A Probabilistic Broken-stick Model For CKD Staging And Risk Stratification

Norman Poh[1], Santosh Tirunagari[1], Simon C Bull[1], Nicholas Cole[2] and Simon de Lusignan[2]

1: Department of Computer Science, University of Surrey

2: Department of Clinical and Experimental Medicine, University of Surrey

Contact: N.Poh@surrey.ac.uk

Introduction

Trends in clinical measurements can provide insight into the expected development of 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. However, modelling long-term trends in biomedical measurements can be complicated by both practical and biomedical considerations. For example, eGFR can be influenced by, amongst other things, the level of protein in the diet, changes in muscle breakdown and the level of hydration.

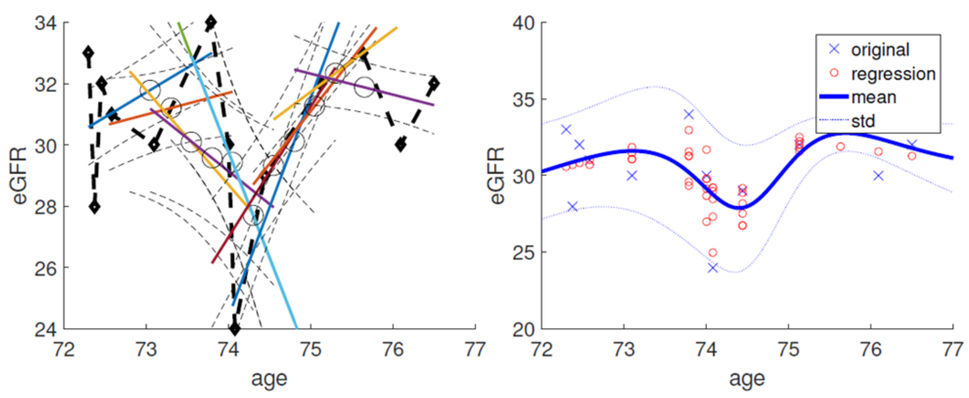

Figure 1: The line segments from a broken-stick model (left) and the raw eGFR time-series from which they were calculated.

The irregular taking of measurements presents an additional difficulty when attempting to model long-term trends, and is solved predominantly by transforming the data in order to enforce regularity. However this forces assumptions on the data, e.g. that the irregularity is random in nature, and can lead to uninterpretable models. In order to strike a balance between model quality and interpretability, broken-stick regression, also known as segmented or piece-wise regression, can be used. However, this introduces discontinuities at the segment boundaries (FIGURE 1). To address this, we take a Bayesian approach to enforce smooth transitions between segments, providing interpretable models over long- and short-term trends.

Windowing

The first step in fitting the broken-stick model is to divide a time-series into windows (FIGURE 2). Here equal window lengths were used, although this is not a requirement. Ideally the window length is set so that each window encompasses a sufficient number of datapoints to capture the local trend, without being so large that it suppresses global trends.

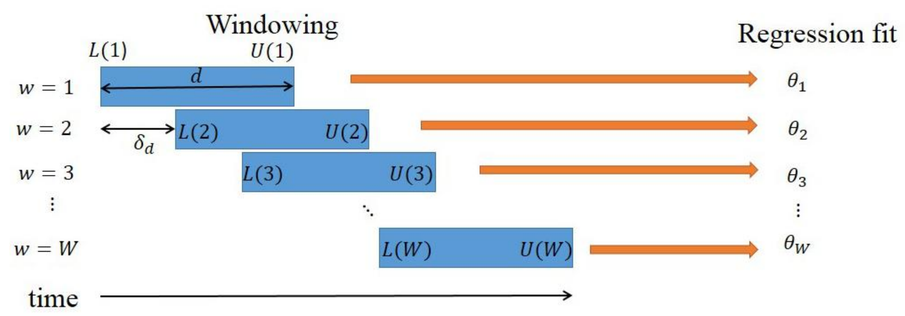

Figure 2: A time series is broken into W windows of length d. For each window, a linear regression model is fit.

In order to smoothly join the fitted line segment we take a Bayesian approach. Ideally the further away in time a window is from a point in time t the less influence its line segment has near t. To achieve this, let $P(w|t)$ be the posterior probability of the w-th window at time t and $p(t|w)$ the distribution of time points across windows. Then we can define $P(w|t)$ such that $P(w|t) \propto p(t|w)$ and the window function $p(t|w)$ is Gaussian (FIGURE 3). We can then use Bayes' theorem to obtain $P(w|t)$:

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

Each window's influence is therefore weighted to diminish as the distance from its midpoint in time increases.

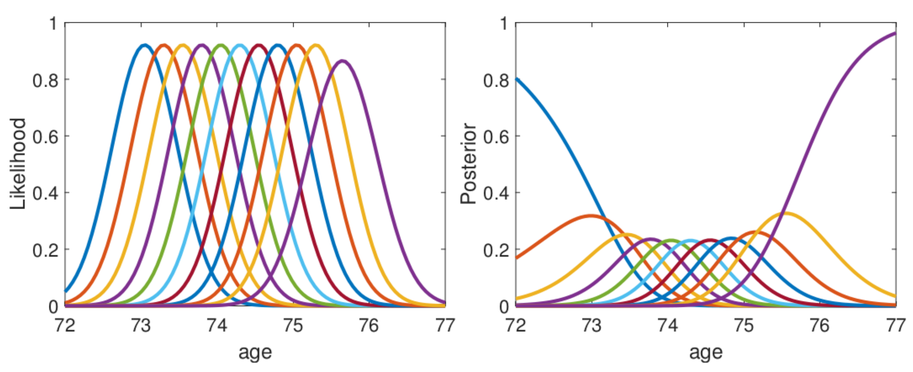

Figure 3: An application of windowing with windows of length two years and a six month gap between windows. Each of the eleven windows is plotted in a separate colour. As $\sum_w P(w|t_*) = 1$ the first and last windows dominate the posterior near the boundaries.

Dataset and Chosen Window

The proposed method was validated using the QICKD [1] dataset. This contains the primary care records of 951,764 patients, of whom 12,297 contain an eGFR measurement. Based on this data, a window length of three years and gap between windows of half a year were chosen. Patients were excluded based on the following criteria:

- 1,546 for having less than three years of measurements.
- 26 for having gaps between measurements of more than three years.
- 10 for having at least one stick with an overly large gradient.
- 1,103 for having experienced an AKI episode (determined using the SAKIDA algorithm [2]).

In total 2,603 patients were excluded, leaving 9,694. Finally, due to the increased variance of larger eGFR measurements, any values above 120 mL/min/1.73m^2 were removed.

Determining CKD Stages and Stratifying Risk

Rather than determine CKD stage using raw eGFR values, we use the estimated mean eGFR value obtained directly from the broken-stick model and compare it to the CKD stages determined using the KDIGO guidelines [3]. To determine whether the staging methods are consistent, we calculated the probability of the stage determined using the estimated mean $CKD(\mu)$ given the raw eGFR value g, i.e. $P(CKD(\mu)|g)$. The probability distribution calculate din this manner for each stage, as determined using the estimated mean, can be seen in FIGURE 4. It is noticeable that the stages determined using the estimated mean largely coincide with those using the KDIGO boundaries, except for a large gap for the boundary between stages 1 and 2.

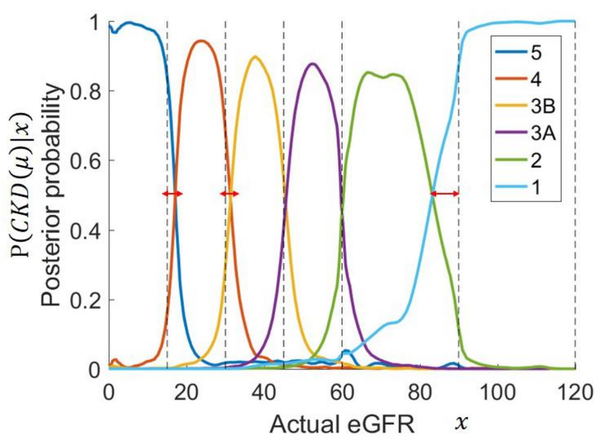

Figure 4: Expected CKD stage probability distributions given raw eGFR values.

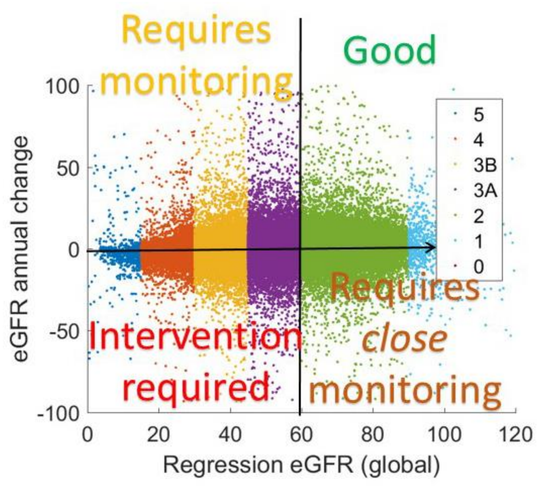

Figure 5: Stratification of patients based on their estimated CKD stage and eGFR slope.

It is possible to use the estimated CKD stage and the eGFR slope calculated from the model to stratify patients according to the trajectory their condition is taking. In FIGURE 5 patients are stratified according to stage and slope into four rough categories based on their outlook.

Discussion

The broken-stick model [4] can estimate short-term and long-term trends from irregularly sampled clinical data, and help understand disease progression by smoothing out local fluctuations. Applied to eGFR measurements and CKD the model presents two distinct use cases:

1. Determining CKD stage. With the broken-stick model a patient's CKD stage can be determined using their entire history, not just local measurements.
2. Determining CKD progression. The general consistency between stages determined using the broken-stick model and the KDIGO guidelines indicates that stratification using the model is likely to prove reliable as it relies on the same model used for the staging.

Taken together, these results could provide useful information when determining the trajectory of a patient's condition and in the retrospective identification of patients for clinical research.

A technical presentation of the broken-stick model is reported in [4].

References

[1] S. de Lusignan et al. 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, Implementation Science 4 (2009) 39.

[2] S. Tirunagari et al. Automatic detection of acute kidney injury episodes from primary care data (2016).

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

[4] Poh, Norman, et al. "Probabilistic Broken-Stick Model: A Regression Algorithm for Irregularly Sampled Data with Application to eGFR." *arXiv preprint arXiv:1612.01409* (2016).

UNIVERSITY OF SURREY

MRC Medical Research Council

www.modellingCKD.org

Acknowledgement: This project entitled "Modelling the Progression of CKD" was awarded to NP by the Medical Research Council, under the New Investigator Grant Scheme MR/M023281/1.