

AUTOMATIC CLASSIFICATION OF IRREGULARLY SAMPLED TIME SERIES WITH UNEQUAL LENGTHS: A CASE STUDY ON ESTIMATED GLOMERULAR FILTRATION RATE

Santosh Tirunagari, Simon Bull, Norman Poh

Department of Computer Science, University of Surrey, Guildford, Surrey GU2 7XH.
 {santosh.tirunagari, s.c.bull, n.poh}@surrey.ac.uk

ABSTRACT

A patient’s estimated glomerular filtration rate (eGFR) can provide important information about disease progression and kidney function. Traditionally, an eGFR time series is interpreted by a human expert labelling it as stable or unstable. While this approach works for individual patients, the time consuming nature of it precludes the quick evaluation of risk in large numbers of patients. However, automating this process poses significant challenges as eGFR measurements are usually recorded at irregular intervals and the series of measurements differs in length between patients. Here we present a two-tier system to automatically classify an eGFR trend. First, we model the time series using Gaussian process regression (GPR) to fill in ‘gaps’ by resampling a fixed size vector of fifty time-dependent observations. Second, we classify the resampled eGFR time series using a K-NN/SVM classifier, and evaluate its performance via 5-fold cross validation. Using this approach we achieved an F-score of 0.90, compared to 0.96 for 5 human experts when scored amongst themselves.

Index Terms— eGFR, Gaussian Process Regression, K-NN, SVM, CKD, AKI.

1. INTRODUCTION

Estimated glomerular filtration rate (eGFR) is a derived measurement that characterises the effective functioning of a kidney. It plays a central role in both the management of people with chronic diseases and epidemiology research involving longitudinal data with ten or more years of observations [1]. Often eGFR time series exhibit irregularities such as missing values and unequal lengths. Missing values are an inevitable consequence of the difficulty of ensuring that patients return for regular follow up measurements [2], while unequal lengths are a result of patients with differing ages and conditions receiving measurements with different frequency (Figure 1). A patient’s eGFR is therefore observed at irregular

time intervals, and will have greater or fewer observations depending on their age and the conditions they suffer from.

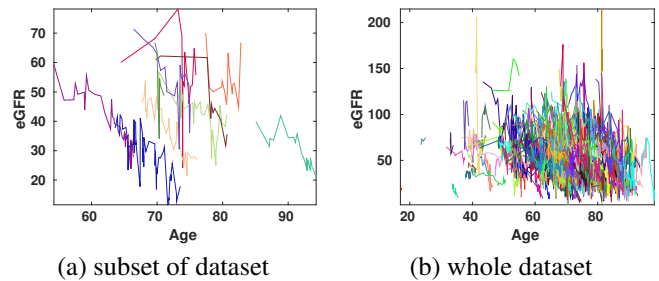


Fig. 1. A patient’s eGFR is observed at irregular time intervals and over different age ranges. Each colour represents a single patient.

For a clinician, having an easily understandable summary of a patient’s eGFR trend can be useful for determining the progression of diseases ranging from diabetes to chronic kidney disease. In many cases, simply distinguishing between stable (non-decreasing) and unstable (decreasing) trends can prove sufficient. Armed with this information a clinician can identify those patients who are most at risk of suffering a deterioration in their renal function. Presently, this trend differentiation is performed by a nephrologist manually analysing and labelling an eGFR time series. Despite manual labelling being time consuming and expensive, automating the process using standard supervised classifiers directly is not possible, as they require an equal number of input features and eGFR time series are of unequal length. Overcoming this requires either developing a framework for classifying the irregular and unequal eGFR time series, or using interpolation to make the time series amenable to standard classification methods.

Prior work in time series analysis has strongly emphasised regularly sampled equal length series, resulting in fewer methods that exist specifically for analysing irregularly sampled or unequal time series data. If the time series is to be analysed directly, then approaches such as spectral analysis [3, 4] and kernel-based methods [5] have been used to extract causal structure or statistics from data in fields such as as-

This work was supported by the Medical Research Council [grant number MR/M023281/1]. We would like to thank our collaborators at East Kent Hospital, UK for providing us with the dataset. The project details can be found at <http://www.modellingckd.org/>

tronomy [6, 7], palaeontology [5] and economics [8]. Despite this, the most common approach when working with irregular time series data is to transform it into regularly spaced data through some form of interpolation. For classification tasks, this has the added advantage of enabling the time series to be equalised to a given length by sampling from the interpolated function, thereby enabling standard classification algorithms to be used.

In this study, classification was performed after equalising eGFR series lengths using Gaussian processes, which are particularly useful for predicting patient outcomes [9], and have been used previously when classifying irregular time series data [10]. By enabling longitudinal data-interpolation and data-extrapolation, Gaussian processes can also directly assist with the problems of missing observations (irregular samples) and unequal series lengths respectively. Therefore, we employed Gaussian process regression (GPR) to equalise the lengths of the eGFR time series across all patients by producing the best linear unbiased predictions of interpolated and extrapolated observations. Following this equalisation, we used K-NN/SVM to classify an eGFR time series as having a stable or unstable trend.

The main objectives of this study are two-fold. First, to examine the feasibility of using machine learning based techniques, such as GPR, for transforming an unequal length irregularly sampled time series to an equal length, and then automatically classifying the equalised time series using a standard classifier. Second, to understand whether the trends classified by our approach correspond to those expected by clinicians. Our contributions can thus be summarised as follows: (i) Novel use of GPR+K-NN/SVM for classifying irregular eGFR time series with unequal lengths. Although GPR has been widely used, its uses in sampling irregular time series with unequal lengths are rarely highlighted or discussed. Our approach represents the first time that individual patient’s eGFR time series have had their trend automatically classified and visualised, and serves to demonstrate the applicability of GPR to such problems. (ii) Improved understanding for clinicians. Through comparison with experts’ trend classifications, we will demonstrate that our approach can serve as a tool for enhancing clinicians diagnostic capabilities.

The organisation of the paper is as follows: In section 2, we study the eGFR dataset. In section 3, we describe our methodological framework. Experiments and results are discussed in section 4. Finally, in section 5, we draw conclusions and discuss their clinical relevance.

2. THE ‘HANNAH’ DATASET

The dataset used in this work contains the eGFR time series from 488 patients treated at East Kent University Hospital, and was collected as part of a study seeking to understand the characteristics of acute kidney injury and its impact on chronic kidney disease. Each patients eGFR time series was

labelled as either stable, linear or step-change by five experts. However, for the purpose of this work, we grouped the unstable (linear and step-change) trends together and sought to distinguish only between stable and unstable time series. This is because clinicians are largely concerned with distinguishing between those patients with and without stable eGFR measurements, as this enables them to identify patients who are likely to need further monitoring (those with unstable measurements). Of the 488 patients, 260 (53.3%) have stable time series and 228 (46.7%) unstable, while 275 (56.4%) were male and 213 (43.6%) female. In total, there were 10,873 eGFR measurements across the 488 patients. Figure 2 summarises the main characteristics of the dataset. Approximately 95% of the patients are between the ages of 60 and 90, with eGFR values between 25 and 95 mL/min/1.73m².

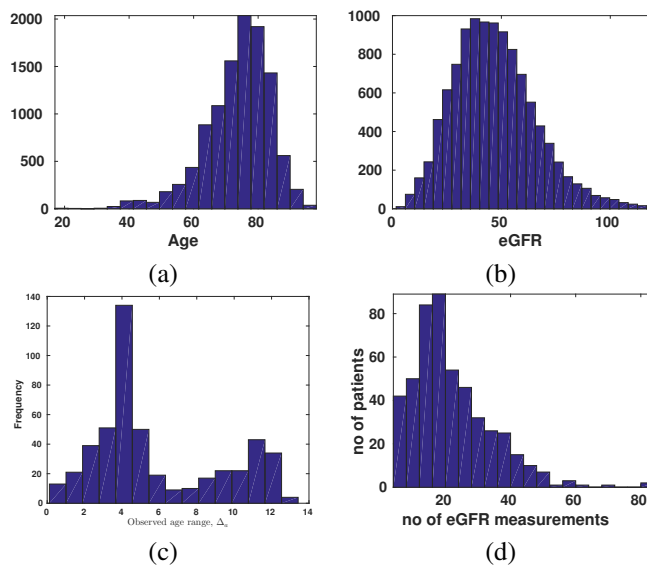


Fig. 2. Dataset characteristics. (a) Histogram showing the number of eGFR measurements recorded across all patients at each age. (b) Histogram showing the number of eGFR measurements over the range of observed eGFR values. (c) Histogram showing the distribution of the age ranges over which patients had eGFR recordings. (d) Histogram showing the number of eGFR measurements recorded per patient.

3. METHODOLOGY

The proposed automatic eGFR-trend classification pipeline consists of a regression step followed by a classification step, as shown in Figure 3. The unequal length eGFR series for each patient is modelled using GPR, which produces a fixed-size vector of 50 observations over the patient’s recorded age. This fixed-size vector is then given as input to a standard supervised classification algorithm in order to classify the eGFR series as stable or unstable.

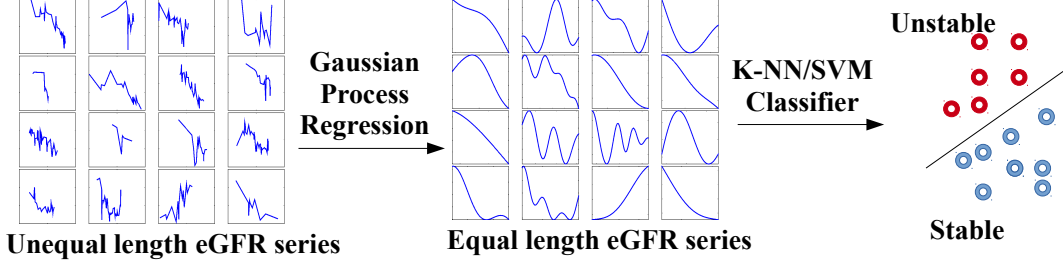


Fig. 3. Each of the eGFR series of 16 patients (left) is modelled using GPR in order to produce a fixed-size vector of 50 observations representing the entire observable time series. A collection of these fixed-size vectors (middle) are then given as input to a standard supervised classification algorithm, such as K-NN or SVM, to determine if an eGFR series is stable or unstable (right).

3.1. Gaussian Process Regression

A Gaussian process (GP) is a collection of random variables, any finite collection of which has a joint Gaussian distribution [11]. GPs characterise the probability distribution over functions by a specified mean function $\bar{d}(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$ [12].

To describe a real process $f(\mathbf{x})$ as a GP, we write: $f(\mathbf{x}) \sim \mathcal{GP}(\bar{d}(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. Here $\bar{d}(\mathbf{x}) = E\{d(\mathbf{x})\}$ and $k(\mathbf{x}, \mathbf{x}') = E\{(d(\mathbf{x}) - \bar{d}(\mathbf{x}))(d(\mathbf{x}') - \bar{d}(\mathbf{x}'))\}$, where $E\{g(\mathbf{x})\}$ denotes the expectation of a function g over the variable \mathbf{x} .

Given a set of measurements $\mathcal{D} = \{(\mathbf{x}_i, d(\mathbf{x}_i))\}_{i=1}^N$, the goal is to estimate the true output $d(\mathbf{x}^*)$ at an arbitrary \mathbf{x}^* given the relation:

$$d_{obs}^i = d(\mathbf{x}_i) + \epsilon(\mathbf{x}_i); \quad \epsilon(\mathbf{x}_i) \sim \mathcal{N}(0, \sigma_n^2) \quad (1)$$

The prior distribution of the observed target $d(\mathbf{x})$ is given by:

$$d(\mathbf{x}) \sim \mathcal{N}(\bar{d}(\mathbf{x}), k(\mathbf{X}, \mathbf{X}')) \quad (2)$$

where $k(\mathbf{X}, \mathbf{X}')$ is the covariance matrix between all pairs of training points. A squared exponential kernel was used to determine the covariance matrix, where the squared exponential kernel (Gaussian/RBF) is defined as: $\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\frac{(\mathbf{x}-\mathbf{x}')^2}{2\gamma^2})$, with γ the length scale of the kernel.

The distribution of the estimated mean value $\mathbf{d}(\mathbf{x})$ can be computed by conditioning on the training data to get $p(d(\mathbf{x})|\mathbf{x}^*, \mathcal{D})$. The joint distribution over $d(\mathbf{x})$ and the new datapoint \mathbf{x}^* is computed using:

$$\begin{bmatrix} \mathbf{d}_{obs} \\ d(\mathbf{x}^*) \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \bar{d}(\mathbf{X}) \\ \bar{d}(\mathbf{x}^*) \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I & K(\mathbf{X}, \mathbf{x}^*) \\ K(\mathbf{x}^*, \mathbf{X}) & k(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix} \right)$$

Here, $\mathbf{d}_{obs} = (d_{obs}^1, \dots, d_{obs}^N)^T$; $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, $\bar{d}(\mathbf{X})_i = \bar{d}(\mathbf{x}_i)$, and $K(\mathbf{X}, \mathbf{X})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

The conditional distribution of Equation 2 allows us to get the distribution of $d(\mathbf{x}^*)$ with the following mean and covariance:

$$d(\mathbf{x}^*) \sim \mathcal{N}(E\{d(\mathbf{x}^*)\}, var\{d(\mathbf{x}^*)\}) \quad (3)$$

where

$$\begin{aligned} E\{d(\mathbf{x}^*)\} &= \underbrace{\bar{d}(\mathbf{x}^*)}_{\text{prior}} + K(\mathbf{x}^*, \mathbf{X}) \left[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I \right]^{-1} (\mathbf{d}_{obs} - \bar{d}(\mathbf{X})) \\ var\{d(\mathbf{x}^*)\} &= \underbrace{k(\mathbf{x}^*, \mathbf{x}^*)}_{\text{prior}} - K(\mathbf{x}^*, \mathbf{X}) \left[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I \right]^{-1} K(\mathbf{X}, \mathbf{x}^*) \end{aligned}$$

The GPstuff toolbox¹ [13], was used for modelling the eGFR time series, with the hyperparameters for the squared exponential kernel tuned using maximum a posteriori estimates.

3.2. Classification Details & Performance Evaluation

The fixed size eGFR series can be used as the features for any standard classification algorithm. Here we have used K-NNs and SVMs for this purpose, as they are known for their high classification accuracy. The Euclidean distance measure was used for the K-NN classifier, with $K = 3$. For the SVMs, the radial basis function kernel was used with $\sigma = 10$. Performance was evaluated using 5-fold cross validation, with the performance for each fold evaluated using the F-score.

Let \mathcal{Y} be the classifier's prediction and ω the groundtruth. Then the stable/unstable classification of an eGFR series is made on the following basis:

$$\text{decision}(\mathcal{Y}) = \begin{cases} \text{stable} & \text{if } y = 1 \\ \text{unstable} & \text{otherwise,} \end{cases} \quad (4)$$

This can result in the following outcomes: (i) a true positive (TP $\equiv y = 1, \omega = 1$); (ii) a false positive (FP $\equiv y = 1, \omega = 0$); (iii) a true negative (TN $\equiv y = 0, \omega = 0$); and (iv) a false negative (FN $\equiv y = 0, \omega = 1$). Using the rates of these four outcomes, recall and precision can be derived. Recall describes the completeness of the classification, and precision the actual accuracy of the classification. They are defined with recall = $\frac{TP}{TP+FN}$ and precision = $\frac{TP}{TP+FP}$. While recall and precision can be individually used to determine the quality of a classifier, it is often more convenient to have a

¹<http://research.cs.aalto.fi/pml/software/gpstuff/>

Table 1. The F-score of each expert of the five experts and their mean

	E1	E2	E3	E4	E5	Mean
F-score	0.8417	1	1	1	0.9938	0.9671

single measure. The F-score achieves this by combining the recall and precision in a single equation:

$$F = 2 * \frac{precision * recall}{precision + recall}$$

3.3. Groundtruth

For each eGFR time-series, five nephrologists annotated with one of three labels: stable, linear or a step-change. Here linear and step-change are both considered to be unstable trends. The class label assigned to an eGFR time-series is based on the consensus of these five annotations. In order to evaluate the performance of an individual expert, we compared their labels to those of the remaining experts, and from this derived the F-score for the given expert. The performance of the five experts can be seen in Table 1.

4. EXPERIMENTS & RESULTS

Our experimental procedures and objectives were as follows:

- **Classification using derived statistics:** In order to provide a baseline for evaluating the performance benefits of equalising the lengths of the eGFR time series, we investigated whether it was possible to classify eGFR trends using only basic statistics derived from the time series.
- **Classification using equalised eGFR time series:** The effects of equalising the lengths of the eGFR time series were evaluated using two approaches. First, we equalised each patient’s time series while taking into account age, by fitting a GPR model and uniformly resampling 50 datapoints from the expected trend between the ages of 30 and 90. Next we used the fitted GPR model to uniformly resample 50 datapoints for each patient, with the sampling restricted to the age range in which the patient had eGFR measurements recorded. Although this ensures that the new datapoints are only generated from the age range with low variance, it may potentially cause the sampling rate to be different for each patient.
- **Classification using derived statistics and equalised eGFR time series:** Classification using only datapoints resampled from GPR can be considered to be an instance of pattern matching, as trends are classified purely based on similarities between GPR models. Therefore,

we investigate whether reintroducing information from derived statistics, such as the age range over which measurements were taken, is beneficial.

- **Comparison with linear interpolation:** Finally, we compare the GPR-based classification method with one where the 50 datapoints are produced by linear interpolation. This experiment examines whether the smoothness of the fitted curve provided by GPR warrants performance generalisation or not.

4.1. Derived Statistics

For each patient, we derived four statistics from their series of eGFR measurements: the age range over which eGFR measurements were made (Δ_a), the range of the eGFR measurement values (Δ_g), the mean age at which a patient had their eGFR measured (μ_a) and their mean eGFR value (μ_g). From Figure 4(a) we can see that μ_a is negatively correlated with μ_g , showing that as patients get older their eGFR tends to decline. In addition, patients with an unstable trend tend to have a lower μ_g value. Patients with a greater μ_g also tend to have their eGFR measured over a longer age range (Figure 4(b)), possibly because a greater mean eGFR value means that the patient is less likely to have suffered from complications that cause their eGFR to stop being measured. As patients with unstable eGFR trends tend to have both a greater μ_a and lower μ_g value than patients with stable trends, the derived statistics are likely to provide some discriminative power to distinguish between the two trends. In order to ascertain the level of this discriminative power, we fed the statistics to K-NN and SVM for classification. We found that this baseline approach gives an average F-score of 0.71 based on 5-fold cross validation (Table 2).

4.2. eGFR Time Series Equalisation

Using the GPR model learned from each patient’s eGFR time series, we resampled 50 uniformly spaced datapoints between the ages of 30 and 90 for each patient. This ensures that not only are all patient’s resampled time series of equal length, but also that the interval between measurements is the same. However, while the model shows relatively low variance in the age range where a patient has measurements, the variance increases markedly outside this range. For example, the patient in Figure 5 had their eGFR recorded between the ages of 55 and 75, and consequently the GPR model fit using their measurements shows lower variance within this age range. The performance of the K-NN classifier and SVM trained using the resampled datapoints was 0.57 and 0.43 respectively; lower than that achieved using only derived statistics (Table 2).

For each patient, we also used the GPR model learned from their eGFR measurements to resample 50 uniformly spaced datapoints within the age range over which they have eGFR

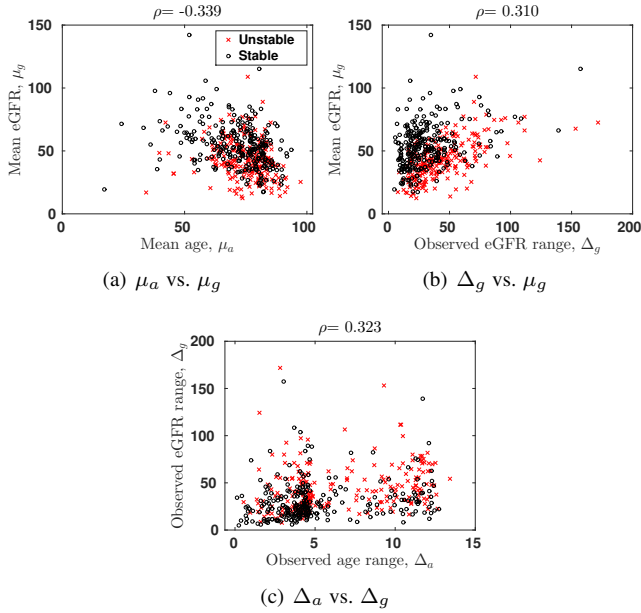


Fig. 4. A scatter plot of different derived statistics.

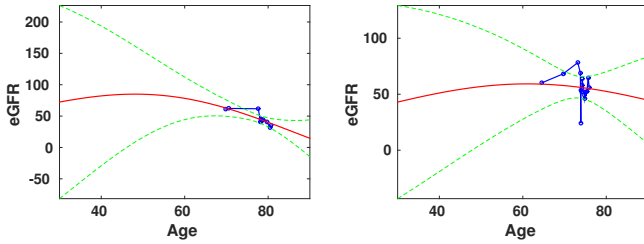


Fig. 5. eGFR resampled between ages 30 and 90 for unstable and stable trends. An increase in variance is seen when eGFR readings are not present.

measurements recorded. The fitted mean curve, along with the 95% variance interval, for four patients can be seen in Figure 6. When compared to the those resampled over the entire 30 to 90 year age range, the resampled curves within the age range can be expected to have lower variance. This decrease in variance is likely responsible for the substantial improvement in performance achieved by both the K-NN classifier and SVM when compared to their performance using data resampled between ages 30 and 90 (Table 2).

4.3. Combining Statistics & Equalisation

In the previous GPR experiments, only the resampled eGFR values are taken into account, thereby equating the problem of trend classification to that of pattern matching, since the notion of time is not considered. As we believe that age information is likely to be important in classifying the trends, we reintroduce this information by creating a feature vector for each patient that consists of the four derived statistics and the

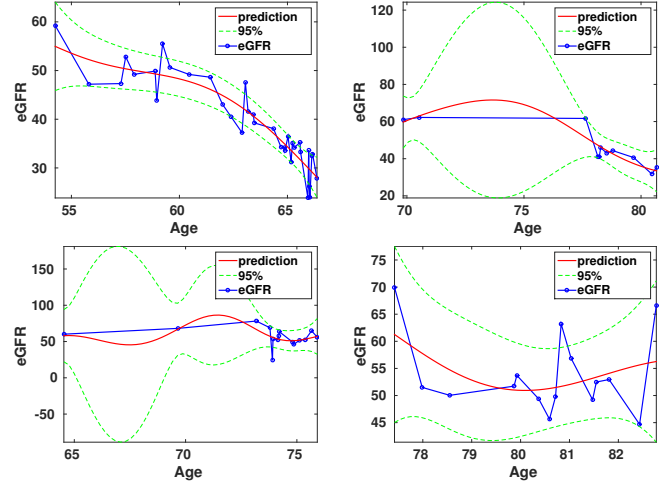


Fig. 6. Unequal length eGFR series (blue) modelled through GPR to produce a fixed-size vector of 50 observations (red) over the patients recorded age. (Top) figure shows the unstable trend and (bottom) figure shows the stable trend.

50 uniformly resampled datapoints within the age range over which the patient has eGFR measurements. By reintroducing the derived statistics, the F-score increased from 0.87 to 0.90 for the K-NN classifier, and from 0.86 to 0.89 for the SVM (Table 2).

4.4. Comparison with Interpolation

Rather than using GPR for equalising the eGFR lengths, we could have used linear interpolation. This allows the structure and trend of the eGFR series to be preserved, while still producing fixed size vectors (as seen in Figure 7). A K-NN classifier and SVM trained on data generated in this manner had an F-score of 0.84 and 0.87 respectively. When coupled with the derived statistics, the F-score of the K-NN classifier improved to 0.86, while the F-score of the SVM remained the same. However, results for both classifiers were lower than for equalisation performed using GPR, possibly due to the smoother fitted curve produced by GPR.

5. CONCLUSION & DISCUSSION

Due to the significance of kidney function in many chronic conditions, it is important for clinicians to be able to quickly and accurately screen patients' eGFR time series to identify those whose kidney function is at risk of deteriorating. Given that many patients with long term conditions are managed in their general practices, the screening process would ideally be automated and able to remotely monitor patients. Despite this, no automated methods exist that solve this problem. Our work is therefore the first reported attempt to design an automated process and compare it with human experts.

Table 2. Classifier F-score

	30-90 GPR		Statistics		GPR		Statistics + GPR		Interpolation		Statistics + Interpolation	
	K-NN	SVM	K-NN	SVM	K-NN	SVM	K-NN	SVM	K-NN	SVM	K-NN	SVM
Fold-1	0.5952	0.4450	0.7526	0.7465	0.8950	0.8643	0.9258	0.8443	0.8762	0.8858	0.8762	0.8760
Fold-2	0.5918	0.4368	0.6927	0.6632	0.8451	0.8450	0.9446	0.9151	0.7343	0.8162	0.8670	0.8776
Fold-3	0.5143	0.4313	0.7427	0.8151	0.9077	0.8732	0.9178	0.9065	0.8976	0.9183	0.8776	0.8876
Fold-4	0.5970	0.4121	0.7142	0.7245	0.8850	0.8757	0.8542	0.8958	0.8164	0.8234	0.8024	0.8182
Fold-5	0.5633	0.4343	0.6561	0.6390	0.8955	0.8655	0.8645	0.8658	0.8792	0.9046	0.8622	0.8752
Average	0.5723	0.4319	0.7116	0.7176	0.8856	0.8647	0.9013	0.8855	0.8407	0.8696	0.8570	0.8669

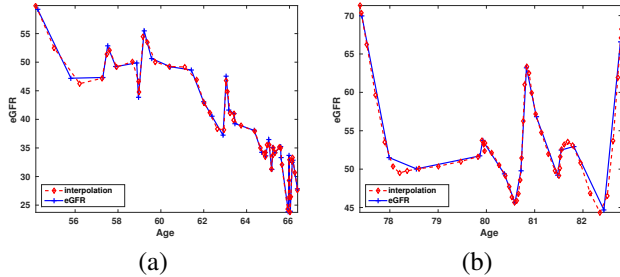


Fig. 7. Unequal length eGFR series (blue) modelled using linear interpolation to produce a fixed-size vector of 50 observations (red) over the range for which a patient has eGFR measurements. The figure shows both (a) unstable and (b) stable trends.

In spite of the complexity of eGFR trends, we found that it was possible to utilise a supervised machine learning approach to automatically determine if an eGFR trend is stable or unstable. We found that by equalising the lengths of the eGFR time series for each patient by uniformly resampling new datapoints, our approach could classify the eGFR trend with an F-score approaching that of human experts (0.90 compared to 0.97 for the experts). The best approach for performing this equalisation is to fit a GPR model, and then resample new datapoints from within the age range of the patient where eGFR measurements are observable. This approach performs better than resampling new datapoints from the fitted GPR across the same age range (30 to 90 years) for all patients, and better than using linear interpolation to resample the new datapoints. This shows that smoothness of the fitted curve plays an important in performance generalisation. Finally, the inclusion of statistics derived from the original eGFR time series of a patient further improves the classification performance for all classifiers except the SVM trained on interpolated data.

6. REFERENCES

- [1] Norman Poh and Simon 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)*, 2012.
- [2] Norman Poh, Santosh Tirunagari, and David Windridge, “Challenges in designing an online healthcare platform for personalised patient analytics,” in *Computational Intelligence in Big Data (CIBD), 2014 IEEE Symposium on*. IEEE, 2014, pp. 1–6.
- [3] Michael Schulz and Karl Stattegger, “Spectrum: Spectral analysis of unevenly spaced paleoclimatic time series,” *Computers & Geosciences*, vol. 23, no. 9, pp. 929–945, 1997.
- [4] Petre Stoica, Prabhu Babu, and Jian Li, “New method of sparse parameter estimation in separable models and its use for spectral analysis of irregularly sampled data,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 1, pp. 35–47, 2011.
- [5] Kira Rehfeld, Norbert Marwan, Jobst Heitzig, and Jürgen Kurths, “Comparison of correlation analysis techniques for irregularly sampled time series,” *Nonlinear Processes in Geophysics*, vol. 18, no. 3, pp. 389–404, 2011.
- [6] Piet Broersen, “Time series models for spectral analysis of irregular data far beyond the mean data rate,” *Measurement Science and Technology*, vol. 19, no. 1, pp. 14, 2008.
- [7] C. Thiebaut and S. Roques, “Time-scale and time-frequency analyses of irregularly sampled astronomical time series,” *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 2486–2499, 2005.
- [8] Ulrich Müller, “Specially weighted moving averages with repeated application of the ema operator,” 2000.
- [9] David Windridge and Miroslaw Bober, “A kernel-based framework for medical big-data analytics,” in *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, pp. 197–208. Springer, 2014.
- [10] Steven Cheng-Xian Li and Benjmain Marlin, “Classification of sparse and irregularly sampled time series with mixtures of expected gaussian kernels and random features,” in *31st Conference on Uncertainty in Artificial Intelligence*, 2015.
- [11] Christopher KI Williams and Carl Edward Rasmussen, “Gaussian processes for machine learning,” *the MIT Press*, vol. 2, no. 3, pp. 4, 2006.
- [12] Carl Edward Rasmussen and Hannes Nickisch, “Gaussian processes for machine learning (gpml) toolbox,” *The Journal of Machine Learning Research*, vol. 11, pp. 3011–3015, 2010.

- [13] Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari, “Gpstuff: Bayesian modeling with gaussian processes,” *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1175–1179, 2013.