Journal of Biomedical Informatics 76 (2017) 69–77

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Probabilistic broken-stick model: A regression algorithm for irregularly sampled data with application to eGFR

Norman Poh[a,b,*], Santosh Tirunagari[a,c], Nicholas Cole[d], Simon de Lusignan[d]

[a] Department of Computer Science, University of Surrey, UK
[b] QuintilesIMS, London, UK
[c] Surrey Clinical Research Center, Guildford, Surrey, UK
[d] Department of Clinical and Experimental Medicine, University of Surrey, UK

ARTICLE INFO

Keywords:
Chronic kidney disease
Electronic medical records
Estimated glomerular filtration rate
eGFR
Regression
Broken-sticks
Clinical time series

ABSTRACT

In order for clinicians to manage disease progression and make effective decisions about drug dosage, treatment regimens or scheduling follow up appointments, it is necessary to be able to identify both short and long-term trends in repeated biomedical measurements. However, this is complicated by the fact that these measurements are irregularly sampled and influenced by both genuine physiological changes and external factors. In their current forms, existing regression algorithms often do not fulfil all of a clinician's requirements for identifying short-term (acute) events while still being able to identify long-term, chronic, trends in disease progression. Therefore, in order to balance both short term interpretability and long term flexibility, an extension to broken-stick regression models is proposed in order to make them more suitable for modelling clinical time series. The proposed probabilistic broken-stick model can robustly estimate both short-term and long-term trends simultaneously, while also accommodating the unequal length and irregularly sampled nature of clinical time series. Moreover, since the model is parametric and completely generative, its first derivative provides a long-term non-linear estimate of the annual rate of change in the measurements more reliably than linear regression. The benefits of the proposed model are illustrated using estimated glomerular filtration rate as a case study used to manage patients with chronic kidney disease.

1. Introduction

The trend in measurements of clinical interest such as blood sugar, cholesterol or kidney function can provide insight into the change over time in a patient's condition. For patients with chronic illnesses such as diabetes and chronic kidney disease (CKD), monitoring of these measurements is necessary in order to effectively manage the condition. For example, in order for clinicians to make effective decisions about drug dosage, treatment regimens or when scheduling follow up appointments, it is necessary to know not only the value of these indicators, but also to have an idea of both the short- and long-term trajectory they are following. However, modelling the trend of biomedical measurements over the long-term can be complicated by both practical, e.g. the irregular taking of measurements and lengthy gaps between them, and biological considerations. For example, the primary indicator of kidney function, the estimated glomerular filtration rate (eGFR), can be influenced by, amongst other things, the level of protein in the diet, changes in muscle breakdown and the level of hydration [1]. This can lead to substantive variability in a patient's eGFR measurements [2,3]. Unfortunately, existing regression algorithms such as linear, polynomial and Gaussian process regression (GPR) [4] either cannot account for these challenges or do not satisfy the key clinical requirements of providing an easily interpretable model that can elucidate short- and long-term trends.

Biomedical measurements are irregularly sampled, posing an additional challenge to analysis. Prior work in time series analysis has strongly emphasised regularly sampled data, resulting in fewer methods that exist specifically for analysing irregularly sampled data. Despite methods for analysing irregular time series data directly having been employed successfully [5–7], the most common approach is still to transform the data to enforce regularity using either interpolation techniques or regression analysis [8]. However, with biomedical time series interpolation can present its own problems due to the measurements not always being taken at random, but rather requested at specific times by clinicians, e.g. as part of routine monitoring or as follow up to treatment. On the other hand, regression imposes a number of assumptions on both the variables and their relationships. For example, linear regression assumes a linear relationships between the dependent

* Corresponding author at: QuintilesIMS, London, UK.
E-mail address: norman.poh@quintilesims.com (N. Poh).

https://doi.org/10.1016/j.jbi.2017.10.006
Received 23 November 2016; Received in revised form 3 October 2017; Accepted 10 October 2017
Available online 16 October 2017

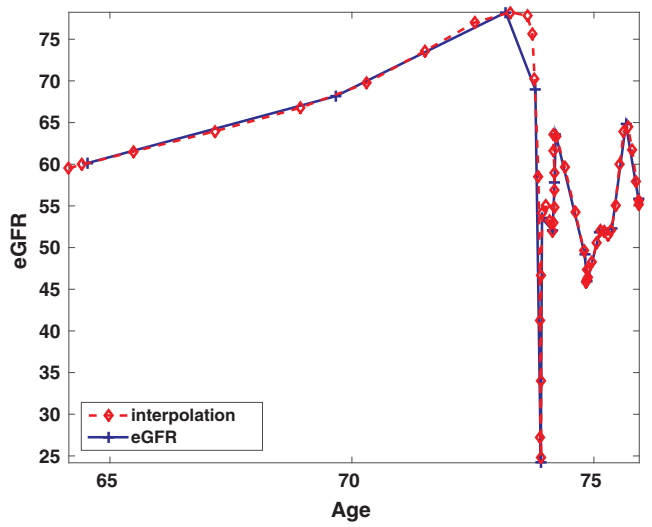


Fig. 1. An eGFR time series (blue) modelled using linear interpolation in order to produce a fixed-size vector of 50 observations (red) over the age range for which the patient has eGFR measurements [13]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and independent variables and independence of the residuals (no auto-correlation); assumptions which are usually violated in biomedical time series. Often linearity is violated due to an acute episode. For example, when a patient suffers an acute kidney injury (AKI) [9–12] their eGFR will drop sharply and potentially recover a short time after (as seen in Fig. 1). Long-term trends may therefore exhibit local fluctuations due to genuine physiological changes as well as external factors.

More flexible models such as Gaussian process regression (GPR) [14], multivariate adaptive regression splines [15] and multivariate additive models [16] can be used instead to provide the desired flexibility. For example, through the use of a kernel function GPR can avoid making the assumptions of linear regression. However, when there are gaps between the data, as is often the case with biomedical time series, the estimated variance of the predicted output can 'explode' [13] (Fig. 2). Consequently, these models are less interpretable, and therefore lose out in situations where a clinician simply needs to know whether a patient's condition is progressing or improving.

In order to strike a balance between interpretability and flexibility, broken-stick regression, also known as segmented or piece-wise regression, can be used to linearly model local trends [17–20]. However,

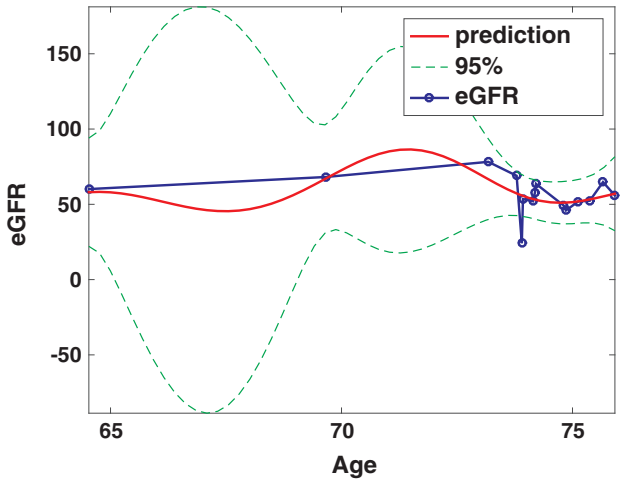


Fig. 2. The GPR model shows relatively low variance when the gap between measurements is small, but the variance increases markedly when the measurements are sparse.

Table 1
Notation.

Variable	Domain	Meaning
$\mathbf{T}$	Vector of real numbers	The time domain
t	Integer	Enumerator of the time domain, from 1 to T
w	Integer	Enumerator of the window, from 1 to W
$\mathbf{L}$	Vector of integers	Indices storing the beginning of window
$\mathbf{U}$	Vector of integers	Indices storing the end of a window
θ_w	Model parameters	Parameters of the w-th line segment
θ	All model parameters	$\{\theta_w \mid \forall_w\}$
Δ_d	Integer	Window interval of length d in year
W	Integer	Number of windows
$\omega_1^{(w)}$	Integer	Line segment gradient
$\omega_0^{(w)}$	Integer	Line segment intercept
$\mu_t^{(w)}$	Integer	Mean value of the time window

in this formulation local discontinuities are introduced at the segment boundaries, resulting in a loss of smoothness and consequently in the inability to infer trends in the boundary regions reliably. To address this we take a Bayesian approach to derive a long-term trend by enforcing a smooth transition between the locally linear line segments, while still preserving the local trends. The ability to capture both long- and short-term trends makes this approach ideally suited to modelling biomedical time series in a clinical context. Additionally, by enforcing smoothness local rates of change can be derived, giving clinicians an indication of whether a patient's condition is progressing or not. Finally, a broken-stick model can accommodate gaps in a time series through choosing the length of each line segment in a manner that ensures that there are a sufficient number of measurements within each segment and can mitigate overfitting as it fits only locally linear line segments.

2. Methodology

Here, $\mathbf{X}$ is used to denote a vector and $\mathbf{X}[t]$ to denote the element in the vector indexed by t. The remainder of the notation used is given in Table 1.

2.1. Windowing

The first step in fitting the broken-stick model is the division of a time series into a number of windows. Here, windows of equal length d were used across all time series, although there is no constraint requiring the windows to be of equal length across or within individual time series. The window length was determined from the data based on the intervals between measurements, as there should be at least three measurements within each window in order to avoid overfitting line segments. In general, having more measurements within each window is preferable. However, it is only possible to influence the number of measurements within a window by increasing d, as the number of measurements in each time series is fixed. Given that larger values of d may result in local fluctuations going undetected, while smaller values of d may lead to measurement noise dominating the model, the window length must be optimised for each application.

2.2. Local fitting

Given d and a specified interval to slide the window by, Δ_d, the number of windows W is also determined. For each window, a linear regression is performed by:

$$\mu^{(w)}(t) = \omega_1^{(w)} \times t + \omega_0^{(w)}, \quad (1)$$

where $\omega_1^{(w)}$ is the gradient and $\omega_0^{(w)}$ is the intercept for the w-th window

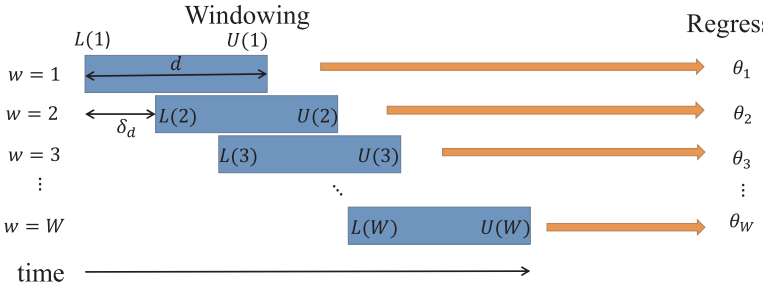


Fig. 3. A time series is broken into W windows of length d. For each window, a linear regression model is fit.

$\theta_w \equiv [\omega_0^{(w)}, \omega_1^{(w)}]$. The fitting of each window is summarised in Fig. 3.

2.3. Window influence

In order to smoothly join the fitted line segment we take a Bayesian approach. Let $p(x|t)$ be the distribution of the repeated measurement x given the time t. The local regression models will then give $p(x|w,t)$ for $w = 1,...,W$. These two quantities can be seen to be related by:

$$p(x|t) = \sum_w p(x|w,t)P(w|t), \tag{2}$$

where $P(w|t)$ is the posterior probability of the w-th window at time t.[1] Ideally the further away in time a window is from t the less influence its line segment has near t. One way to achieve this is to use the Bayes' theorem to define $P(w|t)$ such that $P(w|t) \propto p(t|w)$ and the window function $p(t|w)$ is bell-shaped, e.g. is Gaussian:

$$p(t|w) = \mathcal{N}(t|\mu_t^{(w)}, \sigma^{(w)}) \tag{3}$$

where $\mu_t^{(w)}$ is the mean value, i.e. the mid-point, of the time window (see Fig. 3):

$$\mu_t^{(w)} = \frac{[L(w)] + [U(w)]}{2}$$

and $\sigma^{(w)}$ is the standard deviation. Note that $\sigma^{(w)}$ must be a function of d, not the window, as the standard deviation does not vary from one window to another. Here, $\sigma^{(w)} = \sigma = \frac{1}{\alpha}d$ so that α enables us to define the decay of the function in terms of distance from the mid-point of the time window.

Having defined $p(t|w)$, we can use Bayes' theorem to obtain $P(w|t)$:

$$P(w|t) = \frac{p(t|w)P(w)}{\sum_{w'} p(t|w')P(w')} \tag{4}$$

where the prior, $P(w)$, dictates which windows should be given more weight.[2] When a flat prior is used, we have:

$$P(w|t) = \frac{p(t|w)}{\sum_{w'} p(t|w')} \tag{5}$$

The difference between $p(t|w)$ and $P(w|t)$ is depicted in Fig. 4.

2.4. Predicted values and confidence intervals

Since the expected value of the final regression for a given t is given by

$$\mu(t) \equiv E_{x\sim P(x|t)}[x] = \int x \cdot p(x|t)\, d_x, \tag{6}$$

it follows that we can compute the mean value by plugging (2) into (6).

$$\begin{aligned}\mu(t) &= \int x \cdot p(x|t)\, d_x = \int x \cdot \sum_w p(x|w,t)P(w|t)\, d_x \\ &= \sum_w \int (x \cdot p(x|w,t)\, d_x)P(w|t) \\ &= \sum_w \mu_t^{(w)} P(w|t).\end{aligned} \tag{7}$$

Therefore, the global mean regression function is a weighted sum of the local mean regression functions, with weights determined by $P(w|t)$ at any given time point t. Taking approximately 95% of the probability mass, the intervals around the global mean can be defined by:

$$\mu^L(t) = \sum_w (\mu_t^{(w)} - \sigma_t^{(w)})P(w|t) \tag{8}$$

and

$$\mu^U(t) = \sum_w (\mu_t^{(w)} + \sigma_t^{(w)})P(w|t) \tag{9}$$

respectively, for the lower and upper intervals. Therefore, we have $\mu^L(t) \leqslant \mu(t) \leqslant \mu^U(t)$. The local line fitting for a patient's time series can be seen in Fig. 5(a), with the final fitted curve in (b).

2.5. Computing the rate change

Another useful quantity that can be derived from the global model parametrically is the annual rate change of the time series. To do so, for each of the W line segments fitted using (1), we first compute its first derivative $\omega_1^{(w)}$. The rate change can then be computed as:

$$\mu'(t) = \sum_w \omega_1^{(w)} P(w|t) \tag{10}$$

Even though the local gradient $\omega_1^{(w)}$ for a given stick w does not change over time, the global, derived gradient $\mu'(t)$ still does because it is a result of interpolation weighted by $P(w|t)$, as shown in the equation above. The global mean trend of a patient's eGFR, given by (7), and the slope, computed using (10), can be seen in Fig. 6 for four randomly selected patients.

2.6. Overall algorithm

The overall algorithm consists of two phases, namely the model fitting phase (Algorithm 1) and the inference phase (Algorithm 2). Following the model fitting, the calculated model parameters $\theta \equiv \{\theta^{(w)}|w = 1,...,W\}$ and window bounds, $\mathbf{U}$ and $\mathbf{L}$, can be used in the inference phase to obtain the trend and slope for the time series.

[1] In the above numbered equation, we have used 'P' (in upper-case letter) to denote the probability of a discrete variable whereas 'p' (in lower-case letter) to denote a probability density function which is applicable to a continuous variable.

[2] In (4), we have used w' in the denominator to distinguish w in the numerator which refers to a specific window for which the posterior probability is calculated. w', in contrast, iterates through all windows. The denominator term is known as evidence whereas the numerator is composed of two terms, namely the likelihood term, i.e., $p(t|w)$, and the prior term, i.e., $P(w)$.

Algorithm 1. Model fitting

Input : Upper and lower window bounds, U, L
Repeated measurements, $\{x_t | t = 1, \ldots, T\}$
Output: $\{\theta_w | w = 1, \ldots, W\}$
$W = \text{length}(U)$
% Obtain the window length **for** $w \in 1, \ldots, W$ **do**
  $\theta_w = fit(\{x_{L[w]}, \ldots, x_{U[w]}\})$
end

Algorithm 2. Model inference

Input : Upper and lower time bounds, $t_{\text{U}}, t_{\text{L}}$
$\{\theta_w | w = 1, \ldots, W\}$
Upper and lower window bounds, U, L
Output: trend, $\{\mu_t \pm \sigma_t | t\}$
annual rate change, $\{\mu'_t | t\}$
$W = \text{length}(U)$ % Obtain the window length
% Create the window
for $w \in 1, \ldots, W$ **do**
  $\mathbf{b}_w = \mathcal{N}(t | \frac{t_{L[w]} + t_{U[w]}}{2}, \sigma)$ for $t \in \mathbf{T}$
end
% Normalize the window weight, implementing Equation (5)
for $t \in 1, \ldots, T$ **do**
  for $w \in 1, \ldots, W$ **do**
    $b_w[t] = \frac{b_w[t]}{\sum_{w'} b_{w'}[t]}$
  end
end
% Get the local trends
$\mathcal{T} = \text{sample}(t_{\text{U}}, t_{\text{L}}, 1000)$ %draw 1000 equally-spaced samples
for $w \in 1, \ldots, W$ **do**
  $[\mu^{(w)}(t), \sigma^{(w)}(t)] = \text{infer}(\theta_w)$, for $t \in \mathcal{T}$
  $\mu'_w(t) =$ Calculate first derivative (θ_w), for $t \in \mathcal{T}$
end
% Combine the local trends and the weights as in Equations (7–10)
for $t \in 1, \ldots, T$ **do**
  $\mu[t] = \sum_w \mu_w[t] \times b_w[t]$ % mean value
  $\mu^U[t] = \sum_w (\mu_w[t] + \sigma_w[t]) \times b_w[t]$ % upper interval
  $\mu^L[t] = \sum_w (\mu_w[t] - \sigma_w[t]) \times b_w[t]$ % lower interval
  $\mu'[t] = \sum_w \mu'_w[t] \times b_w[t]$ % annual rate change
end

3. Effect of the window length and interval

In order to illustrate the effect of the window length d and the window interval Δ_d, different parameter pairs (d,Δ_d) were used to fit an eGFR time series in which no AKI was observed. The results of this can be seen in Fig. 7. From this it can be seen that shorter window lengths and shorter intervals produce more sensitive models, as can be seen in the differences between (c) and (f) and between (e), (f) and (g) in terms of the magnitude of the slope. The impact of choosing windows lengths and intervals that are too large can also be seen in the lack of local fluctuations in (g). From this it is reasonable to conclude that the expected eGFR and slope are only comparable between patients when the same fitting parameters are used.

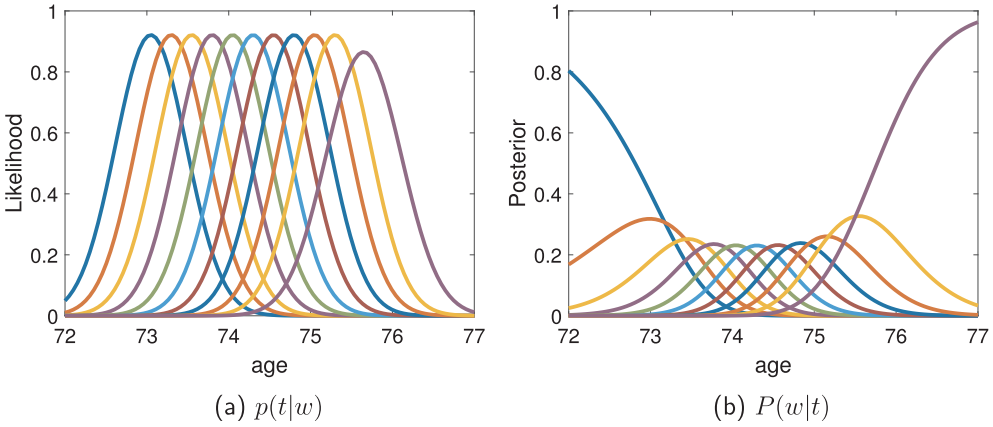
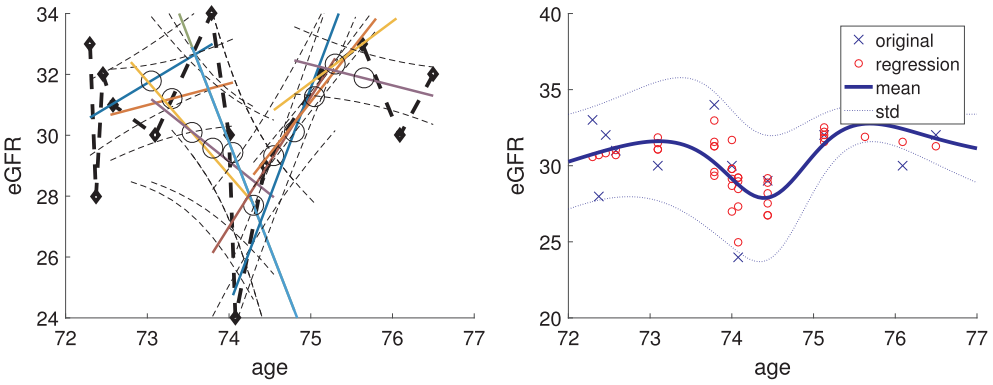


Fig. 4. An application of windowing with $d = 2$ years and $\Delta_d = 1/2$ a year. The result for each window $w = 1,...,11$ is plotted in a different colour. As the posterior probability has the property that $\sum_w P(w|t_*) = 1$, the first and last windows dominate the posterior probability near the boundaries, i.e. near $L[1]$ and $U[11]$.



(a) Local line segments (b) Final model

Fig. 5. Fitting of a broken-stick model to an eGFR time series. (a) 11 locally linear line segments are fitted to the time series. The dark dashed line represents the raw data, with the individual line segments plotted in different colours (along with their confidence intervals as dashed lines). The mid-point of each line segment is marked as an unfilled black circle. (b) The final predicted mean value $\mu(t)$ (blue line) with its confidence intervals (dashed lines). The red circles show the predicted local mean values $\mu^w(t)$ at the time points t where actual eGFR measurements occur. The final fitted curve can be seen to be globally non-linear despite its locally linear constituent lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Case study

In order to demonstrate the utility of the proposed broken-stick model, we applied it to primary care data collected for the Quality Improvement in Chronic Kidney Disease (QICKD) trail [21,22] in order to model the long-term trend of eGFR measurements. For patients with CKD, their eGFR is one of the primary outcomes used by clinicians and is used in determining the stage, and therefore severity, of a patient's CKD. While a true clinical staging of CKD will take kidney damage, as evidenced by the level of albuminuria, into account as well as eGFR, we focus only on eGFR here in order to demonstrate the utility of the proposed broken-stick model. The possible stages of CKD as a function

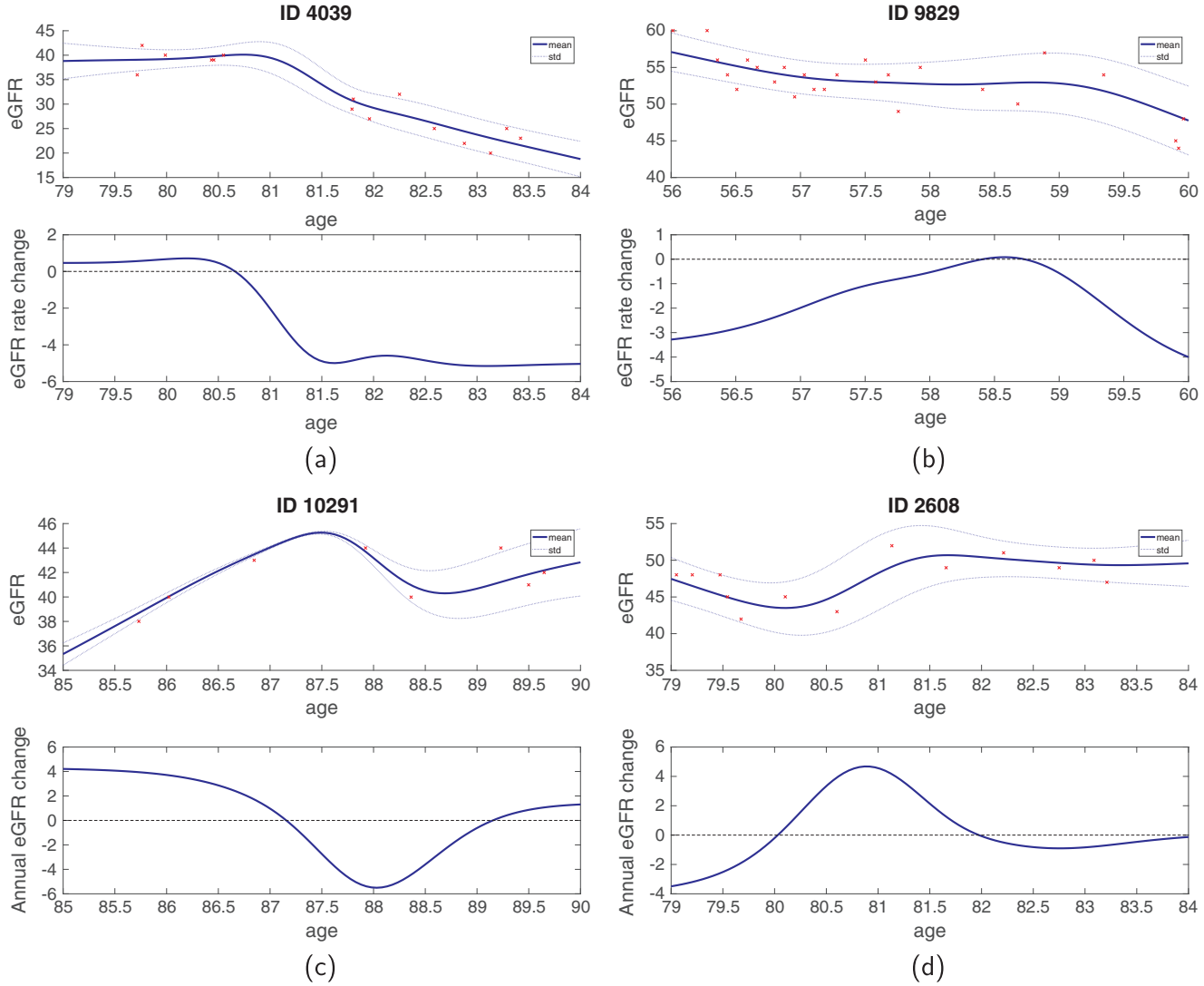


Fig. 6. Four examples of eGFR time series modelled using the broken-stick model (upper diagram) along with their corresponding annual rate change, i.e. the slope of the fitted model, (lower diagram).

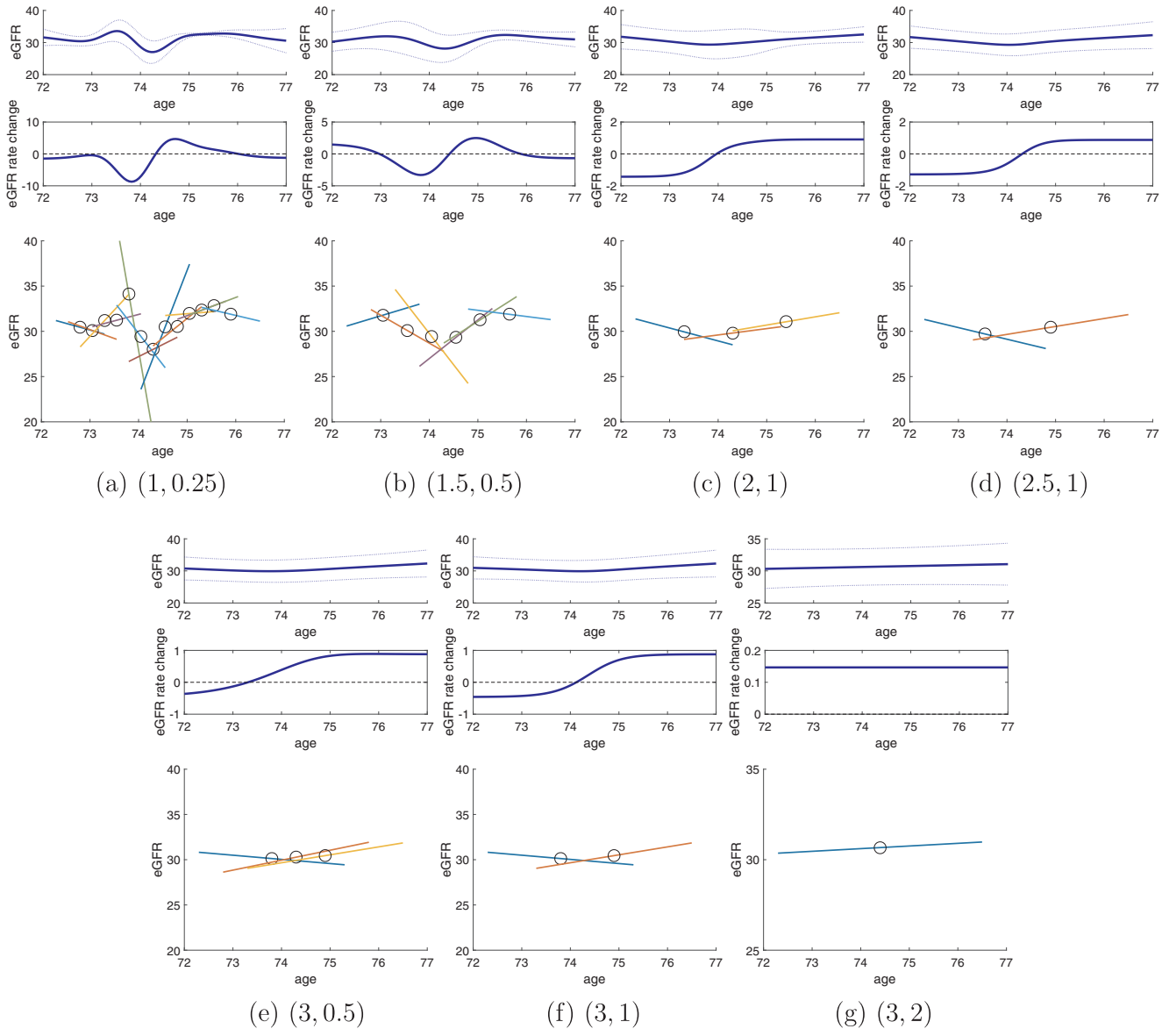


Fig. 7. An illustration of fitting an eGFR time series with different window length d and window interval Δ_d parameters (in years). The raw eGFR time series can be found in Fig. 5(a). For each figure (a)–(g) the top subfigure is the mean eGFR over time, the middle is the eGFR slope over time and the bottom the line segments of the broken-stick model.

Table 2
Definition of CKD stages. Here, g_L and g_U represents lower and upper boundaries of the eGFR values. †: eGFR values greater than 120 mL/min/1.73 m^2 are largely recognised as being inaccurate. For this reason 120 mL/min/1.73 m^2 is used as the cut-off between stages 0 and 1, despite the original Kidney Disease: Improving Global Outcomes (KDIGO) guideline [23] not defining any upper value.

CKD stage	g_L	g_U
0	120	∞
1	90	120†
2	60	90
3	30	60
4	15	30
5	0	15

of a patient's eGFR value can be seen in Table 2, with the caveat that patients without CKD are defined here as having stage 0 CKD.

4.1. The QICKD dataset

The QICKD dataset contains the primary care records of 951,764 patients. Of these records, 12,297 contain an eGFR measurement (45.4% male and 56.6% female). In total, there were 109,397 eGFR measurements across the patients, with approximately 95% between the values of 25 and 120 mL/min/1.73 m^2 and occurring in patients between the ages of 60 and 103. Fig. 8 summarises the main characteristics of the dataset. Based on the patient statistics, a window length of three years and window interval of half a year were chosen as the most appropriate trade-off between smoothness and capturing local trends. Due to this, 1546 of the 12,297 patients with an eGFR measurement were excluded for having less than three years worth of measurements and twenty-six for having gaps between measurements of larger than three years. Ten patients were also discarded for having an overly large gradient, likely as a result of large gaps between measurements isolating individual measurements. Each patient's eGFR sequence was also labelled, using the SAKIDA algorithm [9], with the number of acute kidney injury (AKI) episodes experienced by the patient. As an AKI represents a sudden and substantive change in the eGFR trend, and could therefore interfere with the trend modelling, the 1103 patients identified as having experienced an AKI episode are excluded. In total 2603 patients were excluded, leaving 9694.

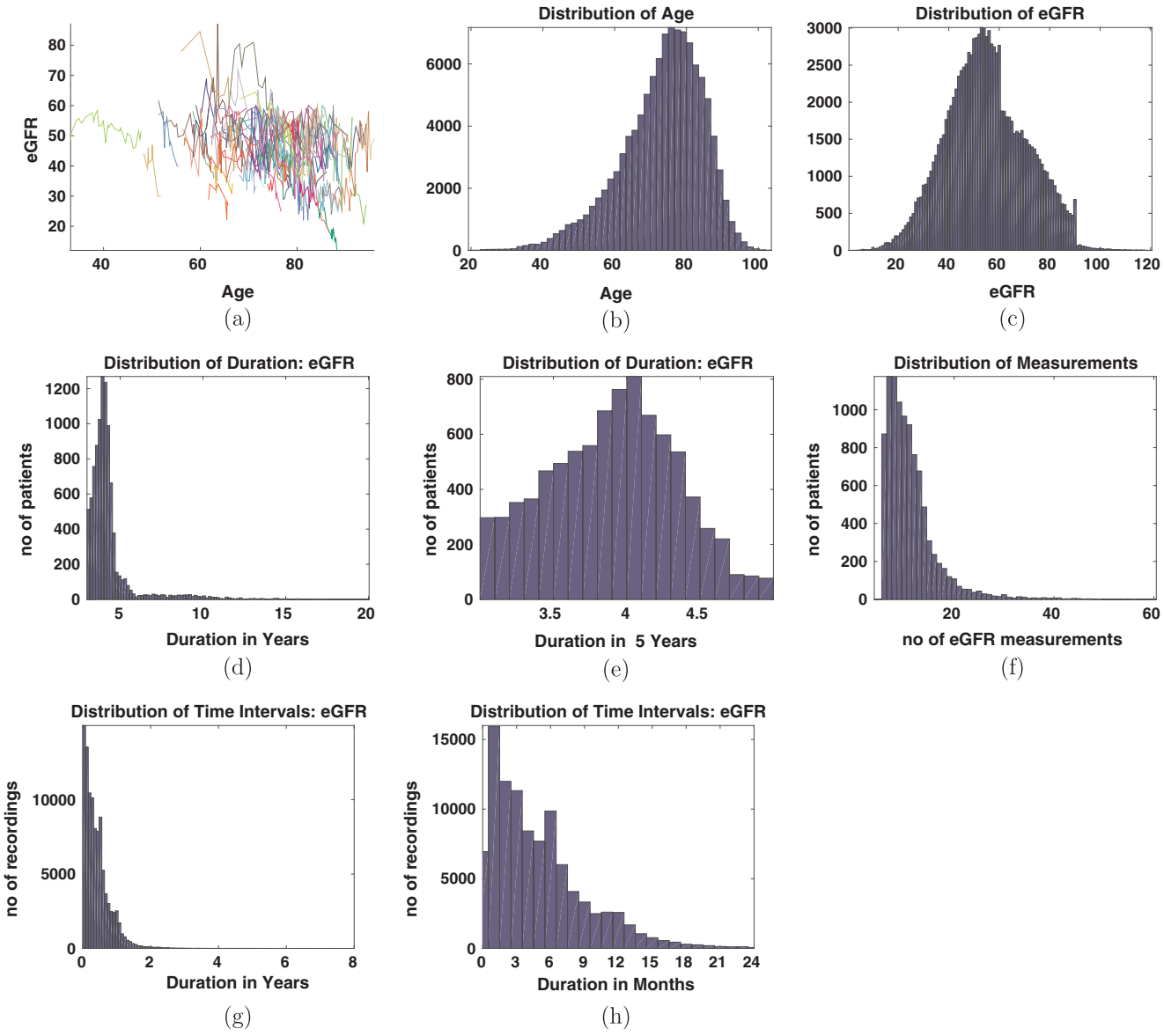


Fig. 8. Patients' eGFR signals are observed at irregular time intervals and over different age ranges. (a) eGFR sequences for a subset of 100 patients (each colour represents a single patient). Dataset characteristics: distributions of the (b) ages over which patients had eGFR measurements recorded, (c) eGFR measurement values, (d) duration of time over which all of a patient's eGFR measurements were recorded, (e) the same as (d) but limited to those patients with all measurements occurring within five years, (f) number of eGFR measurements recorded per patient, (g) time intervals between consecutive eGFR measurements and (h) the same as (g) but limited to consecutive measurements occurring between one and twelve months apart.

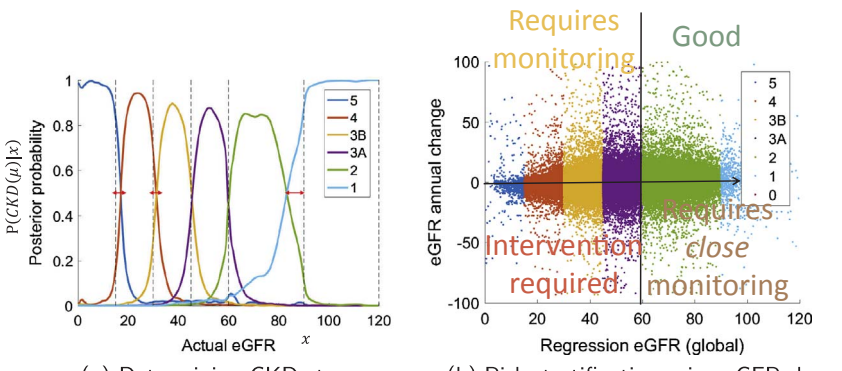


Fig. 9. Two applications of the broken-stick model. (a) Expected CKD stage posterior probabilities given raw eGFR values (the legend shows the expected CKD stage). Gaps between the KDIGO guideline stage (dashed vertical lines) and the boundary between expected CKD stages can be observed for stages 4/5, 3b/4 and 1/2. (b) Stratification of CKD patients using the expected eGFR slope (expected annual rate change). The use of the expected eGFR slope enables both the staging and trajectory of a patient's eGFR measurements to be taken into account.

From Fig. 8(c) it can be seen that there are abrupt changes in the distribution of the measurements at both 60 and 90 mL/min/1.73 m^2. There are likely two reasons for this:

- Measurements above 60 or 90, depending on the testing laboratory, are truncated and reported as either 60 or 90 respectively. This practice results in a peak in the distribution at 60 and 90 mL/min/1.73 m^2.
- The reliability of eGFR measurements decreases as the value of the measurement increases due to the associated variance increasing with the value of the measurement. Therefore, eGFR measurements above certain values, 60 and 90, are often reported as greater than 60 or 90 but a numeric value is not recorded in the patient's record. This likely accounts for the observed drop in the distribution following 60 and 90 mL/min/1.73 m^2.

Due to the increased variance of larger eGFR measurements, any values above 120 mL/min/1.73 m^2 are removed from the patient's time series and excluded from the study.

4.2. Staging and stratifying CKD

Accurate staging of CKD is dependent on being able to take a stable measurement of a patient's eGFR. Therefore, one of the major drawbacks of current approaches to CKD staging is the variability of eGFR measurements due to their sensitivity to natural fluctuations in the breakdown of protein, e.g. from changing levels of protein in the diet, muscle breakdown and hydration, in addition to a range of other clinical factors. It is therefore possible for the trend in a patient's eGFR to be misrepresented due to benign external factors, thereby masking their more serious decline in kidney function and frustrating clinicians' attempts to reliably determine a patient's CKD stage.

As an alternative to relying on the, potentially noisy, raw eGFR values when making staging decisions, here we use the estimated mean eGFR value obtained directly from the broken-stick model. To this end, we use a Bayesian framework to formulate the problem. Let g_t be the eGFR measurement taken at time t and $\{g_t | t \in \mathscr{T}\}$, where $\mathscr{T} \equiv \{1,...,T\}$, be an eGFR time series. Then, after applying the broken-stick model, we have the corresponding regressed mean $\{\mu_t | t \in \mathscr{T}\}$. Using this notation, CKD stages can be determined using the following equation:

$$CKD(g) \equiv g_{\mathrm{L}}^{l} < g \leqslant g_{\mathrm{U}}^{l} \tag{11}$$

where g is a raw eGFR measurement value and $[g_{\mathrm{L}}^{l}, g_{\mathrm{U}}^{l}]$ is the eGFR range that defines a given CKD stage $l \in \{0,1,2,3,4,5\}$. The upper and lower bounds for each CKD stage are shown in Table 2. Given that there are no available ground-truth CKD stages, and as μ is an estimated eGFR value, we can perform the staging using it, rather than g, via the following equation:

$$CKD(\mu) \equiv g_{\mathrm{L}}^{l} < \mu \leqslant g_{\mathrm{U}}^{l} \tag{12}$$

In order to ascertain whether the CKD stages determined using μ and g are consistent with one another, the distribution $p(g|CKD(\mu))$ was calculated for each CKD stage and the following equation used to determine the posterior probability of an expected given the raw eGFR value g:

$$P(CKD(\mu)|g) = \frac{p(g|CKD(\mu))P(CKD(\mu))}{\sum_{CKD(\mu^{\dagger})} p(g|CKD(\mu^{\dagger}))P(CKD(\mu^{\dagger}))}$$

where $CKD(\mu)$ ranges from 0 to 5. In order to prevent the prior dominating the posterior, a uniform prior $P(CKD(\mu))$ was used, resulting in the following equation:

$$P(CKD(\mu)|g) = \frac{p(g|CKD(\mu))}{\sum_{CKD(\mu^{\dagger})} p(g|CKD(\mu^{\dagger}))} \tag{13}$$

The posterior probability distributions calculated using (13) can be seen in Fig. 9(a). It is noticeable from this that the boundaries used to determine CKD stages from raw eGFR values, according to the KDIGO guidelines [23] in Table 2, are not consistent with the expected stage boundaries.

In addition to using the broken-stick model to determine CKD stages, by calculating the eGFR slope using (10) it is possible to both stage and stratify patients according to the trajectory that their condition is taking. By calculating both the expected eGFR value (μ) and slope (μ') at a given point, and recognising that stages 1 and 2 are often considered to be mild CKD, it is possible to stratify patients into four rough categories based on their outlook:

- Good: Patients in this category have mild, or no, CKD and a positive trajectory.
- Requires monitoring: Patients in this category have more severe CKD but show a positive trajectory, and may therefore be less likely to have their CKD worsen.
- Requires close monitoring: Patients in this category are characterised by having mild CKD but a worsening outlook due to the negative eGFR slope.
- Intervention required: In this category patients have advanced and worsening CKD.

5. Discussion and conclusions

The proposed broken-stick model can robustly estimate both short-term and long-term trends simultaneously, while also accommodating the unequal length and irregularly sampled nature of clinical time series. This can provide clinicians with a powerful tool for understanding the overall trajectory of a patient's disease by smoothing out local fluctuations in a parameterised manner. Within the management of CKD, the two primary uses of eGFR are determining the stage of a patient's CKD and determining the likely progression of the condition. While CKD staging is currently based on local trends, in this case the most recent eGFR measurements, by modelling a patient's eGFR time series using a broken-stick model it is possible to base a patient's stage on their entire time series. Conversely, evaluation of CKD progression can be based on both short- and long-term trends, and is difficult to evaluate from raw eGFR values alone. As a demonstration of the utility of the broken-stick model for assisting in the management of CKD, it was applied to the eGFR time series contained in the electronic medical records of approximately 10,000 patients. When compared to the CKD staging following the KDIGO guidelines, the stages determined using the broken-stick model are largely consistent, with the exception of between stages 1 and 2 where eGFR measurements are less frequently recorded as well as less reliable. Given this consistency, our gradient-based patient stratification is likely to prove reliable as it relies on the same model. Taken together, these results could provide useful information when determining the trajectory of a patient's condition and in the retrospective identification of patients with varying rates of decline for clinical research. Additionally, given its flexibility and wide applicability, the probabilistic broken-stick model could be applied to the modelling of other biomedical measurements, such as plasma glucose in diabetes.

Conflict of Interest

Authors declare that there is no conflict of interest.

Acknowledgements

This work was supported by the Medical Research Council under Grant No. MR/M023281/1. The project details can be found at http://www.modellingckd.org/.

References

[1] S. de Lusignan, C. Tomson, K. Harris, J. Van Vlymen, H. Gallagher, 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 (2010) c213–c224.

[2] N. Poh, S. de Lusignan, Calibrating longitudinal egfr in patience records stored in clinical practices using a mixture of linear regressions, in: International Workshop on Pattern Recognition for Healthcare Analytics, 21st International Conference on Pattern Recognition (ICPR).

[3] N. Poh, S. Tirunagari, D. Windridge, Challenges in designing an online healthcare platform for personalised patient analytics, in: 2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD), IEEE, pp. 1–6.

[4] C.K. Williams, C.E. Rasmussen, Gaussian Processes for Machine Learning vol. 2, the MIT Press, 2006, p. 4.

[5] M. Schulz, K. Stattegger, Spectrum: Spectral analysis of unevenly spaced paleoclimatic time series, Comput. Geosci. 23 (1997) 929–945.

[6] P. Stoica, P. Babu, J. Li, New method of sparse parameter estimation in separable models and its use for spectral analysis of irregularly sampled data, IEEE Trans. Signal Process. 59 (2011) 35–47.

[7] K. Rehfeld, N. Marwan, J. Heitzig, J. Kurths, Comparison of correlation analysis techniques for irregularly sampled time series, Nonlinear Process. Geophys. 18 (2011) 389–404.

[8] D. Kreindler, C. Lumsden, The effects of the irregular sample and missing data in time series analysis, Nonlinear Dyn. Psychol. Life Sci. 10 (2006) 187–214.

[9] S. Tirunagari, S.C. Bull, A. Vehtari, C. Farmer, S.d. Lusignan, N. Poh, Automatic detection of acute kidney injury episodes from primary care data, Computational Intelligence (SSCI), 2016 IEEE Symposium Series on, IEEE, 2016, pp. 1–6.

[10] W. Druml, Systemic consequences of acute kidney injury, Curr. Opin. Crit. Care 20 (2014) 613–619.

[11] S. Faubel, P.B. Shah, Immediate consequences of acute kidney injury: the impact of traditional and nontraditional complications on mortality in acute kidney injury, Adv. Chronic Kidney Disease 23 (2016) 179–185.

[12] C.-C. Shiao, P.-C. Wu, T.-M. Huang, T.-S. Lai, W.-S. Yang, C.-H. Wu, C.-F. Lai, V.-C. Wu, T.-S. Chu, K.-D. Wu, National Taiwan University Hospital Study Group on Acute Renal Failure (NSARF) and the Taiwan Consortium for Acute Kidney Injury and Renal Diseases (CAKs), Long-term remote organ consequences following acute kidney injury, Critical Care (London, England) 19 (2015) 438.

[13] S. Tirunagari, S. Bull, N. Poh, Automatic classification of irregularly sampled time series with unequal lengths: a case study on estimated glomerular filtration rate, in: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), ISBN 978-1-5090-0746-2, IEEE.

[14] C.E. Rasmussen, H. Nickisch, Gaussian processes for machine learning (gpml) toolbox, J. Mach. Learn. Res. 11 (2010) 3011–3015.

[15] J.H. Friedman, Multivariate adaptive regression splines, Ann. Stat. (1991) 1–67.

[16] T.W. Yee, C. Wild, Vector generalized additive models, J. Roy. Stat. Soc. Ser. B (Methodol.) (1996) 481–493.

[17] M. Faddy, Follicle dynamics during ovarian ageing, Mol. Cell. Endocrinol. 163 (2000) 43–48.

[18] S. Fattorini, A simple method to fit geometric series and broken stick models in community ecology and island biogeography, Acta Oecologica 28 (2005) 199–205.

[19] H. Ritzema, et al., Drainage principles and applications., ed. 2, International Institute for Land Reclamation and Improvement (ILRI), 1994.

[20] J.D. Toms, M.L. Lesperance, Piecewise regression: a tool for identifying ecological thresholds, Ecology 84 (2003) 2034–2041.

[21] S. de Lusignan, H. Gallagher, T. Chan, N. Thomas, J. van Vlymen, M. Nation, N. Jain, A. Tahir, E. du Bois, I. Crinson, N. Hague, F. Reid, K. Harris, 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 (2009) 39.

[22] S. De Lusignana, H. Gallagher, S. Jones, T. Chan, J. Van Vlymen, A. Tahir, N. Thomas, N. Jain, O. Dmitrieva, I. Rafi, et al., Audit-based education lowers systolic blood pressure in chronic kidney disease: the quality improvement in CKD (QICKD) trial results, Kidney Int. 84 (2013) 609–620.

[23] Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group, Chapter 2: Definition, identification, and prediction of CKD progression, Kidney Int. Suppl. 3 (2013) 63–72.