eGEMs
Generating Evidence & Methods to improve patient outcomes

Extracting Deep Phenotypes for Chronic Kidney Disease Using Electronic Health Records

Duc Thanh Anh Luong;[i] Dinh Tran;[i] Wilson D. Pace, MD, FAAFP;[ii] Miriam Dickinson, PhD;[ii] Joseph Vassalotti, MD;[iii] Jennifer Carroll, MD, MPH;[ii] Matthew Withiam-Leitch, MD;[i] Min Yang, MD, PhD;[i] Nikhil Satchidanand, PhD;[i] Elizabeth Staton, MSTC;[ii] Linda S. Kahn, PhD;[i] Varun Chandola, PhD;[i] Chester H. Fox, MD[i]

ABSTRACT

Introduction: As chronic kidney disease (CKD) is among the most prevalent chronic diseases in the world with various rate of progression among patients, identifying its phenotypic subtypes is important for improving risk stratification and providing more targeted therapy and specific treatments for patients having different trajectories of the disease progression.

Problem Definition and Data: The rapid growth and adoption of electronic health records (EHR) technology has created a unique opportunity to leverage the abundant clinical data, available as EHRs, to find meaningful phenotypic subtypes for CKD. In this study, we focus on extracting disease severity profiles for CKD while accounting for other confounding factors.

Probabilistic Subtyping Model: We employ a probabilistic model to identify precise phenotypes from EHR data of patients who have chronic kidney disease. Using this model, patient's eGFR trajectory is decomposed as a combination of four different components including disease subtype effect, covariate effect, individual long-term effect and individual short-term effect.

Experimental Results: The discovered disease subtypes obtained by Probabilistic Subtyping Model for CKD are presented and their clinical relevance is analyzed.

Discussion: Several clinical health markers that were found associated with disease subtypes are presented with suggestion for further investigation on their use as risk predictors. Several assumptions in the study are also clarified and discussed.

[i]University at Buffalo, [ii]University of Colorado, Denver, [iii]Icahn School of Medicine at Mount Sinai

CONTINUED

Conclusion: The large dataset of EHRs can be used to identify deep phenotypes retrospectively. Directions for further expansion of the model are also discussed.

Introduction

Precision medicine is an emerging medical paradigm that takes into account individual variability in genes, environment, and lifestyle to develop targeted therapy and prevention strategies. The recently launched *Precision Medicine Initiative*[1] underscores the importance of this area. In the context of precision medicine, one of the key scientific challenges is the ability to identify stratified subgroups of individuals who exhibit similar disease-related behavior. In medical terms, this process is known as phenotyping, where the phenotypes are the observable traits exhibited by the subpopulation. Phenotypes are typically the starting point for inquiry in clinical practice and medical research. In the past, the focus has been on identifying coarse phenotypes from patient level data obtained from specific patient cohorts. However, precision medicine demands identification of precise phenotypes, also referred to as deep phenotyping.[3]

Clinical data has the potential to be a cost-effective and large-scale source of deep phenotypes. The rapid growth in *Electronic Health Record* (EHR) technology has the potential to make clinical data much more accessible for analysis. In the context of precision medicine, emergence of networks such as DARTNet[4] and eMERGE[5] are indicative of the importance of identifying phenotypes from clinical data. EHR data collections allow us to study patient populations at an unprecedented scale. For example, the DARTNet data collection corresponds to approximately 12.5 million patient visits per year, 5 million patients, and 5 billion data points (clinical tests, diagnoses, procedures, medications, etc.).

One large collection of EHR data is the DARTNet *Chronic Kidney Disease* (CKD) dataset.[4] CKD is well recognized as a rising problem in global health. According to 2013 Global Burden Disease study,[6] approximately 956,000 deaths were caused by CKD worldwide in 2013. In the same study, CKD was ranked 19th in the top 50 causes of global years of life lost in 2013. In the United States, it is the 9th leading cause of death and affects more than 20% of the U.S. adult population.[7] Analysis of a large EHR dataset shows that there is still room for improvement in clinical care for patients with CKD.[8]

The natural history of CKD often begins with initial kidney damage and progresses through stages of CKD, with decline of glomerular filtration rate (GFR) towards the end stage of renal failure.[9] However, the course of GFR decline among patients is heterogeneous depending on individual, ethnic, and disease specific conditions. The predictors to adverse outcomes of CKD still need to be clarified so that targeted therapy can be implemented based on risk stratification. To elucidate this question, we first need to identify different phenotypes of CKD progression and factors associated with various phenotypes. The goal of this study is to stratify a

large CKD patient population into subgroups, such that each subgroup corresponds to a distinct disease progression profile.

As far as we know, this is one of the first studies that attempts to find disease subtypes of CKD by exploiting large electronic health records. One previous study in CKD with large EHR dataset is Hagar et al.'s study[10] which gives survival analysis on CKD with different types of covariates. However, the paper tends to focus more on survival analysis with CKD as a case study using wide varieties of information on EHRs. In our study, we focus more on temporal aspects of eGFRs, which are a key indicator of severity of CKD. In addition, the problem we are interested in is finding disease's subtypes, which is not mentioned in Hagar et al.'s study.[10]

A study by R. Pivovarov et al.[11] shows a capability of learning phenotypes from heterogeneous EHR data. In that work, the authors considered EHR data as a collection of text including notes, medication orders, diagnosis codes and laboratory tests. Each record contains a bag of words and the relationship between phenotypes and these health records is modeled as a probabilistic graphical model in which a phenotype is a hidden random variable. In our work, we approach the problem of finding phenotypes in EHRs in a different way. Instead of conducting text mining on EHR data, we focus more on laboratory tests and their numerical values. In addition, the temporal aspect of EHRs, which was not considered in that work, is an important element in modeling the disease progression.

Perhaps the closest previous study to our work is by Schulam et al.,[12] in which a probabilistic graphical model is proposed to model different aspects of disease progression including disease subtype, covariate effect, long-term and short-term effect.

In this paper, we apply the same Probabilistic Subtyping Model (PSM)[12] for clustering the disease progression trajectories, while accounting for the effect of patient-specific covariates. In particular, PSM considers factors that contribute to disease progression, including long-term health condition, short-term health condition, patient-specific covariates and trajectory of patient's clinical health marker. In the context of CKD, our selected clinical health marker is estimated glomerular filtration rate. By examining the resulted subtypes identified through PSM-derived model, we can have a better understanding of CKD's phenotypes and therefore provide appropriate care for patients based on their common disease progression. We apply PSM on data available from DARTNet. The subtypes identified using PSM are promising candidates for further study.

Problem Definition and Data

From the precision medicine perspective, our objective is to demonstrate the value of a publicly available clinical data resource for precision medicine. The data resource comes from a collaboration of nine practice-based research networks, called the DARTNet Institute.[4] The partners within the institute are building a national collection of data from electronic health records, claims, and patient-reported outcomes. The nine distinct research networks that make up DARTNet Institute offer access to approximately 12.5 million patient visits per year, five million patient lives, and approximately five billion data points. This big health data resource has immense potential for fostering medical research, and there have been studies that have used the data.[13,14] The data from DARTNet can be used to track patients over several years in terms of disease severity, using information about clinical tests, comorbidities, physiological characteristics, and medications. Moreover, the

active engagement of medical practices allows a pathway for researchers to obtain more information about the patients via genomic sequencing and externally conducted surveys. In particular, we plan to utilize one curated data set that was extracted as a part of a Chronic Kidney Disease Natural History Study, which corresponds to 69,817 patients suffering from CKD. Table 1 lists various available elements in the dataset. It is clear that the richness offered by this data resource both in terms of clinical information and cohort size presents an unprecedented opportunity to understand the role of deep phenotypes for precision medicine. To focus the scope of the proposed research, we targeted extraction of *disease severity profiles* as the phenotypes while accounting for confounding factors such as demographic characteristics.

Target Variable

The target variable that we are interested in is called the *estimated glomerular filtration rate* (eGFR), which is a standard test to measure the level of kidney function in an individual. eGFR is typically estimated from a laboratory test that measures the creatinine level, using the MDRD Study equation.[15] However, estimation of GFR using the MDRD Study equation has limited precision and systematically underestimates GFR at higher values. In 2009, the CKD-EPI group has shown that the CKD-EPI equation is more accurate than the MDRD equation.[16] For this reason, we use the CKD-EPI equation as an estimation for GFR. The formula for CKD-EPI is presented in Table 2. Given that the data was collected from nine practice-based research networks, different clinics used different coding for race information. For this reason, in our calculation of eGFR, we assume that all patients are White.

According to the National Kidney Foundation, eGFR for a normal individual ranges from 90-120. If the eGFR value is below 60 for more than 3 months, it signals a transition to stage 3 CKD. With the above condition, we only focus on patients with eGFR value below 60 for more than three months in this paper.

Probabilistic Subtyping Model

We used the Probabilistic Subtyping Model[12] to explain different factors in variations of eGFR in patients' profiles. We assume that the population consists of M patients. For the i^{th} patient ($1 \le i \le M$), we have an eGFR sequence denoted as a vector $\boldsymbol{y}_i$ consisting of N_i observations collected at times denoted by vector $\boldsymbol{t}_i$. Thus $\boldsymbol{y}_{ij}$ denotes the eGFR reading for patient i taken at time $\boldsymbol{t}_{ij}$. The static covariates for i^{th} patient are represented by a vector $\boldsymbol{x}_i \in \mathrm{R}^C$, where C is the number of available covariates. Conceptually, each patient's eGFR trajectory $\boldsymbol{y}_i$ is modeled as a Gaussian random variable with a mean value that is explained using four different components (See Figure 2 for graphical representation): disease subtype effect, covariate effect, individual long-term effect and individual short-term effect.

Disease Subtype Effect

A disease subtype can be described broadly as hidden traits that a sub-population of patients share. This disease subtype has an effect on disease progression, which can be observed as trajectory of eGFR records. In PSM, the disease subtype associated with a patient is modeled as a hidden discrete variable which is determined probabilistically.

Assuming that there are G disease subtypes, the membership of patient i is modeled using a latent multinomial random variable $z_i \in \{1,2,\cdots,G\}$. A vector $\pi \in \mathrm{R}^G$ parameterizes the multinomial distribution, such that $p(z_i) \sim Mult(z_i;\pi)$. We apply a symmetric Dirichlet prior on the vector $p(\pi) \sim Mult(\pi;\alpha)$ where α is the concentration parameter.

Table 1. List of Data Elements Available in CKD Dataset

BASELINE CHARACTERISTICS	SUMMARY STATISTICS	
	"PREPROCESSED" DARTNET PATIENTS	CKD COHORT
Number of patients	63209	17314
Age at baseline (years)	66.32 (57.50, 74.55)	70.20 (63.13, 76.56)
SEX (%)		
Male	26034 (41.19)	6856 (39.60)
Female	37175 (58.81)	10458 (60.40)
Smoking status		
CKD INDICATORS		
Serum creatinine	1.1 (0.9, 1.3)	1.2 (1.0, 1.5)
Years of last serum creatinine measure	2.76 (0.77, 4.75)	4.48 (3.07, 5.85)
Albumin-to-creatinine ratio	26.5 (8.0, 55.6)	22.3 (7.5, 41.6)
OTHER INDICATORS		
Hemoglobin A1c	6.6 (6.1, 7.4)	6.6 (6.1, 7.3)
Alanine aminotransferase	22 (15, 33)	21 (15, 32)
Aspartate aminotransferase	21 (17, 26)	20 (17, 25)
Fasting Blood Glucose	101 (92, 119)	102 (92, 120)
Non-Fasting Blood Glucose	102 (91, 123)	103 (91, 123)
Triglyceride level	128 (91, 184)	131 (93, 185)
High Density Lipoprotein	47 (39, 58)	47 (39, 57)
Low Density Lipoprotein	95 (75, 120)	91 (72, 115)
Phosphorous	3.6 (3.2, 4.2)	3.5 (3.2, 3.9)
Parathyroid hormone	61.0 (35.6, 111.0)	56.3 (34.4, 92.0)
Height (inch)	66 (63, 69)	66 (63, 69)
Weight (lb)	184.0 (155.7, 217.0)	184.0 (156.0, 215.0)
Systolic blood pressure	130 (120, 140)	130 (120, 140)
Diastolic blood pressure	76 (70, 82)	74 (68, 80)

Note: "Preprocessed" DARTNet patients are extracted by a procedure explained in Figure 1.
Continuous variables are summarized by median while 25th and 75th percentiles are presented in parenthesis. Categorical variables are summarized by number of patients in each category while percentage is presented in parenthesis.

Figure 1. Flowchart of Preprocessing DARTNet Data

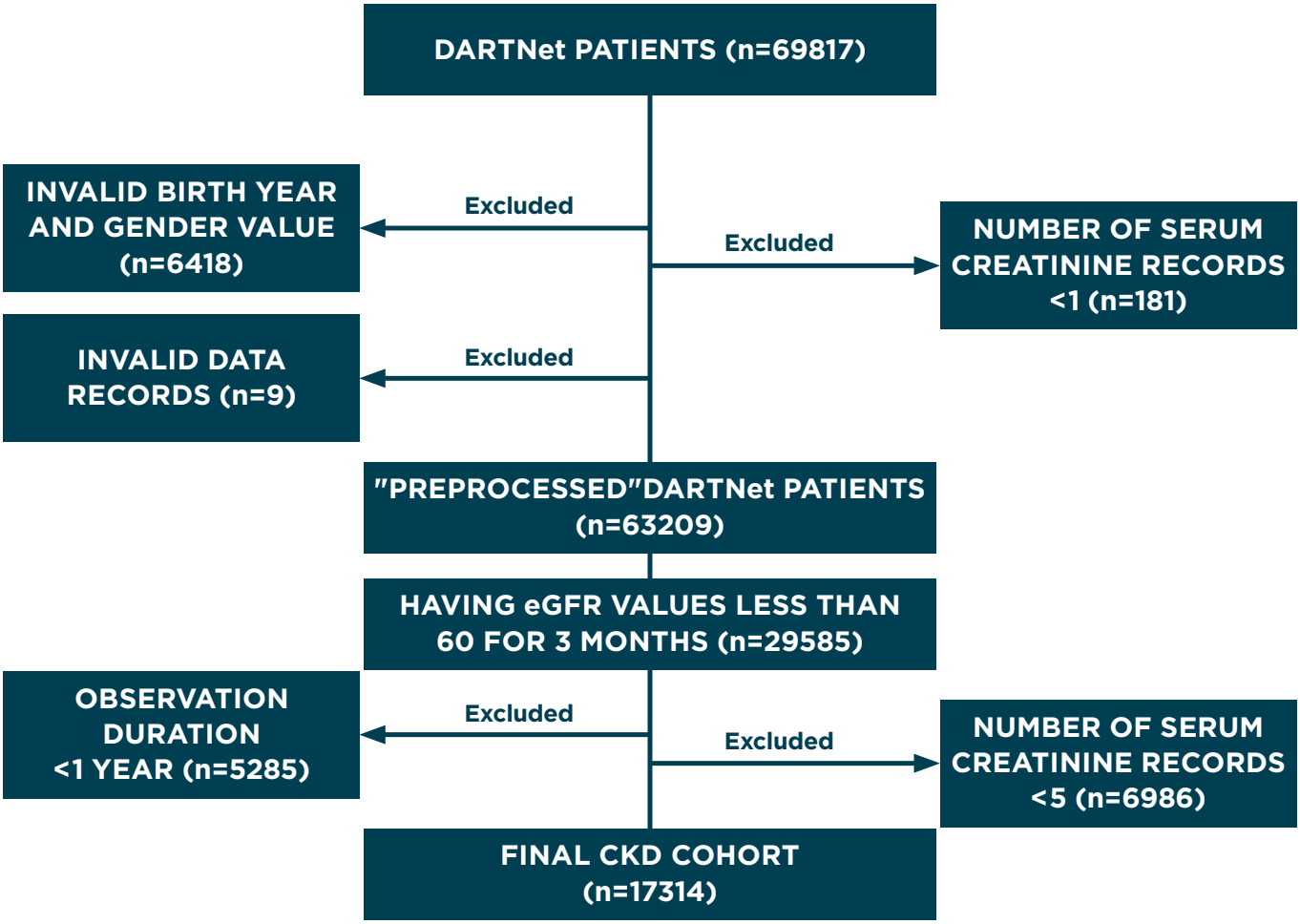
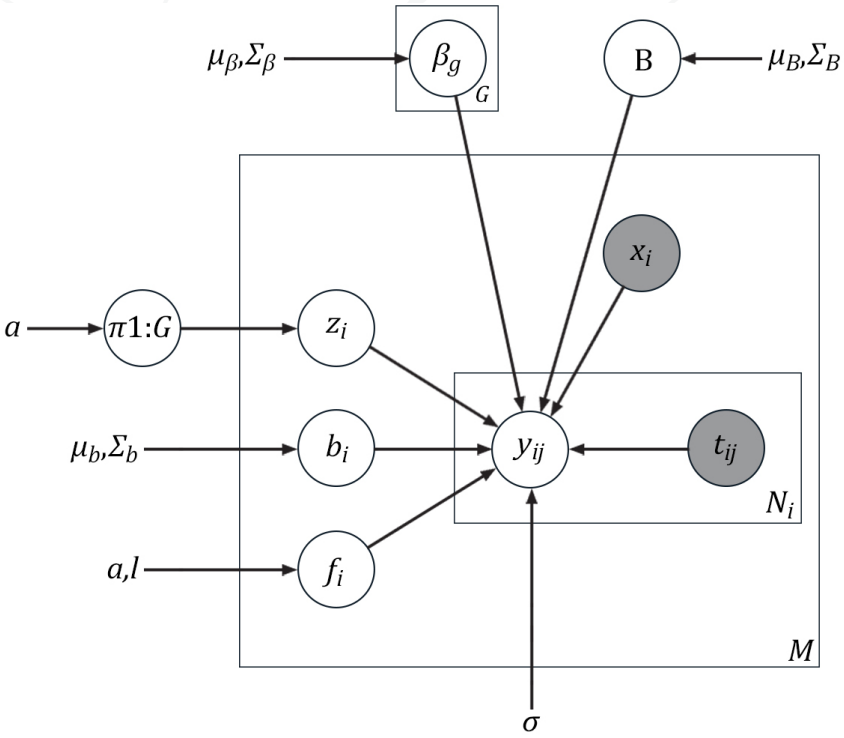


Table 2. The CKD-EPI Equation for Estimating GFR

RACE AND SEX	SERUM CREATININE	EQUATION
BLACK		
Female	≤0.7	$GFR = 166 \times (Scr/0.7)^{-0.329} \times (0.993)^{Age}$
	>0.7	$GFR = 166 \times (Scr/0.7)^{-1.209} \times (0.993)^{Age}$
Male	≤0.9	$GFR = 163 \times (Scr/0.9)^{-0.411} \times (0.993)^{Age}$
	>0.9	$GFR = 163 \times (Scr/0.9)^{-1.209} \times (0.993)^{Age}$
WHITE OR OTHERS		
Female	≤0.7	$GFR = 144 \times (Scr/0.7)^{-0.329} \times (0.993)^{Age}$
	>0.7	$GFR = 144 \times (Scr/0.7)^{-1.209} \times (0.993)^{Age}$
Male	≤0.9	$GFR = 141 \times (Scr/0.9)^{-0.411} \times (0.993)^{Age}$
	>0.9	$GFR = 141 \times (Scr/0.9)^{-1.209} \times (0.993)^{Age}$

Figure 2. Probabilistic Subtyping Model



Each subtype eGFR trajectory is modeled as a weighted sum of P B-spline basis functions with weight vector $\boldsymbol{\beta}_g \in \mathrm{R}^P$. For each k in $\{1,2,\cdots,P\}$, we denote $\boldsymbol{\phi}_k(\boldsymbol{t}_i) \in \mathrm{R}^{P\times 1}$ as the k^{th} B-spline basis function applied on the vector $\boldsymbol{t}_i$. We also denote $\boldsymbol{\phi}(\boldsymbol{t}_i) = [\boldsymbol{\phi}_1(\boldsymbol{t}_i), \boldsymbol{\phi}_2(\boldsymbol{t}_i), \cdots, \boldsymbol{\phi}_P(\boldsymbol{t}_i)]$. With these notations, the contribution of the subtype mechanism in the trajectory for patient i, such that $g = z_i$, can be written as follow:

$$e_i^{subtype} = \boldsymbol{\phi}(\boldsymbol{t}_i)\boldsymbol{\beta}_{zi} \qquad (4.2)$$

The coefficient vectors $\boldsymbol{\beta}_g$ are themselves drawn from a prior multivariate Gaussian distribution, i.e., $p(\boldsymbol{\beta}_g) \sim N(\boldsymbol{\beta}_g; \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$.

Covariate Effect

Specifically modeling the effect of patient-level covariates such as the gender, age, and smoking behavior is important because two patients with similar covariates might appear correlated in terms of the eGFR profiles. For this reason, the covariate effect is captured in the model as a term contributed to the total effect of the target variable. In the context of CKD, relevant covariate data includes race, gender and smoking behavior. Since smoking behavior is a temporal covariate, which may have different status over time, precisely modeling the duration of exposure to smoking within the PSM model is difficult; this will be considered in future extensions of the model. In addition, as race information is not available in our data, we only focus on modeling the covariate effect of gender in our analysis.

Conceptually, covariate values define sub-groups of patients who have similar traits, i.e., gender in this study. We want to model these sub-groups of patients to have similar patterns of disease progression. In PSM, each sub-group covariate effect is modeled as a linear effect that contributes to the eGFRs of patients, as follows:

$$e_i^{covariate} = \boldsymbol{\gamma}(\boldsymbol{t}_i)\boldsymbol{\rho}(\boldsymbol{x}_i) \qquad (4.3)$$

where $\boldsymbol{\gamma}(\boldsymbol{t}_i) = [1, \boldsymbol{t}_i]$ and $\boldsymbol{\rho}(\boldsymbol{x}_i)$ is a patient-specific coefficient vector (slope and intercept) which is obtained through a linear combination of patient specific covariates $\boldsymbol{x}_i$ using a $2 \times \boldsymbol{C}$ loading matrix, $\boldsymbol{B}$, i.e., $\boldsymbol{\rho}(\boldsymbol{x}_i) = \boldsymbol{B}\boldsymbol{x}_i$. The two rows of the loading matrix $\boldsymbol{B}$, denoted as $\boldsymbol{B}_0$ and $\boldsymbol{B}_1$ are modeled using a multivariate Gaussian distribution, i.e., $p(\boldsymbol{B}_k) \sim (\boldsymbol{B}_k; \boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$, $k \in \{0,1\}$.

Individual Long-term Effect

Beside disease's subtype and covariate effect, individual long-term health conditions are also an important factor that can help us explain additional variations in a patient's eGFR's trajectory. For example, eGFR values of a patient who has an unusually weak renal system are expected to decline faster than other normal patients. In PSM, individual long-term effect is modeled as a linear trend. It is also worth noting that patient's eGFR trajectory may not follow a linear trend as shown in Li et al.'s study.[17] However, non-linear residuals in patient's eGFR trajectory which cannot be explained by the above three effects (subtype, covariate and individual long-term) will be later modeled using short-term effect.

The individual long-term effect of patient i can be written as follow:

$$e_i^{long\text{-}term} = \boldsymbol{\gamma}(\boldsymbol{t}_i)\boldsymbol{b}_i \qquad (4.4)$$

where $\boldsymbol{b}_i \in \mathrm{R}^{2\times 1}$ which represents the slope and intercept of individual long-term effect.

Individual Short-term Effect

Sometimes eGFR value can vary beyond the explanation of the subtype effect, covariate effect, and individual long-term effect. We can attribute this variation to temporary changes in a patient's health condition which affect the test results and

subsequently affect the calculated eGFRs. In PSM, these short-term changes in eGFR trajectory are modeled as a Gaussian process[18] with mean $\mathbf{0}$ and kernel function parameterized by hyper-parameter a and l:

$$e_i^{short\text{-}term} = f_i \sim GP(\mathbf{0}, k(\cdot,\cdot)) \quad (4.5)$$

$$k(t_1, t_2) = a^2 \exp\left(-\frac{(t_1 - t_2)^2}{2l^2}\right) \quad (4.6)$$

Experimental Results

In this section, we present experimental results to show how PSM can identify disease subtypes within the CKD cohort by analyzing the clinical data from the DARTNet dataset. First, we explain how we preprocess data so that they can be used as inputs for PSM. After that, we present the subtypes discovered by PSM and examine features associated with each subtype.

Data Preprocessing

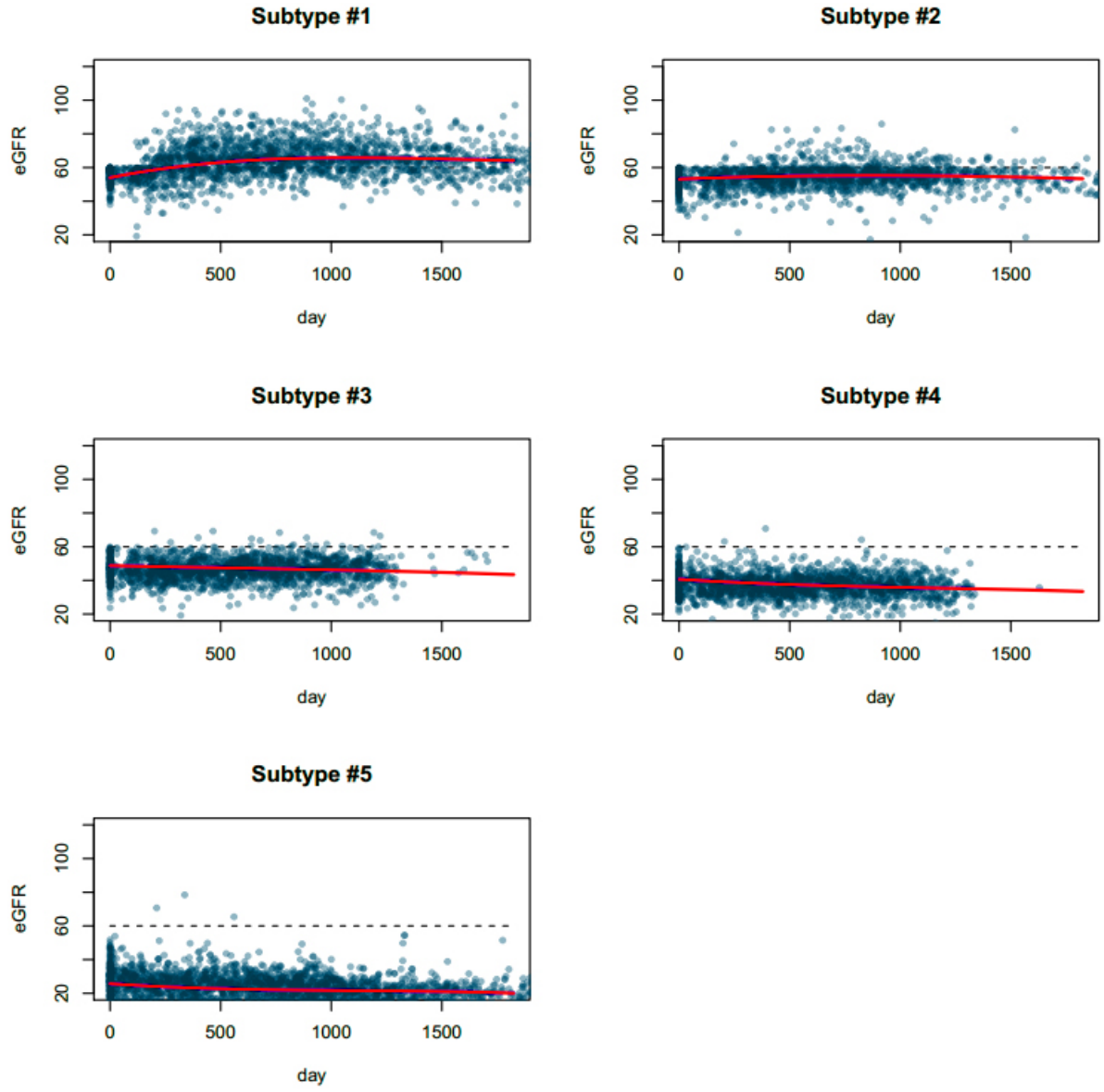
Although the dataset contains data for 69,817 patients, in order to ensure the quality of our analysis, we only choose a subset of patients as a CKD cohort while excluding others whose data do not satisfy our criteria. In particular, we target a group of patients who have eGFR values less than 60 for more than three months. This criterion is usually used in clinical practice to identify patients having CKD. Moreover, it is also viewed as selecting only patients transitioning to stage 3 CKD as well as existing patients in stage 3, 4 and 5 of CKD. In addition, we exclude patients who have invalid birth year and gender value in their records since age and gender are two important values needed to estimate GFR value. We also exclude patients who have less than a year of creatinine data available. Furthermore, having too few data points in eGFR readings can deteriorate the performance of deriving subtype trajectories; thus, in our experiments, only patients with at least five data points of serum creatinine values are considered. This lower bound of number of data points is chosen empirically so that we can ensure data quality while selecting a significant patient population for analysis. The set of patients who have data satisfying all above conditions is our target cohort. This cohort has 17,314 patients, represent 24.80% of total patients in the original dataset. Figure 1 presents a flowchart that shows the preprocessing steps we used to obtain the final CKD cohort. When choosing a cohort, our criteria had been deliberately designed so that the cohort can contain as many patients as possible while maintaining the quality of analysis with enough data.

All eGFR values of patients are computed using CKD-EPI equation as presented in Table 2. Finally, in order to remove outliers of test results from consideration, we remove all eGFR values which are beyond the five standard deviations from the mean eGFR value.

Discovered Disease Subtypes

Using the Probabilistic Subtyping Model, we perform an optimization process as described in Schulam et al.'s work.[9] The result of this optimization is a probability distribution over all subtypes for each patient and the overall covariate effect, which maximizes the complete data log-likelihood. Disease subtype for each patient is subsequently determined by the subtype with highest probability. Given the computed disease subtype and covariate effect, individual long-term and short-term effects are calculated. Using Bayesian Information Criteria (BIC), we determine five as the number of subtypes. Figure 3 shows the result of five subtype trajectories found in the experiment. We have ordered these subtypes from best CKD prognosis to worst. The red lines in Figure 3 represent the subtype's prototype trajectories learned from PSM while the blue dots are eGFR values of 200 sample patients who are probabilistically assigned to that subtype.

Figure 3. Subtype Trajectories



Another way to characterize the subtype trajectories is to use Table 3, which summarizes the rate of change of eGFR as well as the baseline eGFR for each subtype. From both Figure 3 and Table 3, one can observe distinct trends for the different subtypes. In particular, although subtype 1 and subtype 2 have similar baseline eGFR value, they have different rates of change of eGFR per year. While subtype 1 has a slightly upward trend, subtype 2 remains stable over the follow-up period. On the other hand, subtypes 3, 4 and 5 all have downward trends with different rates of decline and different baseline eGFR values. Subtype 5 has a very low baseline eGFR while subtype 3 and 4 have better

baseline eGFR values in comparison with subtype 5. Subtype 3 and subtype 4 can be differentiated by rate of change of eGFR per year and baseline eGFR. In other words, subtype 3 has slower rate of change of eGFR per year and higher baseline eGFR when comparing with subtype 4. In fact, the found subtypes coincide with existing knowledge about patient subgroups in CKD: 1) Subtype 1 corresponds to the group of patients that yields improvement in kidney's function; 2) Subtype 2 indicates a group of patients whose kidney's function is stable in moderate level over the follow-up period; 3) Subtype 3 indicates a set of patients which have slow decline in kidney's function; 4) Subtype 4 represents a set of patients that have a steady decline in renal capability; 5) Subtype 5 coincides with a group of patients having severe damage of kidney's function.

In order to view the resulting subtypes from a different perspective, we look at demographics of each subtype to see if there are any demographic distinctions between different subtypes. Figure 4 shows that the subtypes do not exhibit significant distinction for gender. Figure 5 provides the distribution of baseline age of patients belonging to each subtype. One observation from this figure is that the severity of each subtype is marginally correlated with its corresponding baseline age, with the exception of subtype 5. It is also worth noting that there is a linear trend of baseline age from subtype 1 to subtype 4. For a closer look at each subtype adjusted for age and gender, we give rate of change and baseline eGFR values for each subtype adjusted for age and gender in Table 4 and Table 5 respectively. As shown in Table 4, within the same subtype, female patients on average have better rate of decline and baseline eGFR values in comparison with male patients. Table 5, on the other hand, shows that for a same subtype, the older group of patients on average have worse rate of decline and baseline eGFR values in comparison with younger groups of patients. Finally, Figure 6 shows the distribution of patients among subtypes. In particular, subtype 1 and subtype 2 comprise nearly 50 percent of all patients. On the other hand, subtype 5 in which patients mostly have severe kidney damage only contains five percent of total patients in the CKD cohort.

Table 3. Description of Subtypes in Terms of Their Trajectories

		SUBTYPE 1	SUBTYPE 2	SUBTYPE 3	SUBTYPE 4	SUBTYPE 5
Patient records	Average rate of change of eGFR per year	4.08	0.54	-0.93	-1.54	-1.80
	Average baseline eGFR value	54.71	53.38	48.93	40.90	26.45
Prototype's trajectory	Rate of change of eGFR per year	2.03	0.04	-1.07	-1.45	-1.13
	Baseline eGFR value	53.97	53.13	48.77	40.69	25.69

Figure 4. Distribution of Gender for Each Subtype

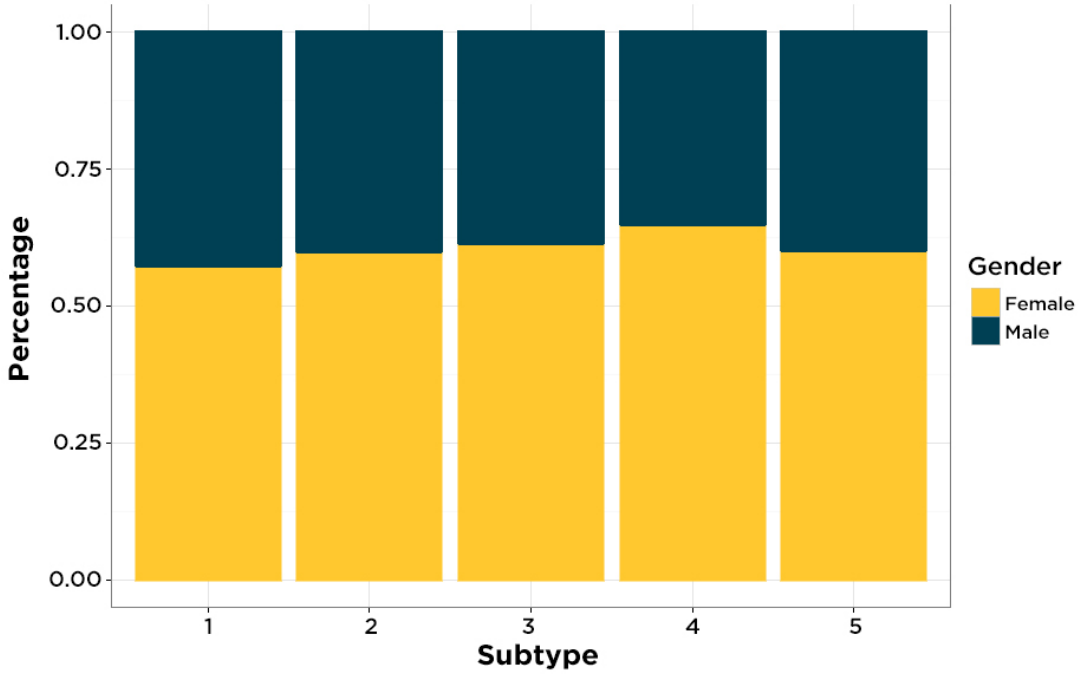


Figure 5. Distribution of Baseline Age of Patients for Each Subtype

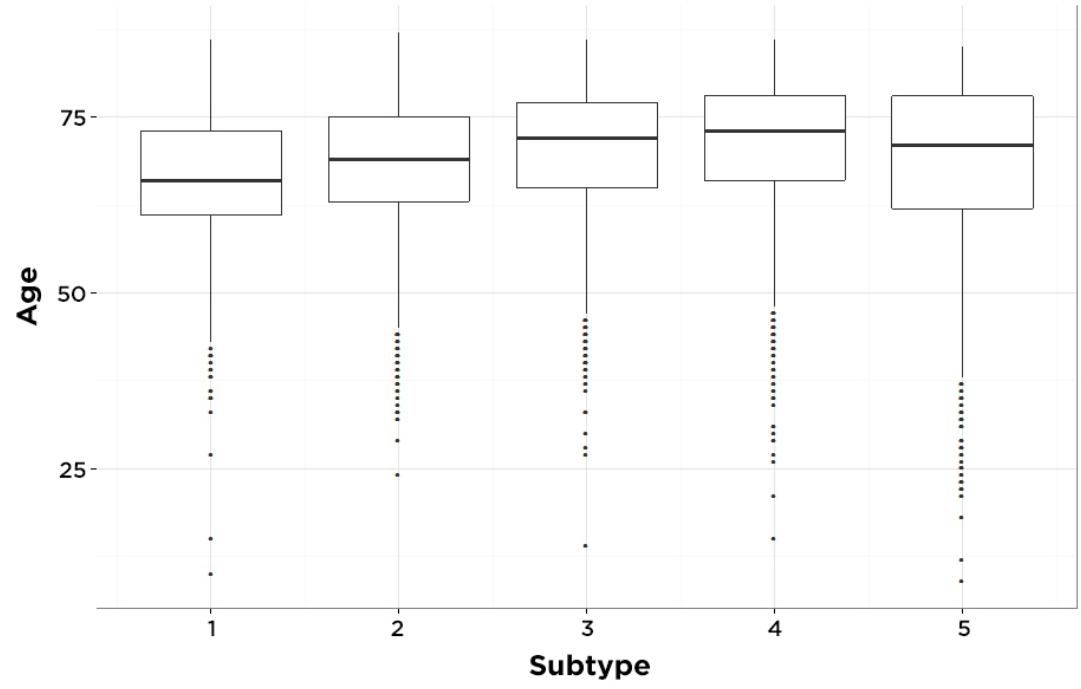


Figure 6. Distribution of Patients for Each Subtype

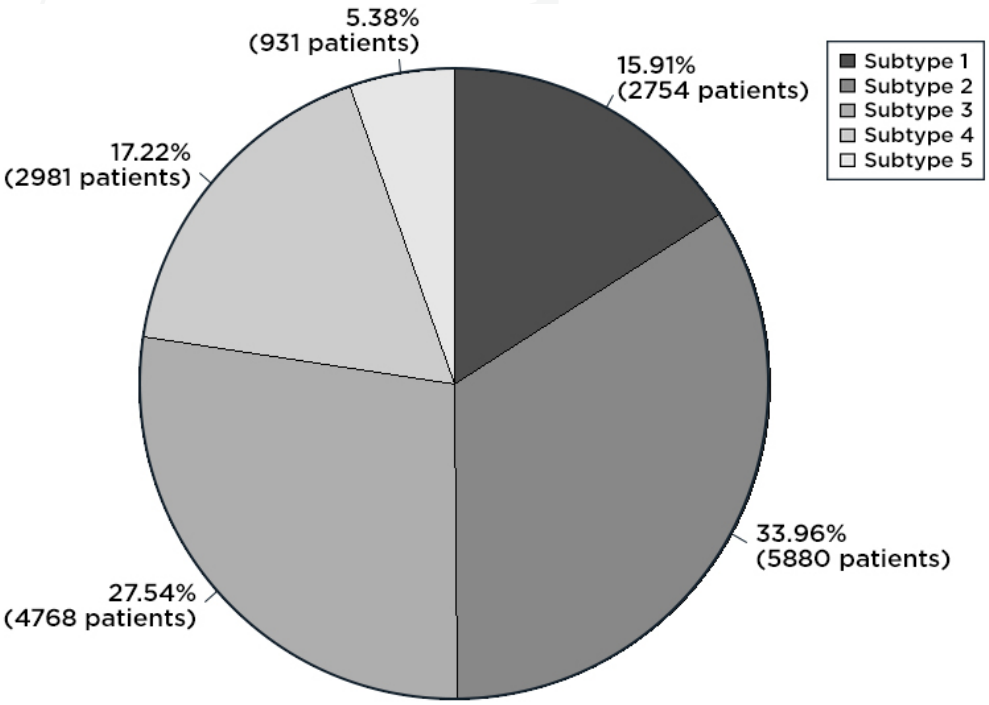


Table 4. Rate of Change and Baseline eGFR of Each Subtype Breaking Down by Gender

SUBTYPE	GENDER	AVERAGE RATE OF CHANGE OF EGFR PER YEAR	AVERAGE BASELINE EGFR VALUE
Subtype 1	Female	4.28	54.52
	Male	3.83	54.96
Subtype 2	Female	0.60	53.15
	Male	0.44	53.72
Subtype 3	Female	-0.87	48.87
	Male	-1.03	49.02
Subtype 4	Female	-1.42	40.72
	Male	-1.76	41.22
Subtype 5	Female	-1.78	27.00
	Male	-1.83	25.63

Table 5. Rate of Change and Baseline eGFR of Each Subtype Breaking Down by Age

SUBTYPE	AGE GROUP	AVERAGE RATE OF CHANGE OF EGFR PER YEAR	AVERAGE BASELINE EGFR VALUE
Subtype 1	< 45	9.10	52.71
	45-65	4.13	54.97
	> 65	3.91	54.56
Subtype 2	< 45	0.46	53.50
	45-65	0.62	53.65
	> 65	0.50	53.24
Subtype 3	< 45	-1.68	49.14
	45-65	-0.84	49.11
	> 65	-0.96	48.86
Subtype 4	< 45	-2.32	42.52
	45-65	-2.24	42.58
	> 65	-1.34	40.43
Subtype 5	< 45	-1.20	13.46
	45-65	-2.22	27.84
	> 65	-1.73	27.61

Table 6 shows the distributions of various test results of patients belonging to each subtype. Among clinical lab measures presented in Table 6, albumin-to-creatinine ratio (ACR) is an important indicator, which is used for predicting CKD progression. As we can see in the table, the more severe the subtype is, the higher the value of albumin-to-creatinine ratio. This indicates that albumin-to-creatinine ratio can also be an indicator for each subtype ranking from best prognosis to worst, which reinforces our understanding that eGFR and albumin-to-creatinine ratio are independent and complementary predictors for CKD progression.[9]

One interesting finding from Table 6 is the correlation between value of alanine aminotransferase (ALT) and the subtypes. In particular, the more severe a subtype is, the less value of ALT it has. A similar observation can also be made with aspartate aminotransferase (AST) in Table 6. We notice that ALT and AST are measures of enzymes that are commonly used to assess liver function. This finding also agrees with previous study,[19] which mentioned the levels of ALT and AST in CKD patients. Another observation that one can have from Table 6 is the relationship of parathyroid hormone (PTH) and the subtypes. As presented in Table 6, in more severe subtypes, we observe slightly higher level of PTH.

Table 6. Summarization of Relevant Clinical Measures with Respect to Each Subtype

LAB MEASURES	SUBTYPE 1	SUBTYPE 2	SUBTYPE 3	SUBTYPE 4	SUBTYPE 5
Albumin-to-creatinine ratio	16.5 (6.8, 30.0)	14.7 (6.1, 30.0)	23.1 (7.7, 40.2)	30.0 (11.0, 93.8)	46.5 (17.0, 318.0)
Hemoglobin A1c	6.5 (6.0, 7.2)	6.5 (6.0, 7.2)	6.6 (6.1, 7.4)	6.7 (6.1, 7.5)	6.7 (6.1, 7.7)
Alanine aminotransferase	23 (16, 34)	22 (15, 33)	21 (15, 32)	19 (13, 29)	18 (12, 27)
Aspartate aminotransferase	21 (17, 26)	21 (17, 26)	20 (17, 25)	20 (16, 24)	19 (15, 24)
Fasting Blood Glucose	102 (92, 116)	101 (92, 117)	103 (92, 120)	104 (91, 127)	105 (92, 132)
Non-Fasting Blood Glucose	102 (91, 120)	101 (91, 118)	103 (92, 126)	105 (91, 132)	106 (91, 138)
Triglyceride level	128 (90, 180)	126 (90, 177)	133 (93, 190)	143 (103, 200)	144 (101, 201)
High Density Lipoprotein	47 (39, 58)	48 (40, 58)	46 (38.5, 57)	45 (37, 55)	43 (36, 53)
Low Density Lipoprotein	92 (72, 116)	93 (73, 117)	90 (71, 114)	89 (70, 113)	88 (68, 113)
Phosphorous	3.4 (3.1, 3.8)	3.4 (3.0, 3.7)	3.4 (3.1, 3.8)	3.6 (3.2, 4.0)	3.8 (3.3, 4.5)
Parathyroid hormone	42.4 (23.0, 59.0)	48.0 (31.7, 75.7)	53.9 (32.0, 84.4)	60.0 (40.5, 102.0)	114.0 (61.0, 203.0)
Systolic blood pressure	128 (120, 140)	130 (120, 140)	130 (120, 140)	130 (120, 142)	130 (120, 142)
Diastolic blood pressure	76 (70, 82)	76 (70, 80)	74 (68, 80)	72 (66, 80)	72 (66, 80)

Note: Clinical measures are summarized by median while 25th and 75th percentiles are presented in parenthesis.

As the role of PTH is to regulate the level of serum calcium and serum phosphate in body, the higher PTH is observed when kidney function decreases. This finding also agrees with previous study from Tomasello's work.[20]

In addition to Table 6, we also perform an analysis to examine which lab measure associates with each subtype. In order to estimate the degree of association, we use hypothesis testing to compare the level of lab measure in two groups: patients in a particular subtype (group 1) and the remaining patients (group 2). Our null hypothesis is that there is no difference between the distribution of level of lab measures in groups 1 and 2. We use t-test with the assumption that the variance is different between two groups. If the resulting p-value is smaller than a threshold of five percent, we can statistically reject the null hypothesis. In Table 7, we present a table of p-value for each lab measure and each subtype. All p-values in Table 7 that are larger than five percent are italicized. We can observe that the distributions of most lab measures are statistically different between patients in a particular subtype and the set of remaining patients. This analysis provides an estimation for understanding the degree of associations between subtypes and lab measures from statistical perspective and it would be useful for further investigation from clinical perspective.

Discussion

CKD is a chronic condition with a strong tendency for progression. Here, we recognized that 15.91 percent of patients with CKD achieved improvement of kidney function and 33.96 percent of them

Table 7. P-value of Hypothesis Testing for Each Lab Measure and Each Subtype

LAB MEASURES	SUBTYPE 1 (%)	SUBTYPE 2 (%)	SUBTYPE 3 (%)	SUBTYPE 4 (%)	SUBTYPE 5 (%)
Albumin-to-creatinine ratio	0.000	0.000	*5.372*	0.001	0.000
Hemoglobin A1c	0.000	0.000	*6.966*	0.000	0.000
Alanine aminotransferase	0.000	0.000	*10.376*	0.000	0.000
Aspartate aminotransferase	1.419	*7.157*	*16.910*	0.002	0.076
Fasting Blood Glucose	0.232	0.647	*10.737*	0.594	0.135
Non-Fasting Blood Glucose	0.000	0.000	1.705	0.000	0.000
Triglyceride level	0.193	0.000	1.785	0.000	0.000
High Density Lipoprotein	0.096	0.000	*25.368*	0.000	0.000
Low Density Lipoprotein	0.537	0.000	0.037	0.311	*0.547*
Phosphorous	3.624	0.000	0.590	1.747	0.000
Parathyroid hormone	0.000	0.000	0.005	*24.463*	0.000
Systolic blood pressure	*98.426*	1.015	*15.010*	*35.429*	*10.178*
Diastolic blood pressure	3.205	*19.894*	*54.079*	0.000	0.000

remain stable during the follow-up period. The rest of the patients displayed various levels of CKD progression. We also compared the distribution of a number of clinical measures among these CKD subtypes. Among these clinical measures, level of albumin-to-creatinine ratio (ACR), alanine aminotransferase (ALT), aspartate aminotransferase (AST) and parathyroid hormone (PTH) were found to be associated with discovered subtypes. A more complexed risk prediction model with multiple weighted factors needs to be developed in the future to further explore the predictors of CKD progression. The CKD subtypes identified in this study can also be utilized to validate some current risk prediction models.[21,22]

More questions can be answered with this large clinical dataset in the future. For example, in CKD, what is the effect of systolic BP control? Or, how detrimental are non-steroidal anti-inflammatory drugs? Having the sub-phenotypes can also be correlated with genomic, proteomic, and microbiome information as we move toward personalized and precision medicine. Patients in each subtype identified through this study are potential candidates for genomic, proteomic, and metabolomics studies in the future to identify new markers or risk factors for CKD progression.

The identified subtypes shown in Figure 3 show distinctive characteristics between the five subpopulations. However, these results are based on an implicit assumption that the clinical data spans the entire disease history starting from the onset of CKD. Clearly, there can be instances when the true onset would have happened much before the data

collection started. This is important in cases such as subtype 4 and 5, where the subtype trajectories have similar slopes and only differ in the baseline values. In the presence of longer histories, some of these subtypes could potentially merge or newer subtypes may arise.

Although we have presented the five subtypes found in CKD for a sub-population of patients transitioning to stage 3 of CKD and existing patients in stage 3, 4 and 5 of CKD, few assumptions and simplifications have been made in our work so that the computation is feasible. In particular, the effect of smoking status is omitted when modeling covariate effect. One may encode a patient's smoking behavior as smoker and non-smoker and then use this variable as a covariate effect. However, since smoking behavior is dynamic as people smoke and stop smoking at different times, it is not suitable to encode smoking behavior as static covariate in our model. Smoking effect may be better modeled when we can estimate the amount of time of one's exposure to smoking but it seems nontrivial when extracting this information only from health records. In future work, we plan to account smoking into the model as a more dynamic feature which can change over time.

Another difficulty when running the model is to choose the prior distributions for parameters. Although a conservative distribution can be made with assumption of no prior information, we can further improve the model in future version by adding more expert knowledge about distributions of some parameters.

In addition, when choosing a cohort set of patients who have enough data with high quality for analysis, the process of filtering un-qualified data removed a substantial amount of data. Moreover, removing patients with little data also introduces selection bias as the subpopulation seems to have better care and have more hospital visits than the general population. A more flexible model should be introduced in the next version so that it can make use of low-quality data for inference as they are a good source of information.

Conclusion

With the ability to collect and normalize data from multiple EHRs, a large amount of longitudinal data can be collected regarding the diagnosis, severity, and natural history of chronic diseases in patients with multiple co-morbidities. This collection of data is efficient and relatively inexpensive. In essence, research data becomes a byproduct of routine clinical care. Another advantage of these datasets is that they are clinical data of real world patients, which is very useful for pragmatic clinical trials.[2] These large datasets can begin to answer some very clinically pertinent questions. In this study, we analyzed a CKD natural history dataset extracted from the DARTNet database and identified five deep phenotypes of CKD trajectories in patients with CKD using the Probabilistic Subtyping Model.

From the perspective of modeling disease progression, the Probabilistic Subtyping Model we used in this paper can be further expanded to cope with more clinical factors in modeling disease, such as medical information, which are abundantly available in EHR dataset. In addition, a joint probabilistic model which uses more than one clinical health marker can also be a possible extension in future research.

References

1. Ashley EA. The precision medicine initiative: a new national effort. JAMA. 2015 Jun 2; 313(21): p. 2119-20.
2. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA. 2003 Sep 24; 290(12): p. 1624-32.
3. Robinson PN. Deep phenotyping for precision medicine. Human mutation. 2012 May 1; 33(5): p. 777-80.

4. Pace WD, Fox CH, 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(2): p. 6.
5. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The electronic medical records and genomics (eMERGE) network: past, present, and future. Genetics in Medicine. 2013 Jun 6; 15(10): p. 761-71.
6. Naghavi M, Wang H, Lozano R, Davis A, Liang X, Zhou M, 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. 2015; 385(9963): p. 117-71.
7. Jha V, Garcia-Garcia G, Iseki K, Li Z, Naicker S, Plattner B, et al. Chronic kidney disease: global dimension and perspectives. The Lancet. 2013 Jun 26; 382(9888): p. 260-72.
8. Drawz PE, Archdeacon P, McDonald CJ, Powe NR, Smith KA, Norton J, et al. CKD as a model for improving chronic disease care through electronic health records. Clinical Journal of the American Society of Nephrology. 2015 Aug 7; 10(8): p. 1488-99.
9. Inker LA, Astor BC, Fox CH, Isakova T, Lash JP, Peralta CA, et al. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for the evaluation and management of CKD. American Journal of Kidney Diseases. 2014 May 31; 63(5): p. 713-35.
10. Hagar Y, Albers D, Pivovarov R, Chase H, Dukic V, Elhadad N. Survival analysis with electronic health record data: Experiments with chronic kidney disease. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2014 Oct 1; 7(5): p. 385-403.
11. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. Journal of biomedical informatics. Journal of biomedical informatics. 2015 Dec 31; 58: p. 158-65.
12. Schulam P, Wigley F, Saria S. Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery. In AAAI; 2015. p. 2956-2964.
13. Anderson HD, Pace WD, Brandt E, Nielsen RD, Allen RR, Libby AM, et al. Monitoring suicidal patients in primary care using electronic health records. The Journal of the American Board of Family Medicine. 2015 Jan 1; 28(1): p. 65-71.
14. Hissett J, Folks B, Coombs L, LeBlanc W, Pace WD. Effects of changing guidelines on prescribing aspirin for primary prevention of cardiovascular events. The Journal of the American Board of Family Medicine. 2014 Jan 1; 27(1): p. 78-86.
15. Rule AD, Larson TS, Bergstralh EJ, Slezak JM, Jacobsen SJ, Cosio FG. Using serum creatinine to estimate glomerular filtration rate: accuracy in good health and in chronic kidney disease. Annals of internal medicine. 2004 Dec 21; 141(12): p. 929-37.
16. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Annals of internal medicine. 2009 May 5; 150(9): p. 604-12.
17. Li L, Astor BC, Lewis J, Hu B, Appel LJ, Lipkowitz MS, et al. Longitudinal progression trajectory of GFR among patients with CKD. American Journal of Kidney Diseases. 2012 Apr 30; 59(4): p. 504-12.
18. Rasmussen CE. Gaussian processes for machine learning. 2006.
19. Ray L, Nanda SK, Chatterjee A, Sarangi R, Ganguly S. A comparative study of serum aminotransferases in chronic kidney disease with and without end-stage renal disease: Need for new reference ranges. International Journal of Applied and Basic Medical Research. 2015 Jan; 5(1): p. 31.
20. Tomasello S. Secondary hyperparathyroidism and chronic kidney disease. Diabetes Spectrum. 2008 Jan 1; 21(1): p. 19-25.
21. Keane WF, Zhang Z, Lyle PA, Cooper ME, deZeeuw D, Grunfeld JP, et al. Risk scores for predicting outcomes in patients with type 2 diabetes and nephropathy: the RENAAL study. Clinical journal of the American Society of Nephrology. .
22. Johnson ES, Smith DH, Thorp ML, Yang X, Juhaeri J. Predicting the risk of end-stage renal disease in the population-based setting: a retrospective case-control study. BMC nephrology. 2011 May 5; 12(1): p. 1.
23. Orth SR, Hallan SI. Smoking: a risk factor for progression of chronic kidney disease and for cardiovascular morbidity and mortality in renal patients—absence of evidence or evidence of absence? Clinical Journal of the American Society of Nephrology. 2008 Jan 1; 3(1): p. 226-36.