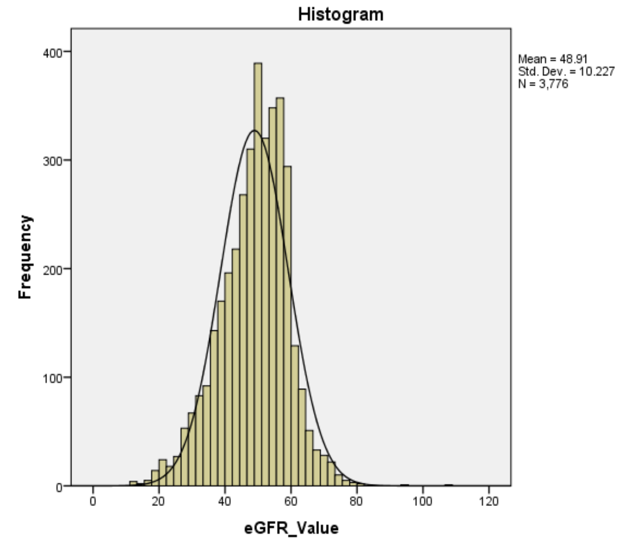

Figure 1: Distribution of eGFR values across all measurements for all patients within our dataset who have been diagnosed with CKD between stages 3 to 5. The solid line is the best-fitting normal distribution.

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistics	**df**	**Sig.**	Statistics	df	**Sig.**
(a) eGFR Data (Figure 1)						
eGFR Value	0.054	3776	<0.001	0.985	3776	<0.001
(b) ln(eGFR) Data (Figure 2)						
ln(eGFR) Value	0.101	3776	<0.001	0.917	3776	<0.001

In this sample of the data set, since the total number of observations number is greater than 2000, the Kolmogorov-Smirnov test is used to test the normality assumption of the dependent variable. From the Table 1 above, it can be concluded that the statistics resulted from Kolmogorov-Smirnov test is significant, meaning that H0 from the hypothesis is rejected and the sample data is assumed to be statistically different from a normal population. The results of the Shapiro-Wilk test also agreed with this. Therefore GLMM models are performed instead of LMM models in order to model the dependent variable which does not follow a normal distribution. However, the dependent variable should follow one of the known distributions from the exponential family.

Even if the dependent variable (which is eGFR) is transformed to the log domain, the distribution is still not normally distributed as can be seen from Figure 2 and Table 1b. Therefore, eGFR itself is used in the model formulation assuming a gamma distribution. This assumption is made in formulation of GLMM models 3-5.

Table 2: Normality test results.
(a) for eGFR data (Figure 1) and
(b) for ln(eGFR) data (Figure 2).

GLMMs are an extension of general linear models (GLM), which take random and fixed effects into account, and are used when the assumption of independence between observations is violated (e.g. in longitudinal studies where repeated measurements are taken from the same individual) [16,17]. GLMM models are the extension of LMMs to account for the response following various non-normal, but standard distributions such as the gamma distribution, inverse Gaussian distribution or binomial distribution. GLMMs allow the linear predictor to have, in addition to fixed effects, one or more random components with assumed normal distribution of mean zero and constant variance. In this way, the correlation between observations from the same individual is taken into account in these models [15].

The general form of the GLMM is the same as in equation (1). However, in a GLMM, instead of modelling the response itself, a link function is used to transform the response into a linear predictor. The covariance structure is also estimated using the link function. The link function is represented by a generic link that is denoted by g (.) and the linear predictor is formed from the combination of fixed and random effects, excluding the residuals.

A linear predictor has the form;

$$\eta = X\beta + Z\gamma \tag{2}$$

Where

X is the matrix of regressors (e.g. independent variables) with corresponding βs, which are the fixed effect parameter coefficient estimates for these regressors,

Z is the matrix of variables having random effects with corresponding random effects denoted by γ [19].

An inverse link function, denoted by h (.)=g-1(.), is used to convert the transformed response back to the original units. In non-Gaussian data, the assumption of correlation between individual measurements is different from that for Gaussian data, and hence results in different interpretations of the regression coefficients in the model [15].

In GLMMs, the effect of a covariate on the mean response for that individual is estimated conditionally based on the random effect for that individual [20]. The type of the link function and corresponding family of the distribution is chosen based on whether the outcome is binary, discrete or continuous [21,22]. In this study, since the response (i.e. eGFR values) are on a continuous scale, only one particular type of link function, namely the logistic link function with a gamma distribution, is used.

The logistic link function [23] is expressed as;

$$g(p) = \ln\left(\frac{p}{1-p}\right) \tag{3}$$

The inverse link function [24] is then expressed as;

$$h(s) = \frac{e^s}{1+e^s} \tag{4}$$

The probability distribution function for the general gamma distribution [19] is expressed as;

$$f(x,a,b) = \frac{1}{b^a\Gamma(a)} x^{a-1} e^{-\frac{x}{b}} \quad a > 0,\ b > 0,\ 0 < x < \infty \tag{5}$$

Where "*a*" is the shape parameter and "*b*" is the scale parameter of the gamma distribution, and Γ(α) is the gamma function, satisfying equation 6 :

$$E(X)=ba$$

$$Var(X)=b^2a \tag{6}$$

The generalized gamma distribution used in this paper represents a general family of distributions, where the exponential distribution and chi-square distribution are special cases with a=1 and with b=2, respectively. The generalized gamma distribution is used to model the product of exponentially distributed random variables. In both linear mixed models and generalized linear mixed models, the coefficients were computed using the restricted maximum likelihood estimation method in the SPSS package.

Results and Discussion

The main purpose of this study is to determine if there is any association between the co-morbidities of interest (i.e. diagnoses of anaemia, diabetes and cardiovascular diseases), and the response (i.e. eGFR values) and its change over time. In this study, the random effect is a correction appropriate for a particular individual patient. Initially, a histogram is drawn in order to look at the distribution of eGFR values across all the measurements for all patients in our sample. In order to make a goodness of fit comparison, a normal curve of appropriate mean and variance is then drawn on top of the same graph (Figure 1). It can be concluded from this that the distribution of this data, for CKD patients only, is negatively skewed.

In total, five models are created and analysed for our data set. Parameter estimates and corresponding standard errors and p-values for each of the five models can be found in Tables 3-7, a table for each model. In these models, only the statistically significant co-morbidities were retained and the rest were removed from the models and the optimal coefficient values are recalculated. The main reason for performing

Model 1 Model Term	Coefficient	Standard Error	p-value
Time (years)	-0.213	0.085	0.013**
Diagnosis of CVD	2.066	0.777	<0.001***
Diagnosis of Diabetes	3.016	0.724	0.008**
Diagnosis of Anaemia*Time	-0.567	-0.567	<0.001***

*p<0.05, **p<0.01, ***p<0.001

Table 3: Results of Model 1.

Model 2 Model Term	Coefficient	Standard Error	p-value
Intercept	3.877	0.016	<0.001***
Time (years)	-0.007	0.002	<0.001***
Diagnosis of CVD	0.043	0.016	0.007**
Diagnosis of Diabetes	0.060	0.015	<0.001***
Diagnosis of Anaemia*Time	-0.013	0.004	0.003**
Diagnosis of CVD*Time	-0.006	0.003	0.021*

*p<0.05, **p<0.01, ***p<0.001

Table 4: Results of Model 2.

Model 3 Model Term	Coefficient	Standard Error	p-value
Intercept	3.869	0.017	<0.001***
Time (years)	-0.006	0.002	0.003**
Diagnosis of CVD	0.050	0.050	0.004**
Diagnosis of Diabetes	0.064	0.064	<0.001***
Diagnosis of Anaemia*Time	-0.015	0.005	0.004**
Diagnosis of CVD*Time	-0.007	0.003	0.014*

*p<0.05, **p<0.01, ***p<0.001

Table 5: Results of Model 3.

Model 4 Model Term	Coefficient	Standard Error	p-value
Intercept	4.060	0.012	<0.001***
Time (years)	0.003	0.001	0.05*
Diagnosis of CVD	-0.033	0.013	0.014*
Diagnosis of Diabetes	-0.052	0.012	<0.001***
Diagnosis of Anaemia*Time	0.009	0.003	0.001***
Diagnosis of CVD*Time	0.005	0.002	0.002**

*p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001

Table 6: Results of Model 4.

Model 5 Model Term	Coefficient	Standard Error	p-value
Intercept	4.060	0.012	<0.001***
Time (years)	0.003	0.001	0.05**
Diagnosis of CVD	-0.033	0.013	0.013*
Diagnosis of Diabetes	-0.050	0.012	<0.001***
Diagnosis of Anaemia*Time	0.003	0.002	0.003**
Diagnosis of CVD*Time	0.002	0.003	0.005**

*p ≤ 0.05, **p<0.01, ***p<0.001

Table 7: Results of Model 5.

Model	Type	AIC	BIC	-2LL
1	LMM	24553.059	24621.572	24530.989
2	GLMM	-4489.937	-4421.424	-4512.007
3	GLMM	-4450.508	-4381.995	-4472.578
4	GLMM	-5724.751	-5656.239	-5746.822
5	GLMM	-5733.991	-5671.701	-5754.049

Table 8: Model comparison.

five models in total is to find the "best" model with fewest parameters, best fit to the data and with the least complex covariance structure. As each of these models are computed, the "goodness of fit" of each one to our data is found using the Akaiki Information Criterion (AIC), -2 LogLikelihood (-2LL) and Bayesian Information Criterion (BIC), in order to compare the models and ensure that a better model is obtained at each step (Table 8). Each of these criteria is such that the lower the value of the statistic, the better the model fits the data [25]. In order to keep consistency between models and to make fair comparisons between them, the same co-morbidities are initially included in all of the models studied. In all five models, the same sample of data is used as for the LMM (i.e. equation (1)), using 472 patients with at least 8 repeated observations for each patient, resulting in 3776 observations of eGFR values in total.

Initially, model 1 is produced assuming a normal distribution with identity link function. This model is essentially that described in equation (1), using the LMM approach. In this model, the co-morbidities found to be significant, and hence taken into account are diagnoses of diabetes and cardiovascular diseases at baseline and time, and its interaction with the diagnoses of anaemia and of cardiovascular disease. The coefficient values, their standard errors and significance levels are given in Table 3.

The coefficients which best fit our CKD data are evaluated and result in the equation for model 1 which is

$$y = 48.592 - 0.213(time) + 3.016(Diabetes\ diagnosis) + 2.066(CVD\ diagnosis)$$
$$-0.567(Anaemia\ diagnosis * time) - 0.328(CVD\ diagnosis * time) \qquad (7)$$

Where y represents the eGFR value and each diagnosis is 1, if the disease is present or 0 otherwise.

In order to investigate whether a multiplicative rather than additive model would be more appropriate for this data, the eGFR values are transformed to the natural logarithmic domain, and, by again assuming normality of the response variable, model 2 is computed using a normal distribution with log link function. The coefficients found, together with their standard errors and significance levels, are given in Table 4.

The equation for model 2, with optimal coefficient values is found to be

$$\ln(eGFR) = 3.869 - 0.006"time" + 0.064(Diabetes\ diagnosis) + 0.050(CVD\ diagnosis)$$
$$-0.015(Anaemia\ diagnosis * time) - 0.007(CVD\ diagnosis * time) \qquad (8)$$

Since model 2 is in the log domain, very different coefficients are observed from before. When the AIC, BIC and -2LL information criteria for models 1 and 2 are compared, it can be observed that transforming the eGFR values using the natural logarithm improved the model fit by a large amount, even though normality assumption was still retained (Table 8).

We had evidence (Figures 2 and 3) that both our eGFR and ln

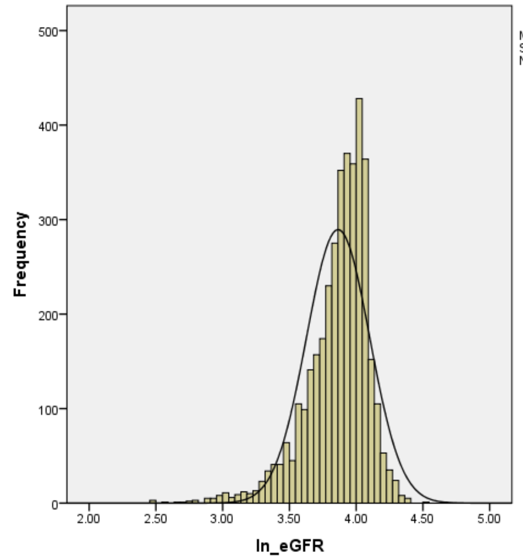

Figure 2: Distribution of ln(eGFR) values across all measurements for all patients within our dataset who have been diagnosed with CKD between stages 3 to 5. The solid line is the best-fitting normal distribution.
If the dependent variable (i.e. eGFR) is subtracted from a constant value (i.e. the highest eGFR value found in our dataset), then it can be assumed that this transformed dependent variable (107-eGFR) will follow a standard gamma distribution. As can be seen from figure 3, even the distribution is still not normal, distribution in figure 3 is closer to normal compared to figure 2 and, hence, the assumption of gamma distribution when using the transformed eGFR value rather than the eGFR value itself is better, which is used in GLMM model 5.

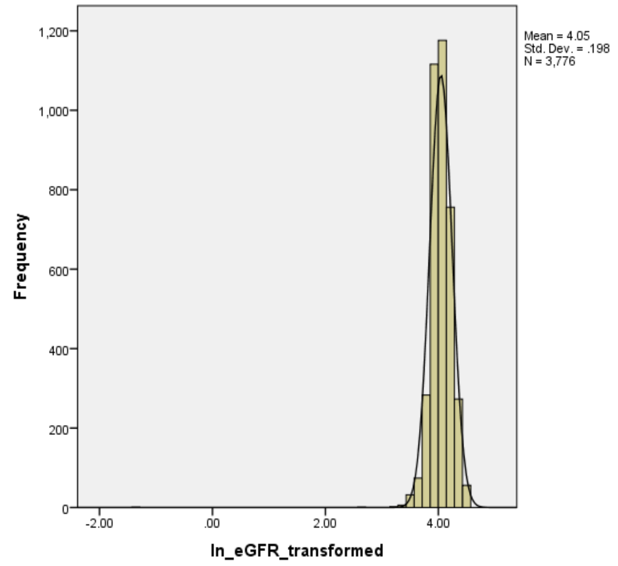

Figure 3: Distribution of ln(Transformed eGFR), i.e. ln(107-eGFR), values across all measurements for all patients within our dataset who have been diagnosed with CKD between stages 3 to 5. The solid line is the best-fitting normal distribution.

(eGFR) data were skewed and not exactly normally distributed. We, therefore, carried out normality tests (Table 2), which showed the data to be non-normal. Thus, in model 3, the normality assumption is removed and, since the distribution of eGFR values is found to be skewed, the eGFR values are therefore modelled using a gamma distribution with log link function. In this way, the natural logarithm of the mean eGFR values is modelled, the coefficients were re-calculated and the model coefficients, their standard errors and significance values are given in Table 5. The equation with best coefficient values for this model (model 3) is found to be

$$\begin{aligned}\ln(107-eGFR) = 3.869 - 0.006\ (\text{time})\ +0.064\ (\text{Diabetes diagnosis})\ +\ 0.050\ (\text{CVD diagnosis})\\ -\ 0.015(\text{Anaemia diagnosis})\ \text{x}\ (\text{time})\ - 0.007\ (\text{CVD diagnosis})\ \text{x}\ (\text{time})\end{aligned} \quad (9)$$

A gamma distribution is usually employed when the data is positively skewed. However, in this study, the data is negatively skewed (Figure 1). Therefore, when the differences in the information criteria between model 3 and model 2 are compared against the corresponding difference between model 2 and model 1, only small improvements are observed in the former case. In order to get a further improved model, the eGFR values are first manipulated to reverse the shape of the distribution from negatively-skewed to positively-skewed. This transformation of eGFR values is carried out by subtracting each eGFR value from the whole number, just greater than maximum eGFR value found amongst our CKD patients (i.e. 107). This ensures any potential problems due to having to find the logarithm of a negative-valued quantity are removed. In this way, the distribution is changed to positively-skewed, and hence will be more appropriate for being modelled using a gamma distribution in the analysis. Therefore, model 4 is formed with the manipulated eGFR values as response variable, using a gamma distribution with log link function. The equation with optimal coefficients for this model is found to be;

$$\begin{aligned}\ln(107-eGFR) = 4.060 + 0.003\text{"}time\text{"} - 0.052(Diabetes\ diagnosis)\\ -0.033(CVD\ diagnosis) + 0.009(Anaemia\ diagnosis * time)\\ +0.005(CVD\ diagnosis * time)\end{aligned} \quad (10)$$

When the information criteria for model 4 and model 3 are compared, a major improvement is observed in the all three measured goodness of fit criteria, indicating that model 4 is a much better model for this data.

The association between the initial eGFR status and the progression of eGFR over time is estimated by calculating the covariance matrix. The "unknown structure" of the covariance matrix is estimated by the SPSS package. In each of the models above (models 1 to 4), the covariance matrix is evaluated with the "unstructured" covariance option selected, and the package then estimated the covariance. However, from each of these models, the covariance between intercept and slope is estimated to be zero. Therefore, a simpler covariance matrix structure, such as a variance component (diagonal) matrix can possibly be used to achieve a better model with lower computational requirements. In model 5, the process of model 4 is repeated, but with the "variance component" option selected for the form of the covariance matrix rather than "unstructured" for the calculations. In this way, it can be seen that a better fit to the data (in terms of information criteria) can be achieved by using the simpler covariance matrix (Table 8). The coefficients, their standard errors and significance levels for this simpler model (model 5), are given in Table 7. The equation for model 5 is given by;

$$\begin{aligned}\ln(107-eGFR) = 4.060 + 0.003\text{"}time\text{"} - 0.050(Diabetes\ diagnosis)\\ -0.033(CVD\ diagnosis) + 0.008(Anaemia\ diagnosis * time)\\ +0.005(CVD\ diagnosis * time)\end{aligned} \quad (11)$$

When comparing all five models, the lowest AIC, BIC and -2 LL values are found for model 5, and we conclude that this is the best-fitting model for our data. The results from all five models indicated that statistically significant parameters are diagnoses of CVD and of diabetes to account for the changes in initial value of eGFR (i.e. the intercept) across patients, whereas the parameters included to describe the progression of CKD (i.e. the slope), and the effect of co-morbidities on this are the interaction with time terms of the diagnoses of anaemia and of CVD in all cases.

For evaluation and interpretation in terms of eGFR values, values obtained using model 5 are transformed back by exponentiation (i.e. the inverse of taking the natural logarithm), and then subtracting the result from 107 to give values to give meaningful model-predicted eGFR values. The mean eGFR value at time zero is found from model 5 to be 49.0257, given that the patient has not being diagnosed to have CVD or diabetes. If the patient has been diagnosed to have only CVD at time zero, this eGFR value rises to 50.9076, whereas if the patient has been diagnosed to have only diabetes at time zero, the resulting eGFR value is 51.8531. This means that both diagnoses of CVD and of diabetes tend to increase the initial eGFR value, with the effect of having diabetes being more influential than that of diagnosis of CVD at baseline. However, each year increase in time results in a decrease in this predicted eGFR value by a factor of 0.9964, if the patient has none of these co-morbidities. Thus the patient has not being diagnosed to have CVD or diabetes at time zero, then this initial eGFR value (i.e. 49.0257) would be expected to decrease to 48.8515 after one year. This decrease will be more if the patient has either anaemia, CVD or both (Figure 8).

Each regression coefficient is estimated by using a robust method, hence resulting in the corresponding standard errors being low. The regression coefficients for the parameters affecting the progression of CKD are lower standard errors (less than 0.005) than those for the regression coefficients for the parameters affecting the initial eGFR value (between 0.010 and 0.015).

The higher eGFR values observed for patients with CVD indicates that for some initial period, they will have better (i.e. higher) eGFR than patients without that disease. However, the fast rate of eGFR decrease for patients with CVD results in lower eGFR values than non-CVD patients after some time, typically around 4.125 years. The predicted eGFR values obtained from each model for "typical" patients with each disease combination are shown in Figures 4-8.

Conclusion

The most efficient ways of analysing longitudinal data with repeated measurements when the data is incomplete and unbalanced are by using the methodologies known as Linear Mixed Models (LMMs) and Generalized Linear Mixed Models (GLMMs). LMMs are used when the outcome measure can be assumed to follow a normal distribution, whereas GLMMs are applied otherwise, when this normality assumption is removed. However, some standard distribution should be assumed in order to perform the GLMM approach, and here a gamma distribution is used since the distribution of the outcome is skewed. Furthermore, a natural logarithm link function is used to transform the response. Here, the result obtained from the GLMM approach used in model 5 indicates that when a patient has been diagnosed to have CVD or diabetes, that patient will have a higher initial eGFR value compared with a patient without those diseases. However, having either CVD or anaemia will increase the rate of decline of eGFR, and hence the progression of CKD. The models could be improved if a distribution that better fits the data is used, instead of assuming a gamma distribution.

The results of this study are consistent with those of previous research on the progression of CKD [26]. However, our work is based on a large sample of routinely-collected General Practice patient records, in contrast to the cross-section controlled studies or clinical trials. Our results provide evidence that the methodological approach presented here applied to this routinely collected data is a useful and appropriate mechanism for investigating dynamic relationships within health-related data.

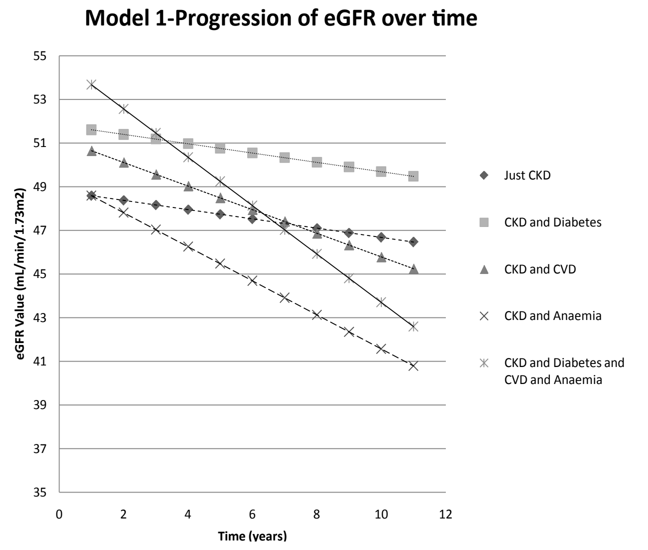

Figure 4: Graphs of eGFR against time for typical patients with particular disease combinations, as predicted by Model 1.

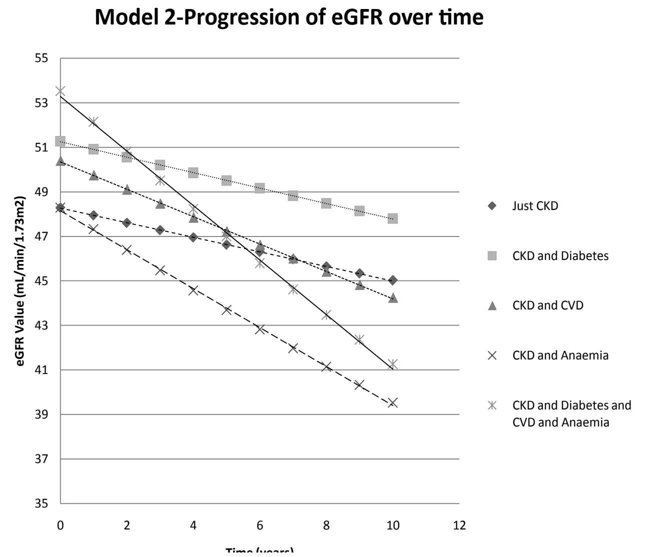

Figure 5: Graphs of eGFR against time for typical patients with particular disease combinations, as predicted by Model 2.

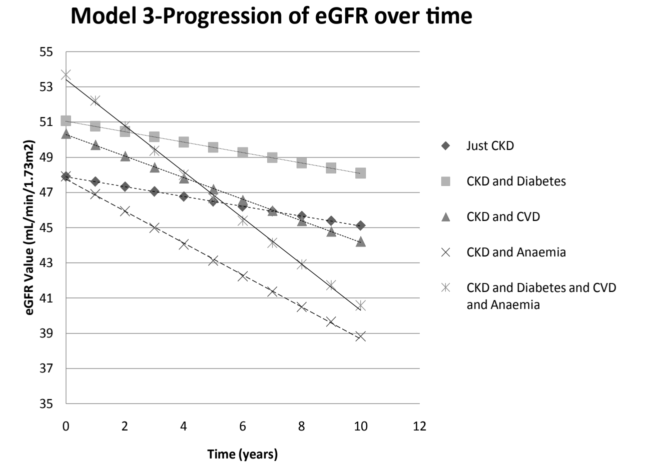

Figure 6: Graphs of eGFR against time for typical patients with particular disease combinations, as predicted by Model 3.

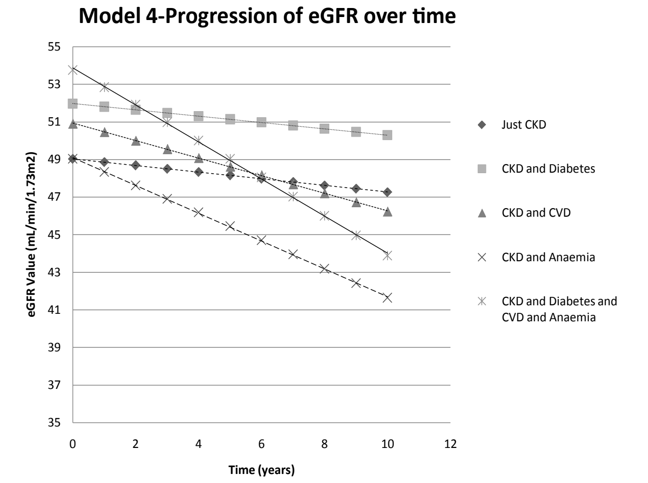

Figure 7: Graphs of eGFR against time for typical patients with particular disease combinations, as predicted by Model 4.

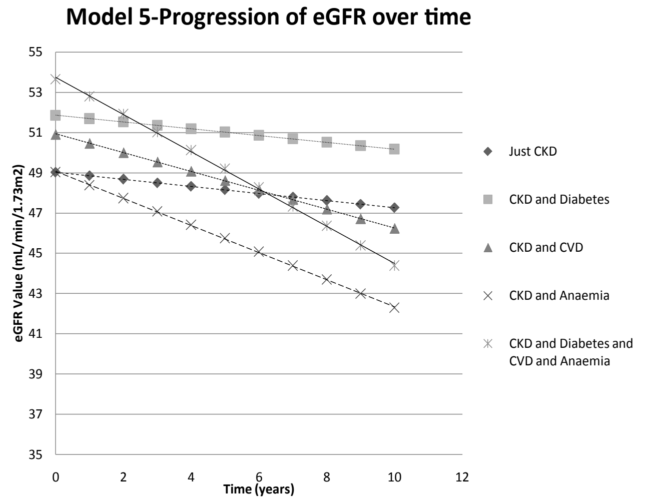

Figure 8: Graphs of eGFR against time for typical patients with particular disease combinations, as predicted by Model 5.

Acknowledgement

We are grateful to Kingston University, Faculty of SEC, for financial support of this research and the Health Foundation for providing the data. We would also like to thank Dr. Penelope Bidgood and Prof. Simon Jones for helpful comments on the work described here.

References

1. Ryder NB (1965) The cohort as a concept in the study of social change. Am Soc Rev 30: 301-328.
2. Fitzmaurice GM, Molenberghs G (2006) Advances in longitudinal data analysis: An historical perspective. Chapman & Hall/CRC Press, Boca Raton, USA.
3. Fitzmaurice GM, Laird NM, Ware JH (2004) Applied longitudinal analysis. Wiley & Sons, Hoboken, New Jersey, USA.
4. National Kidney Foundation (2002) K/DOQI clinical practice guidelines for chronic kidney disease: Evaluation, classification and stratification. Am J Kidney Dis 39: S1-S266.
5. Jones G (2005) Routine reporting of eGFR: Laboratory implementation guidelines. Med J Aust 183: 138-141.
6. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, et al. (1999) A more accurate method to estimate glomerular filtration rate from serum creatinine: A new prediction equation. Ann Intern Med 130: 461-470.
7. http://www.connectingforhealth.nhs.uk/systemsandservices/ssd/prodserv/vaprodmiquest
8. de Lusignan S, Gallagher H, Chan T, Thomas N, van Vlymen J, et al. (2009) The QICKD study protocol: a cluster randomised trial to compare quality improvement interventions to lower systolic BP in chronic kidney disease (CKD) in primary care. Implement Sci 4: 39.
9. de Lusignana S, Gallagher H, Jones S, Chan T, van Vlymen J, et al. (2013) Audit-based education lowers systolic blood pressure in chronic kidney disease: The Quality Improvement in CKD (QICKD) trial results. Kidney Int 84: 609-620.
10. Ekart R, Ferjuc A, Furman B, Gerjevic S, Bevc S, et al. (2013) Chronic kidney disease progression to end stage renal disease: A single center experience of the role of the underlying kidney disease. Therapeutic Apheresis and Dialysis 4: 363-367.
11. de Lusignan S, Tomson C, Harris K, van Vlymen J, Gallagher H (2011) Creatinine fluctuation has a greater effect than the formula to estimate glomerular filtration rate on the prevalence of chronic kidney disease. Nephron Clin Pract 117: c213-24.
12. Dmitrieva O, de Lusignan S, Macdougall IC, Gallagher H, Tomson C, et al. (2013) Association of anaemia in primary care patients with chronic kidney disease: Cross sectional study of quality improvement in chronic kidney disease (QICKD) trial data. BMC Nephrol 14: 24.
13. Brow H, Prescott R (2006) Applied mixed models in medicine. John Wiley & Sons, Chichester, UK.
14. Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38: 963-974.
15. Diggle PJ, Heagerty P, Liang K, Zeger S (2002) Analysis of longitudinal data. (2nd Edn), Oxford University Press, New York, USA.
16. Berslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88: 9-25.
17. Zeger SL, Karim MR (1991) Generalized linear models with random effects: A Gibbs sampling approach. J Am Stat Assoc 86: 79-86.
18. Furukawa M, Io H, Tanimoto M, Hagiwara S, Horikoshi S, et al. (2011) Predictive factors associated with the period of time before initiation of hemodialysis in CKD stages 4 and 5. Nephron-Clin Pract 117: c341-c347.
19. Hedeker D (2005) Generalized linear mixed models. Encyclopedia of statistics in behavioral science. John Wiley & Sons, USA.
20. McCulloch C, Searle S, Neuhaus J (2008) Generalized, linear, and mixed models. (2nd Edn), John Wiley & Sons, Inc., Hoboken, NJ, USA.
21. Molenberghs G, Verbeke G (2005) Models for discrete longitudinal data. Springer-Verlag, New York, USA.
22. Zhang P, Song PXK, Qu A, Greene T (2008) Efficient estimation for patient-specific rates of disease progression using nonnormal linear mixed models. Biometrics 64: 29-38.
23. Azuero A, Pisu M, McNees P, Burkhardt J, Benz R, et al. (2010) An Application of longitudinal analysis with skewed outcomes. Nurs Res 59: 301-307.
24. Anderson CJ, Verkuilen J, Johnson T (2010) Applied generalized linear mixed models: Continuous and discrete data, for social and behavioural sciences. Springer, USA.
25. LV J, Liu JS (2013) Model selection principles in misspecified models. J Royal Stat Soc Ser B: Stat Methodol.
26. de Lusignan S, Chan T, Gallagher H, van Vlymen J, Thomas N, et al. (2009) Chronic kidney disease management in southeast England: A preliminary crosssectional report from the QICKD – Quality Improvement in Chronic Kidney Disease study. Prim Care Cardiovasc J.
27. Bierens HJ (2006) Information criteria and model selection. Manuscript, Penn State University, USA.
28. Ruth G, Thomas M (1993) Anova for unbalanced data: An Overview. Ecol Soc Am 74: 1638-1645.
29. Huda MN, Alam KS, Harun-Ur-Rashid (2012) Prevalence of chronic kidney disease and its association with risk factors in disadvantageous population. Int J Nephrol.
30. Keser N, Cihan N, Dogu O, Gunduz H, Akdemir R, et al. (2012) Any difference in sociodemograpic variables and risk factors of patients hospitalised with cardiovascular disease (CVD)? Health Med 7: 2325-2331.
31. Rule AD, Glassock RJ (2013) GFR estimating equations: Getting closer to the truth? Clin J Am Soc Nephrol 8: 1414-1420.
32. Olive DJ (2013) Plots for generalized additive models. Commun Stat Theory Methods 18: 3310-3328.
33. Wyatt K, Henley W, Anderson L, Anderson R, Nikolaou V, et al. (2012) The effectiveness and cost-effectiveness of enzyme and substrate replacement therapies: A longitudinal cohort study of people with lysosomal storage disorders. Health Technology Assessment 39: 1-543.
34. De Wet C, Johnson P, O'Donnell C, Bowie P (2013) Can we quantify harm in general practice records? An assessment of precision and power using computer simulation. BMC Med Res Methodol 1: 39.
35. Peralta CA, Vittinghoff E, Bansal N, Jacobs Jr D, Muntner P, et al. (2013) Trajectories of kidney function decline in young black and white adults with preserved GFR: Results from the coronary artery risk development in young adults (CARDIA) study. Am J Kidney Dis 2: 261-266.
36. Daniel K, Cason CL, Shrestha S (2011) A comparison of glomerular filtration rate estimating equation performance in an older adult population sample. Nephrol Nurs J 38: 351-356.
37. Stack AG (2005) Coronary artery disease and peripheral vascular disease in chronic kidney disease: An epidemiological perspective. Cardiol Clin 23: 285-298.
38. Jones RH (2011) Bayesian information criterion for longitudinal and clustered data. Stat Med 30: 3050-3056.

This article was originally published in a special issue, **Longitudinal Data Analysis** handled by Editor. Dr. Mohamed Alosh, Center of Drug Evaluation and ResearchU.S. Food and Drug Administration/FDA, USA