Real-time monitoring of progression towards renal failure in primary care patients

PETER J. DIGGLE
CHICAS, Medical School, Lancaster University, Lancaster, LA1 4YG, UK and Institute of Infection and Global Health, University of Liverpool, Liverpool, L69 7BE, UK

INÊS SOUSA
Department of Mathematics and Applications, University of Minho, Guimarães, 4800-058, Portugal

ÖZGÜR ASAR*
CHICAS, Medical School, Lancaster University, Lancaster, LA1 4YG, UK
o.asar@lancaster.ac.uk ozgurasar@yahoo.com

SUMMARY

Chronic renal failure is a progressive condition that, typically, is asymptomatic for many years. Early detection of incipient kidney failure enables ameliorative treatment that can slow the rate of progression to end-stage renal failure, at which point expensive and invasive renal replacement therapy (dialysis or transplantation) is required. We use routinely collected clinical data from a large sample of primary care patients to develop a system for real-time monitoring of the progression of undiagnosed incipient renal failure. Progression is characterized as the rate of change in a person’s kidney function as measured by the estimated glomerular filtration rate, an adjusted version of serum creatinine level in a blood sample. Clinical guidelines in the UK suggest that a person who is losing kidney function at a relative rate of at least 5% per year should be referred to specialist secondary care. We model the time-course of a person’s underlying kidney function through a combination of explanatory variables, a random intercept and a continuous-time, non-stationary stochastic process. We then use the model to calculate for each person the predictive probability that they meet the clinical guideline for referral to secondary care. We suggest that probabilistic predictive inference linked to clinical criteria can be a useful component of a real-time surveillance system to guide, but not dictate, clinical decision-making.

Keywords: Dynamic modeling; Kidney failure; Longitudinal data analysis; Non-stationarity; Real-time prediction; Renal medicine; Stochastic processes.

1. Introduction

In this paper, we consider the problem of using routinely collected data from a large sample of people in primary care to monitor the progression of undiagnosed incipient renal failure. The problem is important...
because chronic renal failure can be asymptomatic for many years. Early detection followed by initiation of ameliorative treatment can slow the rate of progression to end-stage renal failure and so postpone the need for expensive, invasive, and often scarce renal replacement therapy (dialysis or transplantation). Progression towards renal failure is characterized by a sustained fall in a person’s glomerular filtration rate (GFR). However, direct measurement of GFR is expensive. For this reason, many specialist renal treatment centres now use as a clinical indicator of a person’s renal function an estimated glomerular filtration rate (eGFR). This is calculated from a person’s age, sex, ethnicity, and their level of serum creatinine (SCr) as determined by a blood sample. A widely used formula is the modification of diet in renal disease equation (Levey and others, 1999),

\[
eGFR = 175 \times \left( \frac{SCr}{88.4} \right)^{-1.154} \times \text{age}^{-0.203} \times 0.742 \times I(\text{female}) \times 1.21 \times I(\text{black}). \tag{1.1}
\]

In common with other published formulae, (1.1) expresses a multiplicative relationship between eGFR (in mL/min per 1.73 m² of body surface area) and SCr (in µmol/L), adjusted by age, sex, and ethnicity. In our study, information on ethnicity is not available but the population is mostly Caucasian, and we have ignored the ethnicity component of (1.1).

The data that we analyze consist of repeated measurements of eGFR taken at irregular, person-specific follow-up times for 22,910 primary care patients who have been diagnosed with pre-disposing conditions for renal failure; these co-morbidities, and other relevant baseline information, are included in the data as a set of explanatory variables attached to each person’s record. The data were collected as part of a longitudinal cohort study run by the Salford Royal Hospital Foundation Trust (SRFT), Greater Manchester, UK.

Our proposed strategy, anticipating to some extent results from the preliminary analysis of the data as reported in Section 5, is to build a dynamic regression model in which a subject’s rate of change in log-transformed GFR, relative to the expected profile for all people with the same values of the explanatory variables, is modeled as a stochastic process \( B(t) \), which is realized independently for each person. Using the fitted model, we then evaluate the predictive distribution of \( B(t) \) for each person, conditional on their data up to and including time \( t \). Under this approach, a person would be flagged as a candidate for referral to a specialist treatment unit if and when the predictive probability that their current rate of decrease in GFR exceeds 5% per year is at least \( p_c \), where \( p_c \) is a threshold value to be specified by the clinician. The choice of \( p_c \) will determine the balance between sensitivity and specificity.

The remainder of the paper is organized as follows. In Section 2, we give a more detailed description of the SRFT data. Section 3 describes our proposed model. Section 4 describes associated methods for parameter estimation and for prediction of \( B(t) \). Section 5 describes the application to the SRFT data, including model diagnostics and individual predictions. Section 6 presents two simulation studies conducted to investigate the properties of the estimators of the model parameters and the influence of distributional and variance structure misspecifications on the predicted probabilities. Section 8 is a concluding discussion.

2. Data

The SRFT dataset contains information on 22,910 patients who entered the study between March 7, 1997 and March 22, 2007 and met the criterion of being at risk for renal failure. The patients provided a total of 392,870 values of eGFR. Of the 22,910 patients, 11,833 (51.65%) were male. Their baseline age ranged between 13.74 and 102.10 years with a median of 67.19 years. The number of eGFR values per patient ranged between 1 and 305 with a median of 12. Total follow-up time ranged from 0 (i.e., only one eGFR value) to 10.02 years with median of 4.46 years. Figure 1 shows the complete set of log-transformed eGFR values as a gray scatterplot, with longitudinal trajectories for a representative sample of six patients.
Fig. 1. Log-transformed eGFR measurements against follow-up time (in years). Data from a representative sample of six patients are highlighted as black lines.

highlighted as black lines. The data exhibit substantial variation in eGFR, both between patients and over time within patients.

3. **Model formulation**

We consider a general model for the longitudinal eGFR trajectories, of the form

\[ Y_{ij} = \mu_i(t_{ij}) + U_i + W_i(t_{ij}) + Z_{ij}. \]  (3.1)

In (3.1), \( Y_{ij} \) denotes the log-transformed eGFR response for subject \( i \) (\( i = 1, \ldots, M \)) at time \( t_{ij} \) (\( j = 1, \ldots, n_i \)). The function \( \mu_i(t) \) is the expected value of the response, which we represent as a multiple linear regression, hence \( \mu_i(t_{ij}) = X_i(t_{ij})\alpha \), where \( X_i(t_{ij}) \) denotes a set of explanatory variables and \( \alpha \) denotes the corresponding set of fixed effects regression parameters to be estimated. The \( U_i \) is independent \( N(0, \omega^2) \) random variables that represent time–constant differences among patients that cannot be explained by the linear regressions. The \( W_i(t) \) is independent copies of a zero-mean, continuous-time stochastic process representing change in a patient’s GFR over time that cannot be explained by the linear regressions. We assume that this continuous-time process is Gaussian, and therefore specified by its covariance function, \( \gamma(s, t) = \text{Cov}(W_i(s), W_i(t)) \). Finally, the \( Z_{ij} \) is mutually independent \( N(0, \tau^2) \) random variables representing measurement error in the determination of \( Y_{ij} \). Note that the model expresses eGFR as a noisy version of GFR.

The scalar-valued random effects \( U_i \) in (3.1) could be replaced by a second multiple linear regression, \( X^*_i(t_{ij})U_i \), where now the \( U_i \) are mutually independent, multivariate normal random variables. The
term $W_i(t_{ij})$ in (3.1) might then be omitted. A widely used example of such specification is the random-intercept-and-slope model in which $X_i(t_{ij}) = (1, t_{ij})$ (Laird and Ware, 1982). However, this is seldom realistic for long series (Henderson and others, 2000). Also, both the clinical context and visual inspection of longitudinal trajectories for individual patients, typical examples of which are shown in Figure 1, suggest that a random slope is too inflexible. Instead, we model $W_i(t)$ as the integral of a continuous-time random walk,

$$W_i(t) = \int_0^t B_i(v) \, dv,$$

where $B_i(v)$, the rate of change at time $v$, is Brownian motion. We set $B_i(0) = 0$ for every patient $i$, so that the random effects $U_i$ represent each patient’s deviation from their expectation at the time of their first eGFR measurement. The conditional distribution of $B_i(t)$ given $B_i(s)$ for some $s < t$ is normal, with mean $B_i(s)$ and variance $\sigma^2(t - s)$. It follows that unconditionally, and using $\lceil \cdot \rceil$ to mean “the distribution of”, $[B_i(t)] = N(0, \sigma^2 t)$ and $\text{Cov}(B_i(s), B_i(t)) = \sigma^2 \min(s, t)$. It then follows in turn that $[W_i(t)] = N(0, \sigma^2 t^3 / 3)$ and

$$\text{Cov}(W_i(s), W_i(t)) = \sigma^2 \frac{\min(s, t)^2}{2} \left( \frac{\max(s, t) - \min(s, t)}{3} \right).$$

The process $B(t)$ is Markov, but $W_i(t)$ is not, that is, in general $[W_i(t) | W_i(s), W_i(q)] \neq [W_i(t) | W_i(s)]$ for $q \leq s \leq t$. The bivariate process $(B_i(s), W_i(t))$ is bivariate Gaussian with zero means and cross-covariance structure

$$\text{Cov}(B_i(s), W_i(t)) = \sigma^2 \frac{\min(s, t)^2}{2},$$

and is Markov. For details, see Chapter 8 of Ross (1996) and Robinson (2010).

In what follows, we write the model equation (3.1) in the following condensed form,

$$Y_i = X_i \alpha + U_i + W_i + Z_i.$$

(3.4)

Here, $Y_i = (Y_i(t_{i1}), \ldots, Y_i(t_{in}))^T$, $X_i = (X_i(t_{i1})^T, \ldots, X_i(t_{in})^T)^T$ with $X_i(t_{ij}) = (1, X_{i1}(t_{ij}), \ldots, X_{ip}(t_{ij}))$, $\alpha = (\alpha_0, \ldots, \alpha_p)^T$, $U_i = U_i K_i$ where $K_i$ denotes an $n_i \times 1$ matrix of ones, $W_i = (W_i(t_{i1}), \ldots, W_i(t_{in}))^T$ and $Z_i = (Z_i(t_{i1}), \ldots, Z_i(t_{in}))^T$.

4. INFERENCE

4.1 Estimation

The distributional properties of $U_i$, $B_i$, $W_i$, and $Z_i$ as defined in Section 3 induce a multivariate normal distribution for $Y_i$,

$$[Y_i] = \text{MVN}(X_i \alpha, V_i(\phi)),$$

where $X_i$ and $\alpha$ are as before. Also,

$$V_i(\phi) = \omega^2 J_i + \sigma^2 R_i + \tau^2 I_i,$$

(4.1)

where $\phi = (\omega^2, \sigma^2, \tau^2)^T$, $J_i$ is an $n_i \times n_i$ matrix of ones, $R_i$ is an $n_i \times n_i$ matrix with $(j, k)$th element

$$\frac{\min(t_{ij}, t_{ik})^2}{2} \left( \frac{\max(t_{ij}, t_{ik}) - \min(t_{ij}, t_{ik})}{3} \right),$$

and $I_i$ is an $n_i \times n_i$ identity matrix.
We assume that repeated observations belonging to different patients are independent, that is, Cov($Y_i, Y'_j$) = 0 for $i \neq i'$. Then, the log-likelihood function can be written as

$$L(\theta) = \text{Constant} - \frac{1}{2} \sum_{i=1}^{M} \log(\det(V_i(\phi))) - \frac{1}{2} \sum_{i=1}^{M} (Y_i - X_i \alpha)^T V_i^{-1}(\phi) (Y_i - X_i \alpha), \tag{4.2}$$

where $\theta = (\alpha^T, \phi^T)^T$ and “det” denotes determinant of a square matrix. We obtain the maximum likelihood estimates (MLEs) of the parameters, $\hat{\alpha}$ and $\hat{\phi}$, by a Fisher-Scoring algorithm as described in Jenrich and Schluchter (1986). Let $\phi^m$ and $\hat{\alpha}^m$ denote the values of $\phi$ and $\alpha$ at the $m$th scoring step. Given $\hat{\phi}^m$, set

$$\hat{\alpha}^{m+1} = \left(\sum_{i=1}^{M} X_i^T V_i^{-1}(\hat{\phi}^m) X_i\right)^{-1} \left(\sum_{i=1}^{M} X_i^T V_i^{-1}(\hat{\phi}^m) Y_i\right), \tag{4.3}$$

and update $\phi$ to

$$\phi^{m+1} = \hat{\phi}^m + I_{\phi, \phi}^{-1} \phi^{m+1} S_{\phi, \phi}. \tag{4.4}$$

In (4.4), the $(r, s)$th element of $I_{\phi, \phi}$ is

$$\{I_{\phi, \phi}\}_{rs} = \frac{1}{2} \sum_{i=1}^{M} \text{trace}\left(V_i(\hat{\phi}^m)^{-1} \frac{\partial V_i(\phi)}{\partial \phi_r} V_i(\hat{\phi}^m)^{-1} \frac{\partial V_i(\phi)}{\partial \phi_s}\right), \quad r, s = 1, 2, 3 \tag{4.5}$$

and the $r$th element of $S_{\phi, \phi}$ is

$$\{S_{\phi, \phi}\}_r = \frac{1}{2} \sum_{i=1}^{M} \text{trace}\left(V_i(\hat{\phi}^m)^{-1} ((Y_i - X_i \hat{\alpha}^{m+1})(Y_i - X_i \hat{\alpha}^{m+1})^T - V_i(\hat{\phi}^m) V_i(\hat{\phi}^m)^{-1} \frac{\partial V_i(\phi)}{\partial \phi_r}\right), \quad r = 1, 2, 3. \tag{4.6}$$

In (4.5) and (4.6), the first partial derivatives of $V_i(\phi)$ are calculated by $\partial V_i(\phi)/\partial \omega^2 = J_i$, $\partial V_i(\phi)/\partial \alpha^2 = R_i$, and $\partial V_i(\phi)/\partial \tau^2 = I_i$. In our data analysis and simulations, we assess convergence by the criterion,

$$\sqrt{(\phi^m - \hat{\phi}^{m+1})^T (\phi^m - \hat{\phi}^{m+1})} < 10^{-10}.$$

At convergence, the large sample variance–covariance matrices of $\hat{\alpha}$ and $\hat{\phi}$ can be obtained by

$$\text{cov}(\hat{\alpha}) = \left(\sum_{i=1}^{M} X_i^T V_i^{-1}(\hat{\phi}) X_i\right)^{-1} \tag{4.7}$$

and $\text{cov}(\hat{\phi}) = I_{\phi, \phi}^{-1}$, respectively.

### 4.2 Prediction

The conditional distributions $[U_i | Y_i, \theta]$, $[W_i(t_{i,k}) | Y_i^k, \theta]$ and, especially, $[B_i(t_{i,k}) | Y_i^k, \theta]$, where $Y_i^k = (Y_{i1}, \ldots, Y_{ik})^T$, are of scientific interest. The first two are relevant to an individual’s prognosis, whilst the third is the predictive distribution of the underlying rate of change, which is the primary target in our application. The explicit forms of these distributions can be obtained using the properties of multivariate
normal distribution (Anderson, 1984) as

\[
[U_i | Y_i; \theta] = N(\omega^2 K_i^T V_i^{-1} (Y_i - X_i \alpha), \omega^2 (1 - \omega^2 K_i^T V_i^{-1} K_i)),
\]

\[
[W_i(t_{ik}) | Y_i^k; \theta] = N \left( \frac{\sigma^2}{2} F_i^{kT} (V_i^k)^{-1} (Y_i^k - X_i^k \alpha), \sigma^2 \left\{ \frac{t_{ik}^2}{3} - \frac{\sigma^2}{4} F_i^{kT} (V_i^k)^{-1} F_i^k \right\} \right),
\]

\[
[B_i(t_{ik}) | Y_i^k; \theta] = N \left( \frac{\sigma^2}{2} L_i^{kT} (V_i^k)^{-1} (Y_i^k - X_i^k \alpha), \sigma^2 \left\{ t_{ik} - \frac{\sigma^2}{4} L_i^{kT} (V_i^k)^{-1} L_i^k \right\} \right),
\]

where \( V_i^k \) is the variance–covariance matrix of \( Y_i^k \), \( K_i \) is as before, \( F_i^{k} = (t_{i1}^2(t_{ik} - t_{i1}/3), \ldots, t_{ik}^2(t_{ik} - t_{ik}/3))^T \) and \( L_i^{k} = (t_{i1}^2, \ldots, t_{ik}^2)^T \). Here, we suppress the dependence of \( V_i \) on \( \phi \). In our application, we substitute the MLVs of \( \theta \) into (4.8–4.10), because estimation errors are negligible compared with prediction errors. In smaller samples, Bayesian prediction with diffuse priors would provide a convenient, albeit pragmatic, means of accommodating estimation error.

Similarly, we use properties of the multivariate normal distribution to obtain the predictive distributions needed for forecasting \( W_i \) and \( B_i \) at time \( t_{ik} \) with lead-time \( u \) as

\[
[W_i(t_{ik} + u) | Y_i^k; \theta] = N \left( \frac{\sigma^2}{2} F_i^{k,uT} (V_i^k)^{-1} (Y_i^k - X_i^k \alpha), \sigma^2 \left\{ \frac{(t_{ik} + u)^3}{3} - \frac{\sigma^2}{4} F_i^{k,uT} (V_i^k)^{-1} F_i^{k,u} \right\} \right),
\]

\[
[B_i(t_{ik} + u) | Y_i^k; \theta] = N \left( \frac{\sigma^2}{2} L_i^{k,uT} (V_i^k)^{-1} (Y_i^k - X_i^k \alpha), \sigma^2 \left\{ (t_{ik} + u) - \frac{\sigma^2}{4} L_i^{k,uT} (V_i^k)^{-1} L_i^{k,u} \right\} \right),
\]

where \( F_i^{k,u} = (t_{i1}^2(t_{ik} + u - t_{i1}/3), \ldots, t_{ik}^2(t_{ik} + u - t_{ik}/3))^T \). Note that the expectation of \( [B_i(t_{ik} + u) | Y_i^k; \theta] \) is the same for all \( u \), but, its variance increases with \( u \). Whenever a new response \( Y_{i,k+1} \) becomes available, at time \( t_{ik} + u \) say, we update the predictions accordingly.

5. Application: SRFT dataset

Using log(eGFR) rather than eGFR as our response variable is better-matched to the clinical criterion for referral to specialist secondary care and improves the empirical fit of our data to the linear model. On the log-transformed scale eGFR is equivalent to SCr adjusted for sex and age. Nevertheless, we include sex and age as explanatory variables because standard formulae such as (1.1) are not optimal for prediction in particular sub-populations; see, for example, page 606 of Levey and others (2009). It follows that our predictive inferences are unaffected by whether we use log(eGFR) or log(SCr) as the response variable. Our main reason for working with eGFR rather than directly with SCr is that this is more easily interpretable by renal physicians.

We decompose age into age at entry and follow-up time in order to separate cross-sectional and longitudinal effects of age. This decomposition is strongly supported by a likelihood ratio criterion: for our final model, the maximized log-likelihood is 312, 251.81 when age at measurement-time is included as a single explanatory variable, 312, 425.83 when the cross-sectional and longitudinal effects of age are separated. Based on our preliminary analyses, we use a piece-wise linear model for the longitudinal age effect with a change of slope at the age 56.5 years. We code sex as 0 for males, 1 for females. The explicit form of our
Table 1. MLEs of the model parameters and the corresponding standard errors (SE)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>4.6006</td>
<td>0.0203</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>-0.0877</td>
<td>0.0048</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>-0.0048</td>
<td>0.0004</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>-0.0232</td>
<td>0.0011</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>-0.0075</td>
<td>0.0006</td>
</tr>
<tr>
<td>$\omega^2$</td>
<td>0.1111</td>
<td>0.0012</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.0141</td>
<td>0.0002</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>0.0469</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

linear model for a single patient at the $j$th measurement is therefore

$$
\log(eGFR)_{ij} = \alpha_0 + \alpha_1 \text{Gender}_i + \alpha_2 \text{Baseline age}_i + \alpha_3 t_{ij} + \alpha_4 \max(0, \text{age} - 56.5)_{ij} + U_i + W_i(t_{ij}) + Z(t_{ij}),
$$

and the corresponding rate of change in kidney function is

$$
\alpha_3 + \alpha_4 I(\text{age} > 56.5) + B_i(t_{ij}).
$$

5.1 Estimation

Table 1 shows the MLEs and standard errors for the regression parameters; all are (unequivocally) significantly different from zero. The estimate $\hat{\alpha}_0 = 4.6006$ establishes the average level of kidney function on entry, while $\hat{\alpha}_1 = -0.08768$ indicates that at study entry, females had 8.4% ($= 100 \times (\exp(-0.0877) - 1)$) worse kidney function on average than males. The estimate $\hat{\alpha}_2 = -0.0048$ indicates that the cross-sectional effect of age is a loss of kidney function at a rate of $\sim 0.5\%$ per year of age at entry; this is more or less in line with the general population, as would be expected. The estimate $\hat{\alpha}_3 = -0.0232$ corresponds to an average loss of kidney function at a rate of 2.3% per year until and including the age 56.5. This value, together with the estimate $\hat{\alpha}_4 = -0.0075$, indicates that the loss of kidney function accelerates to 2.9% after the age 56.5. These estimates of the loss of kidney function reflect the fact that all of the patients have been diagnosed with pre-disposing conditions for renal failure. MLEs of the covariance parameters are $\hat{\omega}^2 = 0.1111$, $\hat{\sigma}^2 = 0.0141$, and $\hat{\tau}^2 = 0.0469$.

5.2 Diagnostics

We apply diagnostic checks on our fitted model, using empirical residuals calculated as follows. First, for each subject $i$, let $r_i = Y_i - X_i\hat{\alpha}$. Now, denote by $\hat{V}_i$ the fitted variance matrix of $Y_i$, obtained by substituting $\hat{\phi}$ into (4.1), and write $\hat{V}_i = S_iS_i^T$, where $S_i$ is lower triangular. Finally, define the transformed (or standardized) empirical residual vector for patient $i$ as $r_i^* = S_i^{-1}r_i$.

To check the assumed form of the regression model, we inspect scatterplots of the residuals against the fitted values $\hat{Y}_{ij} = X_{ij}\hat{\alpha}$, and against the follow-up times, $t_{ij}$. Figure 2 shows the two scatterplots with superimposed LOWESS curves (Cleveland, 1979) obtained using the R (R Development Core Team, 2014) function lowess with the default value for the smoothing parameter. There is no discernible systematic pattern in either of the scatterplots and the fitted smooth curves are close to zero, suggesting a reasonable fit.
Real-time monitoring of progression towards renal failure

Fig. 2. Left panel: Scatterplot of fitted values versus standardized residuals. Right panel: Scatterplot of follow-up time (in years) versus standardized residuals. The dashed line is the zero line, the solid line a LOWESS smooth.

Fig. 3. The empirical variogram based on the transformed residuals against the lag based on the transformed time-scale. The variogram ordinates are averaged over bins with width 0.01. Bins with fewer than 30 residuals are omitted.

To check the assumed form of the covariance structure, we use the variogram of the empirical residuals (Diggle and others, 2002). The theoretical variogram is the function \( \delta(u) \), where \( \delta(u_{ijk}) = \frac{1}{2} E(r_{ij}^* - r_{ik}^*)^2 \) and \( u_{ijk} = |t_{ij} - t_{ik}| \). The empirical variogram is obtained by calculating empirical variogram ordinates, \( g_{ijk} = \frac{1}{2} (r_{ij}^* - r_{ik}^*)^2 \) and averaging all \( g_{ijk} \) corresponding to each unique value of \( u_{ijk} \) or, if follow-up times are completely irregular, by averaging all \( g_{ijk} \) corresponding to values of \( u_{ijk} \) within a pre-specified set of intervals. In a well-fitting model, the empirical variogram of the standardized residuals should fluctuate randomly \( \sim 1 \) when drawn against the lag, \( u_{ijk}' = |t_{ij}' - t_{ik}'| \), where \( t_{ij}' \) is elements of \( t_i' = S_i^{-1} t_i \) (Fitzmaurice and others, 2011). Figure 3 shows the empirical variogram for the fitted model. The empirical variogram ordinates show a decreasing trend from \( \sim 1.5 \) to \( 1 \) over the range 0–2 of differences of transformed times, and thereafter fluctuate \( \sim 1 \).
Fig. 4. The variances of the raw residuals over follow-up time, in years, (dots) and the theoretical variance function of the fitted model (solid line). Residuals are binned through time with bin size of 1 week and bins with <30 elements are omitted. Baseline data are treated separately.

We further assess the appropriateness of the assumed variance structure by comparing the variances of the (unstandardized) empirical residuals $r_{ij}$ and the theoretical variance implied by our model, $\text{var}(Y_{ij}) = \omega^2 + \sigma^2 t_{ij}^3 + \tau^2$, plugging-in the estimates $\hat{\omega}^2$, $\hat{\sigma}^2$, and $\hat{\tau}^2$. The empirical and theoretical variances are drawn in Figure 4. The empirical variances were calculated from binned residuals through time, with bin size of 1 week, but with baseline data treated separately, that is, variances are calculated separately at baseline and over follow-up measurements between 0+ and 7 days after baseline. The empirical variance of the baseline residuals is substantially smaller than the variances of the rest of the residuals, whose variance increases with time but at a substantially slower rate than the theoretical variance of the fitted model.

We also inspect the distributional assumptions of our model by examining the standardized residuals which are expected to follow standard normal distribution. Figure 5 shows four diagnostic plots. Collectively, these indicate that the residual distribution has the expected properties in the main body of the distribution, but heavier tails than the standard normal distribution.

Overall, the diagnostic plots show discrepancies between the data and fitted model, whose influence on predictive performance we investigate in Section 6.

5.3 Prediction

Figures 6 and 7 show predictions for four selected patients. Point predictions and prediction intervals for log(eGFR) were calculated from the mean, 2.5 and 97.5% quantiles of the conditional distribution, $[Y_{ij}|X_{ij}, \hat{\alpha}, \tilde{U}_i, \tilde{W}_i(t_{ij}), \tilde{\tau}^2] = N(X_{ij}\hat{\alpha} + \tilde{U}_i + \tilde{W}_i(t_{ij}), \tilde{\tau}^2)$, where $\tilde{U}_i$ and $\tilde{W}_i(t_{ij})$ correspond to the means of $[U_i|Y_i; \theta]$ and $[W_i(t_{ik})|Y_i^{ik}; \theta]$, given in (4.8) and (4.9), respectively. Point and interval predictions for the rate of change were similarly calculated from the conditional distribution $[B_i(t_{ik})|Y_i^{ik}; \theta]$ given in (4.10). The predictive probability, at each follow-up time, for the underlying rate of change being less than $-0.05$ was calculated as

$$p^*_i(t_{ik}) = P(B_i(t_{ik}) < -0.05 - \hat{\alpha}_3 - \hat{\alpha}_4 I(\text{age}_{ik} > 56.5)|Y_i^{ik}; \theta).$$  (5.3)
Fig. 5. Diagnostics plots on distributional assumptions based on standardized residuals. Upper left panel: quantile-quantile plot. Upper right panel: histogram with normal density superimposed. Lower left panel: empirical (solid line) and theoretical (dashed line) cumulative distribution functions. Lower right panel: the difference between the empirical and theoretical distribution functions.

The individual predictions for log(eGFR) seem very reasonable. The predicted means smooth out the erratic fluctuations in measured log(eGFR) and almost all the observed log(eGFR) measurements are covered by the 95% confidence intervals.

The predictive probability graphs of $p^*_{it}(t)$ as defined by (5.3) are shown in the third column of Figures 6 and 7. Our view is that these should guide, rather than determine, clinical decision-making. From this point of view, the four selected patients show interestingly different patterns.

1. For patient $i = 100$, the predictive probability $p^*_{it}(t)$ rises to a value slightly $>0.5$ after 2 years of follow-up, and thereafter fluctuates between $\sim0.3$ and 0.6. The appropriate clinical response would likely depend on factors other than those that can be encoded in a statistical model, for example, the patient’s general frailty and any co-morbidities.
Fig. 6. Plots of the predictions for two selected patients. Rows 1 and 2 correspond to patients $i = 100$ and $i = 9000$, respectively. Column 1 shows observed values of log(eGFR) (solid dots), predictive means (solid lines) and predictive 2.5 and 97.5% predictive quantiles (dashed lines). Column 2 shows predictive means (solid lines) and 2.5 and 97.5% predictive quantiles (dashed lines) of the underlying rate of change in log(eGFR). Column 3 shows the predictive probabilities, $p_i^*(t)$ that the underlying rate of change is less than $-0.05$.

Fig. 7. Plots of the predictions for two more selected patients, $i = 9600$ (row 1) and $i = 1278800$ (row 2). Details as for Figure 6.
2. For patient \( i = 9000 \), \( p^*_i(t) \) rises sharply to \( \sim 0.8 \) after 2 years, drops equally sharply between 2 and 4 years, then rises again. This is not atypical of patients experiencing progression towards renal failure. Indeed, this patient may well have received treatment within their primary care setting to reverse an acute loss of kidney function. This pattern is one example of something that the model cannot be expected to capture, but which does not necessarily negate its value as a predictive tool.

3. For patient \( i = 9600 \), \( p^*_i(t) \) rises inexorably from 0 to 1 during the first 2 years of follow-up. For this patient, referral to secondary care is clearly indicated.

4. For patient \( i = 1278800 \), the progress of \( p^*_i(t) \) is qualitatively similar to that for patient \( i = 113 \), but on a shorter time-scale and with an unequivocal indication of referral by 2.5 years.

6. Simulations

6.1 Simulation Study I

We first conducted a simulation study to investigate the properties of the parameter estimates under the model given by (5.1). For each simulation, the values of the explanatory variables were those of a random sample of 500 from the 22 910 patients in the SRFT data. The random effects, \( U_i \), \( W_i(t_{ij}) \), and \( Z_{ij} \) were then simulated from their assumed distributions as described in Section 3. The true parameter values for the simulations were the estimated values from our analysis of the SRFT data, as reported in Table 1. The simulation was replicated 500 times. The mean and standard deviations of the total number of repeated measurements in the simulated datasets were 8599 and 410, respectively, the variation being a consequence of the variation in the number of repeated measurements per patient in the SRFT data. Table 2 summarizes the results. The parameter estimates are approximately unbiased. The empirical standard deviations of the parameter estimates are close to the mean of the nominal asymptotic standard errors. Coverage is close to the nominal rate of 95%.

6.2 Simulation Study II

We conducted a second simulation study to examine the robustness of the predictive probabilities \( p^*_i(t) \) to the covariance structure misspecification as depicted in Figures 3 and 4, and to the heavy-tailed residual distribution as depicted in Figure 5.

For each simulation, the values of the explanatory variables were those of a random sample of 500 from the 22 910 patients in the SRFT data. We considered three data-generating mechanisms as follows. In the first case, we generated data according to (5.1) as for simulation Study I, again setting the parameter values as the estimates reported in Table 1. In the second case, we added a random effect, \( U^*_i I(t > 0) \), where \( U^*_i \sim N(0, 0.25) \) and is independent of \( U_i \). The value 0.25 for the variance of \( U^*_i \) was based on the empirical variances displayed in Figure 4, while the assumed independence of \( U_i \) and \( U^*_i \) was based on a preliminary analysis of the SRFT dataset using a linear mixed model with \( U_i \) and \( U^*_i \) as the random effects. In the third case, we reverted to the model given by (5.1), but with the \( Z_{ij} \) simulated from a \( t \) distribution with degrees of freedom 3, scaled by \( \tau = \sqrt{0.0469} \). In all three cases, we simulated \( B_i(t_{ij}) \) together with \( W_i(t_{ij}) \) as a bivariate process whose properties are as described in Section 3.

Properties of the predictive probabilities, \( p^*_i(t) \), for the underlying rate of change being less than \(-0.05\) are summarized by their area under the receiver operating characteristics (ROC) curve. Area under the ROC curve is calculated as time-invariant, since the random effects are simulated from their marginal distribution. The simulations are replicated 500 times for each case. The mean numbers of repeated measurements, over 500 replicate simulations, were 8574, 8533, and 8582 in cases 1, 2, and 3, with standard deviations 448, 398, and 409, respectively. Table 3 shows the mean and standard deviation of the area under
Table 2. Results of the simulation Study I

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Mean</th>
<th>Bias (%)</th>
<th>SD</th>
<th>meSE</th>
<th>CP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>4.6006</td>
<td>4.5874</td>
<td>−0.2869</td>
<td>0.1458</td>
<td>0.1379</td>
<td>94.0</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>−0.0877</td>
<td>−0.0842</td>
<td>−4.0437</td>
<td>0.0323</td>
<td>0.0323</td>
<td>94.0</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>−0.0048</td>
<td>−0.0046</td>
<td>−4.5885</td>
<td>0.0029</td>
<td>0.0027</td>
<td>94.0</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>−0.0232</td>
<td>−0.0227</td>
<td>−1.9979</td>
<td>0.0071</td>
<td>0.0072</td>
<td>95.6</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>−0.0075</td>
<td>−0.0078</td>
<td>3.3830</td>
<td>0.0041</td>
<td>0.0038</td>
<td>93.4</td>
</tr>
<tr>
<td>$\omega^2$</td>
<td>0.1111</td>
<td>0.1103</td>
<td>−0.6764</td>
<td>0.0078</td>
<td>0.0079</td>
<td>95.4</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.0141</td>
<td>0.0141</td>
<td>−0.0806</td>
<td>0.0010</td>
<td>0.0010</td>
<td>94.2</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>0.0469</td>
<td>0.0469</td>
<td>−0.0120</td>
<td>0.0008</td>
<td>0.0008</td>
<td>95.8</td>
</tr>
</tbody>
</table>

Columns give the parameter name (Parameter), the mean (Mean), percentage bias (Bias (%)), and standard deviation (SD) of the parameter estimates, the mean of the nominal standard errors according to standard likelihood asymptotic theory (meSE), and the percentage coverage of the corresponding approximate 95% confidence intervals (CP%).

Table 3. Results of the simulation Study II

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>0.7574</td>
<td>0.0186</td>
</tr>
<tr>
<td>Case 2 (variance structure misspecification)</td>
<td>0.7373</td>
<td>0.0278</td>
</tr>
<tr>
<td>Case 3 (heavy-tailed residuals)</td>
<td>0.7558</td>
<td>0.0186</td>
</tr>
</tbody>
</table>

Columns give the mean (Mean) and standard deviation (SD) of the area under the ROC curve, calculated from 500 replicate simulations for each cases 1, 2, and 3.

the ROC curve in each case. The influence of either covariance structure misspecification or heavy-tailed residuals is negligible.

7. Computational aspects

All computations were programmed in R and run on a PC with Windows 7 32bit, 4.00 GB RAM and 3.00 GHz processor. We have written an R package, `lmenssp`, available at [http://CRAN.R-project.org/package=lmenssp](http://CRAN.R-project.org/package=lmenssp), that implements parameter estimation and plug-in prediction for a range of non-stationary Gaussian process models. The supplementary material available at *Biostatistics* online includes the SRFT dataset and gives an exemplary R script for data analysis reported in Section 5. The computational time required for the estimation of the parameters was 60 min for the SRFT dataset, 34 s for a simulated dataset with 500 patients and 8462 repeated measurements. The predictions require less computational time, since we estimate the model parameters once and plug-in these estimates into the predictive distributions for each patient separately. For example, computational times required for predictions were 0.9, 1.0, 2.8, and 8.8 s for patients with 10, 100, 203, and 305 observations, respectively.

8. Discussion

We have used a large set of longitudinal clinical data to develop a statistical model for real-time monitoring of progression towards end-stage renal failure. Our specific objective was to provide predictive probabilities for the event that the underlying rate of change in a patient’s kidney function is less than $-0.05$, that is, a loss of at least 5% of kidney function per year. The value $-0.05$ is taken from current UK guidelines for
referral of a primary care patient to specialist secondary care. Our model is a linear mixed effects model in which between and within patient heterogeneities are captured by a random intercept and integrated Brownian motion, respectively.

We found discrepancies between the assumed and empirical distribution and variance structure of the residuals. However, simulations showed that the impact of these discrepancies on the predictive performance of the model is negligible. The finding regarding the influence of heavy-tailedness is in agreement with the results of Sweeting and Thompson (2012). We considered capturing the behaviour of the empirical variance depicted in Figure 4 by adding a random effect to (5.1) for post-baseline measurements, that is, \( U^*_i I(t_{ij} > 0) \) where \( U^*_i \sim N(0, \omega^*_2) \), but this resulted in a non-identifiable model. We also considered modeling only post-baseline data, differences from baseline and differences between successive observations. However, none of these gave any improvement in diagnostic performance. Similarly, inclusion of the baseline co-morbidity variables did not improve the diagnostics. As an alternative to the integrated random walk specification for the serially correlated random effects, we considered specifying \( W_i(t) \) in (5.1) as an integrated Ornstein–Uhlenbeck process (Taylor and others, 1994). For this specification, the underlying rate of change follows a stationary Gaussian process with exponential correlation function. However, the resulting fit suggested that the rate of change in kidney function behaves as white noise, which is biologically implausible.

The discussion of Figures 6 and 7 in Section 5.3 makes the general point that the evolution of kidney function in individual patients shows features that are unlikely to be captured by any statistical model. These features arise for a number of reasons, including the imperfection of eGFR as a measure of true kidney function (e.g., the underlying SCr assay is affected by changes in muscle mass) and the transient effects of minor acute illnesses that go unrecorded. For observational time series of this kind, fit to the data is less important than the ability to address the primary research question which is the probability that a patient is losing kidney function at a rate of 5% or more per year.

Our current algorithm for prediction requires inversion of matrices of dimension \( n_i \times n_i \), where \( n_i \) is the current number of eGFR measurements available for the \( i \)th patient. In principle, it would be computationally more efficient to use a Kalman filter algorithm (Kalman, 1960) to update predictive probabilities based on the most recent results, rather than re-calculating from scratch whenever new eGFR measurements are added to the data. By exploiting the result that the bivariate process \((B_i(t), W_i(t))\) is Markov, we can represent our model as a local linear trend model (Robinson, 2010; Durbin and Koopman, 2012). However, this formulation assumes that the rate of change is constant between successive measurements, which introduces an element of approximation. In our data, the maximum value of \( n_i \) is 305, most are much smaller and the associated exact computations are not burdensome. Similarly, in principle we might want to allow for parameter uncertainty when calculating the predictive probabilities, for example, by assigning Bayesian priors to the model parameters, \( \theta \) say, and replacing the plug-in predictive distribution, \([B_i(t_{ik})|Y_i; \hat{\theta}]\), by \( \int [B_i(t_{ik})|Y_i; \theta][\theta|Y_i]d\theta \). However, in our application the difference between the two is negligible as the prediction error in \( B_i(t) \) dominates the estimation error in \( \theta \).

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

This work was supported by a Lancaster University Health e-Research Centre studentship for Ö.A. We thank Philip Kalra, James Ritchie, and John New (Salford Royal Foundation Trust) for helpful discussions. Conflict of Interest: None declared.
References


[Received June 19, 2014; revised October 21, 2014; accepted for publication October 27, 2014]