Automatic Extraction of Deep Phenotypes for Precision Medicine in Chronic Kidney Disease*

Prerna Singh
Biomedical Engineering
Johns Hopkins University
USA
psingh26@jhu.edu

Varun Chandola
Computer Science and Engg
University at Buffalo, SUNY
USA
chandola@buffalo.edu

Chester Fox
Family Medicine
University at Buffalo, SUNY
USA
cfox@buffalo.edu

ABSTRACT

Chronic Kidney Disease (CKD) is one of the deadliest diseases in the world, with 10% of the global population affected by the disease. Identifying subpopulations with characteristic disease progressions is important to find more efficient treatments for patients with this disease. The abundance of electronic health records (EHR) data can be used to find meaningful subtypes for CKD but comes with challenges during analysis, including irregular data sampling, and skewness in the data collected over time. In this paper, multiple regression techniques were used to fill in the missing estimated glomerular filtration rate (or eGFR – a key measure for kidney function) trajectory data, so it can be clustered effectively. Clustering is applied to the enhanced data to obtain six subtypes, which capture crucial trends in the disease progression of patients. Moreover, the characteristics of patients in each of the subtypes had minor differences from others. These characteristics demonstrate risk factors and positive lifestyles choices of patients with CKD, which can help develop new treatments for CKD.

CCS CONCEPTS

•**Computing Methodologies** → Machine Learning
•**Applied Computing** → Life and medical sciences

KEYWORDS: Time-series Clustering, Partitioning around Medoids, Regression, Spline.

ACM Reference format:

P. Singh, V. Chandola, and C. Fox, 2017. Automatic Extraction of Deep Phenotypes for Precision Medicine in Chronic Kidney Disease. In *Proceedings of 7th International Conference on Digital Health, London, UK,* July 2017, 5 pages. DOI: http://dx.doi.org/10.1145/3079452.3079489

*This work was done while Prerna Singh was a senior at Williamsville East High School.

DH'17, July 02-05, 2017, London, United Kingdom

ACM ISBN 978-1-4503-5249-9/17/07…$15.00
http://dx.doi.org/10.1145/3079452.3079489

1 INTRODUCTION

Clinical Electronic Health Record (EHR) data is rapidly growing and has the potential to be a cost-effective and large-scale source for extracting deep phenotypes. However, the extraction of detailed disease and drug-related phenotype information hidden in clinical data is a challenging task.

One large collection of EHR data is the DARTNet *Chronic Kidney Disease* (CKD) dataset [7]. CKD is well recognized as a rising problem in global health. According to the 2013 Global Burden of Disease study, there were approximately 956,000 deaths caused by CKD worldwide in 2013 [4]. Furthermore, CKD was ranked 19th in the top causes of global years of life lost in 2013. In the US, it is the 9th leading cause of death and impacts over 20% of the US adult population [2].

In this paper, we describe a methodology to identify precise disease subtypes for CKD, by clustering EHR data obtained from the DARTNet collection. Clustering EHR data is challenging due to the irregularity of lab tests and clinical observations. Time-series clustering methods, however, require consistent time intervals between data points for all patients. We employ a statistical spline regression method to convert the irregularly sampled test results into a uniformly sampled time series. The imputed time series data is then used for clustering to identify the disease subtypes in the target population. We obtain six subtypes, which are then analyzed in terms of demographic characteristics, patient lifestyle, and co-existing conditions. While, there have been past studies that have used EHR for understanding CKD progression, the temporal aspects of eGFRs, a key indicator of CKD severity and CKD subtypes, has not been addressed previously [2,5].

2 DATA

The data comes from a collaboration of nine research networks, the DARTNet Institute. It stores data for approximately 12.5 million patient visits per year and over 5 million patient lives, leading to approximately five billion data values [1,3]. The data is used to track patients over several years as a time series in terms of disease severity, physiological characteristics, and medications.

One curated dataset that has been extracted from the larger data set, consists of 69,817 patients suffering from Chronic Kidney

Disease is used in this study. Table 1 summarizes the data elements in this set.

Table 1: Data elements in the CKD dataset.

Patient Information	Age, Gender, Race, Ethnicity
Test Results	Alanine Aminotransferase (ALT) Hemoglobin HDL Level LDL Level Triglycerides Aspartate Amino-transferase (AST) albumin/creatinine ratio HbA1c 25 OH Vitamin D Serum Phosphorus intact PTH eGFR Creatinine
Clinical Information	All medications All diagnoses Total physician visits
Physiological Measurements	Blood pressure Weight Weight/BMI

The target variable for this study was the estimated Glomerular Filtration Rate (eGFR), which is a standard derived test value that measures a patient's kidney function. The eGFR value is estimated from a clinical laboratory test that measures the *creatinine* level through a standard blood test. The eGFR rate estimates how much blood passes per minute through the glomeruli, which filter waste from the blood. It is standardized based on measurements of age, race, and gender through the following equation [7]:

$$eGFR = 186 * S_{cr}^{-1.154} * age^{-0.203} * \kappa_g * \lambda_r$$

where S_{cr} is the serum creatinine level obtained through the lab test. The constant κ_g is used to factor in gender information: κ_g = 0.742 for females and κ_g = 1 for males. The constant λ_r is used to account for ethnicity information: λ_r = 1.212 for African Americans and 1, otherwise.

According to the National Kidney Foundation, the eGFR value for healthy individual's ranges between 90-120. Patients suffering from Chronic Kidney Disease typically have an eGFR value below 60 for over 3 months.

Table 2: Criteria for valid measurement.

Value	Valid Threshold
Weight	< 500 pounds
Height	< 99 inches
Blood Pressure	10 - 300 mm Hg
eGFR	< 300 mL/min/1.73 m2
AST	< 1000 IU/L
Triglyceride	< 3165 mg/dl
Creatinine	< 100 mg/dL
ALT	< 1000 IU/L
Blood Glucose	< 2656 mg/dl

3 METHODS

The first step was to validate and clean the data. Table 2 displays the exact criteria for the results to be considered valid. All values that were invalid were discarded before proceeding with the analysis.

Three lab tests were used to predict the standardized GFR value for a patient with said test done on the same day.

- Creatinine
- GFR lab result for African Americans, and
- GFR lab result

If there was a creatinine test value and a standardized GFR value, available for the same day, the relationship between the points was used to create an inverse relationship between creatinine and GFR, which was then used to predict GFR when only a creatinine value was present. Based on 437,719 pairs of creatinine and standardized GFR tests done on the same day, a relationship between 1/Creatinine and GFR was identified. The R^2 value is the coefficient of determination and a measure of how well the regression line approximates the data (See Table 3). Figure 1 illustrates a patient's GFR values over time.

For applying the time-series clustering method, the observations for all patients must start and end on the same days and all of the points used for clustering must be an equal distance (time) apart. Based on the frequency of eGFR values over time available in the CKD data set, data was available for most patients between March 14, 2005 and March 7th, 2012 (nearly 7 years). If there was no value before/on March 14, 2005 or after/on March 7th, 2012, a linear regression model was applied to extrapolate. The spline regression model was used to "fill" the values for days within consecutive observations to obtain data at 30-day sampling rate, starting from March, 2005.

Table 3: Coefficient of determination for the linear regression of each variable, where y = Standardized GFR and x = variable (test result value)

Variable	Coefficient of Determination	Equation
Inverse Creatinine	0.67522	y = 56.8461x + 5.4281
African American GFR	0.41475	y = 1.0309x + 0.2113
GFR	0.68323	y = 0.7714x + 12.034

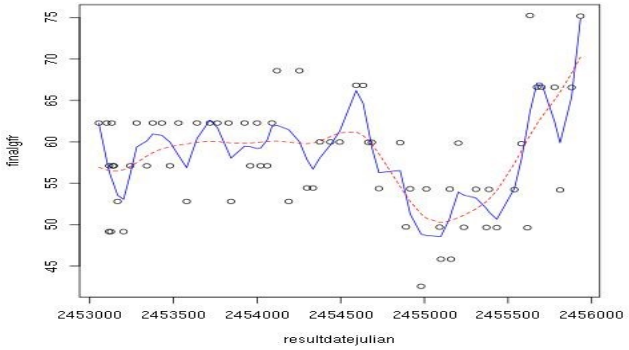

Figure 1: A sample patient's GFR values over time (x-axis as Julian time and y-axis as finalgfr). The dots signify data points; blue and red lines denote spline interpolations with different smoothing characteristics.

Next, the TSclust method in R was applied to create a dissimilarity matrix of the disease severity (eGFR). The Partitioning Around Medoids (PAM) [6] algorithm was used for clustering. The PAM algorithm partitions the dataset of *n* objects into *k* clusters, by minimizing the distance between points assigned to a cluster and a point evaluated as the center of the cluster (medoids] by creating a dissimilarity matrix.

4 RESULTS AND DISCUSSION

Six distinct subtypes were discovered from the Chronic Kidney Disease data set. Figure 2 shows the result of the six subtypes' prototype trajectories found in the experiment. In fact, the found subtypes' prototype trajectories coincide with existing knowledge about patient subgroups in CKD:

- Subtype #1 indicates a group of patients whose kidney function was stable, at a moderate level over the course of 5 years with an early fluctuation downward, then upwards, and later kidney function stabilizes.
- Subtype #2 represents a set of patients who have a slow decline in renal capability.
- Subtype #3 indicates a set of patients who have an increase in kidney function early on, but then kidney function stagnated at a lower level.
- Subtype #4 corresponds to the group of patients that yielded a constant improvement in kidney function.
- Subtype #5 indicates a set of patients whose kidney function originally decreased, but over time increased significantly.
- Subtype #6 coincides with a group of patients who had an increase in kidney function in early stages, but had severe damage in kidney function over time.

The characteristics associated with these six subtypes were further analyzed. The aim of this analysis was to find characteristics such as ALT, AST, HDL, Creatinine, LDL, triglyceride, and gender, and phenotypes associated with each subtype.

Subtype 1: Alanine Aminotransferase (ALT) and Aspartate Aminotransferase (AST) values are average in comparison to the other subtypes, but there were outliers with ALT and AST values of more than 50. The HDL value was average in this subtype. The LDL value was average when compared to other subtypes. The triglyceride values in this subtype were similar to those in other subtypes, but there were a significant number of outlier patients with extremely high triglyceride values as well.

Subtype 2: ALT and AST values were average in comparison to the other subtypes, but there were a significant number of outliers with ALT and AST values of more than 50. The HDL value was relatively higher in subtype 2 when compared to the other subtypes. The creatinine values in this subtype were comparable to the values in the other subtypes, and the triglyceride values in this subtype were similar to those in other subtypes. This subtype was 97% female and 3% male.

Subtype 3: ALT and AST values were slightly lower in comparison to the other subtypes, but there were also outliers with ALT and AST values of more than 50. The HDL values were relatively lower in subtype 3 when compared to the other subtypes. The creatinine values in this subtype were comparable to the values in the other subtypes. The LDL value was average when compared to other subtypes. The triglyceride values in this subtype were significantly higher than triglyceride values in other subtypes. This subtype was 27% male and 73% female.

Subtype 4: ALT and AST values were significantly higher in comparison to the other subtypes. The HDL value was average when compared to other subtypes. The creatinine values in this subtype were generally higher than those of other subtypes, and there were a significant number of outliers with even higher creatinine values in this subtype. The LDL value was average when compared to other subtypes. The triglyceride values in this subtype were similar to those in other subtypes. The male to female ratio in this subtype was about 50-50.

Subtype 5: The HDL value was average in this subtype. The creatinine values in this subtype were comparable to the values in the other subtypes. The LDL value was average when compared to other subtypes, and the triglyceride values in this subtype were similar to those in other subtypes. This subtype was 100% female.

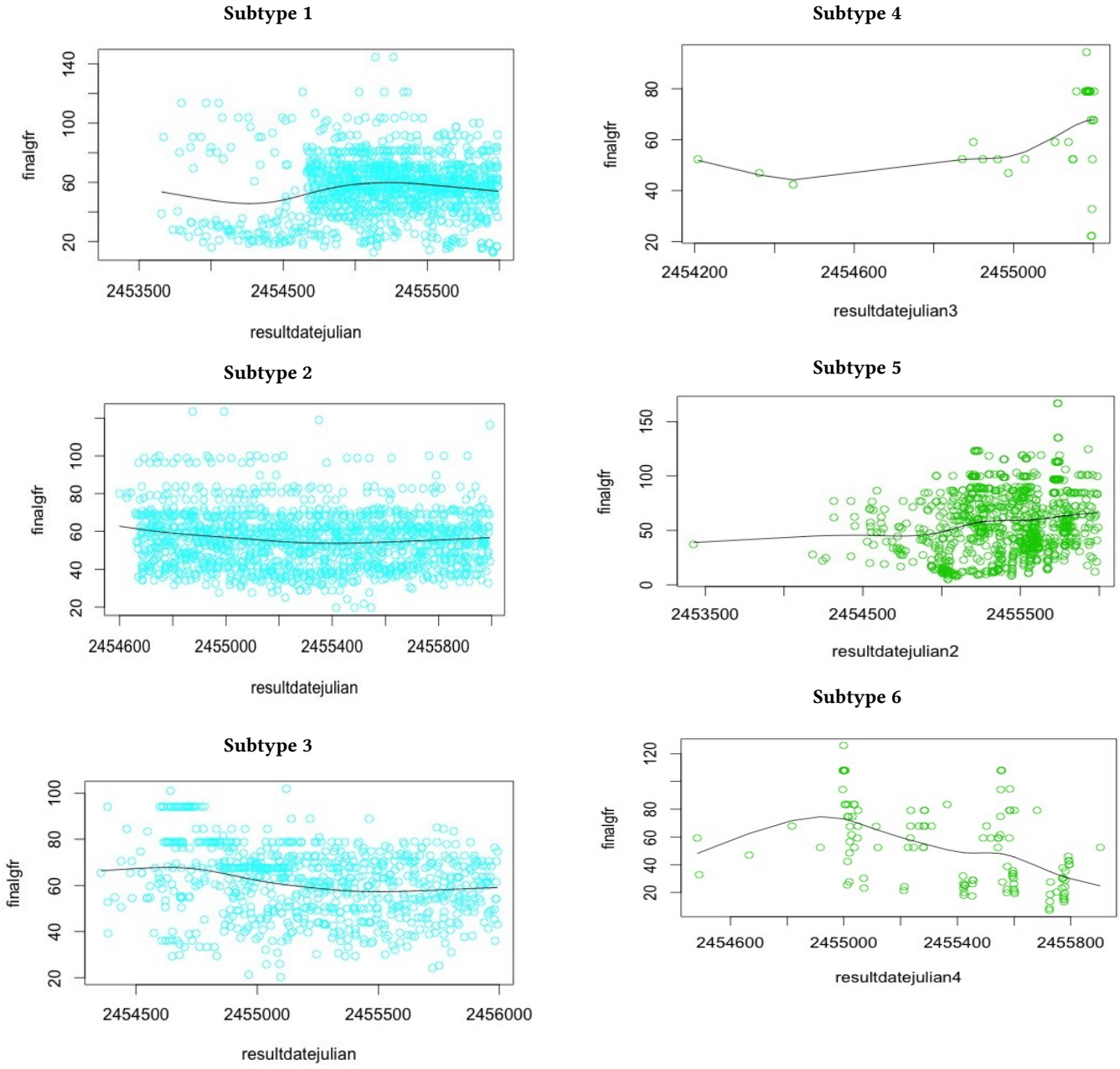

Figure 2: The graphs show disease severity trajectory for each derived subtypes (x-axis is Julian time and y-axis is finalgfr)

Subtype 6: The HDL value was significantly lower in subtype 2 when compared to the other subtypes. The creatinine values in this subtype were also generally much higher than those of other subtypes. The LDL value was significantly higher when compared to other subtypes, and the triglyceride values in this subtype were slightly higher than triglyceride values in other subtypes. All of the subtypes except subtype 6 had more female patients than male. However, in this subtype it was 65% male to 35% female. For instance, for subtype 5, the HDL values were average. The creatinine values in this subtype were comparable to the values in the other subtypes. The LDL value was average when compared to other subtypes, and the triglyceride values in this subtype were similar to those in other subtypes. Additionally, this subtype was 100% female.

4.1 Analysis of Results

By creating a trajectory for patient's eGFR values, a disease severity trajectory was created. Cohorts of similar severity

trajectory gave insights into the causality of such changes and led to risk factors. Of all of the lab results per subtype in only ALT, AST, HDL, Glucose, LDL, Triglycerides, and gender had a significant difference in value between subtypes, while birth year, blood glucose levels, blood pressure, UACR, and hemoglobin levels did not show significant trends.

The ALT values were relatively higher in cluster 4. This could potentially show that high ALT levels coincide with a recovery from CKD.

A similar conclusion can be drawn from the AST distribution by subtype. The AST values were higher than average in subtype 4, which shows a recovering trend from CKD, and the AST values were lower in subtypes 3 and 6, which had a trend towards deteriorating kidney function. High AST levels could indicate a recovery from CKD, while low AST levels indicate a deterioration in kidney function, or the progression of CKD.

The only clear observation from the HDL graph was that subtype 6 had extremely low HDL levels for all patients, and these patients had a severe decline in kidney function.

The creatinine levels by subtype showed that patients in Subtype 5 had low creatinine levels while patients in subtype 6 had higher creatinine levels. This could indicate that patients that recovered from CKD had lower creatinine values while patients whose CKD progresses severely had high creatinine values.

The LDL results by subtype showed that patients in Subtype 2 and 6 had high LDL levels. Both Subtype 2 and 6 described patients with decreased kidney function over time, which could show that high LDL levels coincided with the progression of CKD.

Triglyceride lab result values by subtype displayed that patients in subtypes 3 and 6 had higher Triglyceride levels than the other patients. Furthermore, subtypes 3 and 6 described patients with decreased kidney function over time, which indicates that high triglyceride levels were a risk factor for CKD.

The gender distribution by subtype showed that there were more female patients in the subset than male patients. This could be because female patients were more likely to follow-up on doctor appointments and therefore were more likely to have over 20 non-null eGFR values. Given that the majority of patients were female in the subset taken into consideration, the subtypes with the most males, 4 and 6 both had upward trajectories and recovered from CKD. So, it was concluded that males were more likely to recover from CKD than females.

A general profile of patient who recovered from CKD was: high ALT level, high AST level, low Creatinine level, and male, while a general profile of a patient whose CKD progresses was: Low AST level, low HDL level, high Creatinine level, high LDL level, and high triglyceride level.

5 CONCLUSIONS

Chronic Kidney Disease is the 9th leading cause of death in the US [4] with little understanding about the progression of the disease. Finding accurate and detailed phenotypes can lead to identification of medically relevant disease subtypes, which is a vital component of modern precision medicine, and application of the precision medicine philosophy can be extremely helpful in developing new treatments for patients with CKD. In this paper, we presented a methodology to use EHR data records to extract patient clusters corresponding to potential disease subtypes. The irregular data sampling issue with EHR measurements was solved by using a regression model to impute the desired lab test value (e.g., eGFR) using other lab test results. The six subtypes identified through this analysis can be used to better understand phenotypes for CKD and combined with other patient information such as genomic surveys for developing better treatments.

REFERENCES

[1] H. D. Anderson, W. D. Pace, E. Brandt, R. D. Nielsen, R. R. Allen, A. M. Libby, D. R. West, and R. J. Valuck. Monitoring suicidal patients in primary care using electronic health records. The Journal of the American Board of Family Medicine, 28(1):65–71, 2015.

[2] Y. Hagar, D. Albers, R. Pivovarov, H. Chase, V. Dukic, and N. Elhadad. Survival analysis with electronic health record data: Experiments with chronic kidney disease. Statistical Analysis and Data Mining: The ASA Data Science Journal, 7(5):385–403, 2014.

[3] V. Jha, G. Garcia-Garcia, K. Iseki, Z. Li, S. Naicker, B. Plattner, R. Saran, A. Y. Wang, and C. W. Yang. Chronic kidney disease: global dimension and perspective. Lancet, 382(9888), 2013.

[4] M. Naghavi, H. Wang, R. Lozano, A. Davis, X. Liang, M. Zhou, S. E. Vollset, A. A. Ozgoren, S. Abdalla, F. Abd-Allah, et al. Global, regional, and national age- sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013. Lancet, 385(9963):117–171, 2015.

[5] P. N. Robinson. Deep phenotyping for precision medicine. Human Mutation, 33(5), 2012.

[6] Leonard Kaufman, P. R. (1987). Clustering by means of medoids. Statistical Data Analysis Based on the L1-Norm and Related Methods, 405-416.

[7] Pace W, Fox C, White T, Graham D, Schilling LM, West DR. The DARTNet Institute: Seeking a sustainable support mechanism for electronic data enabled research networks. eGEMs (Generating Evidence & Methods to improve patient outcomes), 2014; 2:1063.