CVD prediction of CKD Patients using Window-based Laboratory Tests and Drug Usage Covariates

Akira Koseki[1], Reitaro Tokumasu[1], Daijo Inaguma[2]
IBM Research – Tokyo [1]
Department of Internal Medicine, Fujita Health University Bantane Hospital [2]
{akoseki, rtoku}@jp.ibm.com, daijo@fujita-hu.ac.jp

Abstract

Modeling disease progression is the key to the understatings of diseases and the development of efficient clinical treatments. By virtue of recent advancement of machine learning technologies, a lot of models have been studied to predict and explain the status of disease and find important factors. One of the important aspects of those modelings is designing covariates. While many studies use non-temporal covariates, an important scope lies in examining temporal structures of covariates such as tendency, long-term influence, and others. This paper discusses cardiovascular disease (CVD) prediction and explanation of chronic kidney disease (CKD) patients using such temporal structures of covariates of laboratory tests and drug usages based on time windows.

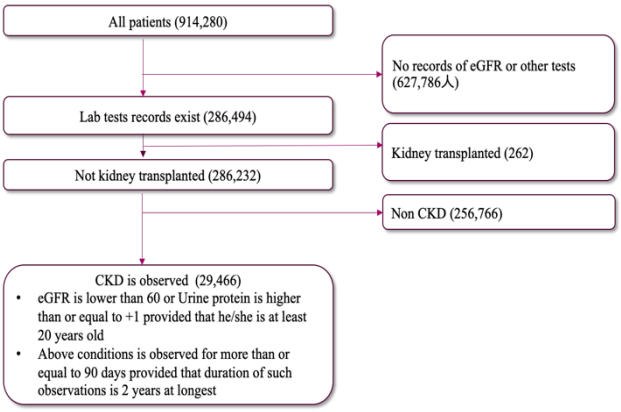
1 Introduction

Growing computing power and digital accumulation of electrical medical record (EMR) enabled to find complicated statistical relations among medical data. Various applications have been found in medicine, pathology, diagnostic and others [Niel, 2019; Hamet, 2017; Johnson, 2018; Yu, 2018; Liu, 2018; Xiao, 2019]. Leveraging such a computational environment, modeling chronic diseases using machine leaning technology has been widely studied using large-scale cohorts. Chronic kidney disease (CKD) is one of the chronic diseases and its complications in later CKD stages largely affect the lifestyle. Among such complications, cardiovascular disease (CVD) is regarded as serious one so creating predictive or explanatory models which identify important risk factors with high classification performance is definitely beneficial for patients and physicians. This paper is thus specifically motivated in creating such models leveraging temporal information of lab tests and drug usages. A patient suffering from a chronic disease such as CKD repeatedly takes medical checks and intakes drugs for a long term so the EMR holds such important temporal information to be analyzed with machine learning models. To this end, we tried an intriguing approach where such temporal information is summarized by computing statistics for several pre-defined time windows, as forming explanatory variables for machine learning models. Using Random Forest and other interpretable models, we successfully found the risk factors for CVD onset taking time structure into account, with obtaining high AUC scores.

2 Background and Motivation

Growing computing power and digital accumulation of electrical medical record (EMR) enabled to find complicated statistical relations among medical data. Various applications have been found in medicine, pathology, diagnostic and others [Niel, 2019; Hamet, 2017; Johnson, 2018; Yu, 2018; Liu, 2018; Xiao, 2019]. The major purposes of those applications include the construction of predictive and explanatory models of the target outcome such as an onset of a certain disease or medical condition changes using the explanatory variables as well as to detect the importance of such variables as to how they affect the outcome. Leveraging such a computational environment, modeling chronic diseases has been widely studied using large-scale cohorts. Chronic kidney disease (CKD) is one of the chronic diseases and its complications in later CKD stages resulting in dialysis largely affect the lifestyle. Risk analyses of the kidney function have been then conducted to identify several conditions including proteinuria, hypertension, and comorbidity of diabetes, which are related to estimated glomerular filtration rate (eGFR) declines [Yang, 2014; Inaguma, 2017; De Nicolca, 2015; Toto, 2010]. Several clinical trials also identified renin angiotensin system blockers and other related drugs are related to control eGFR decline [Brenner, 2001; Lewis, 2001; Wanner, 2018; Perkovic, 2019]. Motivated by such studies, this paper discuses models to predict and explain cardiovascular disease (CVD) onset of CKD patients, which is one of the serious complications of CKD. Our interests lie in how temporal status of human body measured by lab tests, and temporal usage of CKD-related drugs are related to CVD onsets. To this end, we tried an intriguing approach where such temporal information is summarized by computing statistics for several predefined time windows, as forming explanatory variables for machine learning models. For high interpretability, we don't adapted recent method processing temporal date using RNN and generative point process models, which are worth studying further [Du, 2016; Xiao, 2017]. Using interpretable machine learning models, we identify the importance of such temporal information taking time structures into account, as well as obtaining high classification performance.

Figure 1: Flow of extracting CKD patients from EMR



3 Experiments Design

This section describes our design of experiments regarding necessary dataset construction and problems to be solved as well as other analytics details.

3.1 Cohort Creation

We used a real-world EHR dataset of one of the largest Japanese hospitals [Inaguma, 2020]. For our experiments, we first constructed a cohort of chronic kidney disease (CKD) patients as shown in Figure 1. Note that estimated glomerular filtration rate (eGFR), a key factor of determining the CKD stages, is largely affected by kidney transplant so we excluded such transplanted cases. Using a usual CKD determination criteria, we lastly constructed a CKD cohort with 29,466 subjects.

3.2 Problem Definition

Using the CKD cohort, we then defined the outcome, forming the dependent variable of our analytical models. Our interests lie in modeling cardiovascular disease (CVD) so we counted the number of CVD appearance after CKD onset. We defined CVD onset if a patient is diagnosed as hear failure diseases listed in DPC system [JMHLW, 2018]. Among 29,466 CKD patients, 1,814 are diagnosed as CVD after CVD onset. We thought CVD appearance mechanisms are quite different if the time to CVD is largely different so we take CVD onset within 5 years as our target, which amounts to 1,277, constructing our CVD-labeled samples. For extracting non-CVD subjects, we considered two groups. One group consists of the patients who are diagnosed as CVD more than 5 years after CKD onset, while the other group consists of the patients who are not diagnosed as CVD and medical records exist at least 5 years after CKD onset. Note that, in the second group, we excluded the CKD patients who are censored with not being diagnosed as CVD within 5 years. The former amounts to 537, while the latter comes to 7,438, constructing our non-CVD-labeled samples, 7,975 in total.

Our problem is then to create models to predict whether a patient is diagnosed as CVD within 5 years from CKD onset.

In this paper, we study two models. One is a predictive model which classifies CVD-samples from non-CVD samples mostly using lab test time-series features observed before CKD onset. The other is an explanatory model which uses the label and features which are the same as above with drug time-series features before CVD added additionally. Note that the former model is to predict the future CVD appearance using past lab test results while the latter model is focusing on how recent drug usages explain the CVD onset. Note that the predictive model is to be use of advanced clinical treatment while the explanatory models is to solicit pharmacological interests to directly see relations between drug usages and CVD onsets.

3.3 Feature Engineering and Construction

To construct the features, we make use of the following types of raw data in EMR:

- Demographic information
- Lab test results
- Drug usages

First, the demographic information includes sex, age, and diabetes mellitus episode, which forms the static features used for both models.

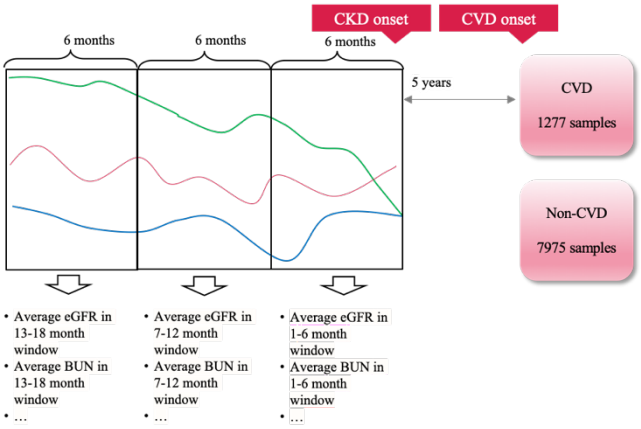
For lab test results, we chose relevant ones which are considered to be related to CKD or CVD as follows:

- Kidney function
  - Estimated glomerular filtration rate (eGFR)
  - Serum creatinine
  - Serum albumin
  - Blood urea nitrogen (BUN)
  - Urine protein
- Anemia indicator
  - Hemoglobin
  - Ferritin
  - Iron saturation
- Lipid, liver function, vital and others

Large category	Small category	# products
Anemia drug	Iron compound preparation	24
	ESA formulation	80
Hypertension or heart failure drug	Cardiac stimulant	159
	Diuretic	100
	ACE	229
	Coronary vasodilator (includes combination products with Ca antagonist and statin)	707
	Beta blocker	105
	Alpha Beta blocker	60
	ARB (includes combination products with diuretic and Ca antagonist)	1039
Dyslipidemia	Statins (includes combination products with small intestinal cholesterol transporter inhibitor and Ca antagonist)	384
	SGLT2 inhibitor	11
Diabetes	DPP-4 inhibitor (includes combination products with insulin resistance improving drug, biguanide formulation, and SGLT2 inhibitor)	45

Table 1: Drug categories

Figure 2: Time-series data processing for predictive model



- Total cholesterol
- Body mass index (BMI)
- Hemoglobin A1c
- C-reactive protein (CRP)
- Systolic blood pressure (SBP)
- Diastolic blood pressure (DBP)

All those test results are recorded as time-series data and we use statistics for couple of time windows as features. Details are explained below.

As for drugs, we have a myriad of product variations including generic products so we categorized those into groups which are considered to be related to CVD as explained in Table 1. Note that category and products are coded using a Japanese drug coding system named YJ code [JAPIC, 2020]. Same as lab tests, such drug usages are recorded as time-series data. The followings describe the details of processing those along to the time windows. We also included some combination drugs, because they contain chemical substances we want to see. As for those combination drugs, the number of products are double counted. Each single drug in the combination agent is used independently as an explanatory variable as explained later.

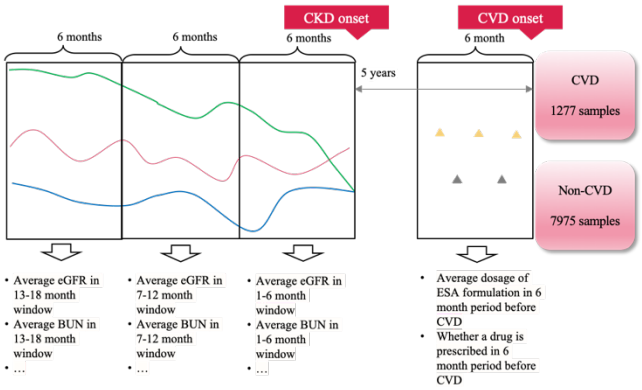
Time-Series Features for CVD Predictive Model

For conducting risk assessment of time-varying features like lab tests taking the time structure into account, this study uses a time-window-based approach. We provide several time windows in which we compute statistics of time-series data, which are separately used as explanatory variables inputted into the models. By doing so, when using interpretable machine learning models, influence, the other word, the risk of such window-based explanatory variables can be illustrated using the importance.

For the CVD predictive model, we constructed time-series features based on three of time windows before CKD onset. Figure 2 depicts such feature formulation. We first make three six-month-long time windows which include the first six-month-long window which ends at CKD onset date, being preceded by second and third six-month long windows. Note that those three windows are disjoint.

For each window, we then compute average values of time-series-data of lab tests results for the kidney function

Figure 3: Time-series data processing for explanatory model



and others listed in Section 3.3, forming 14 x 3 = 42 explanatory variables.

Time-Series Features for CVD Explanatory Model

Additional to the aforementioned lab tests time-series explanatory variables, we consider examining time-series values of drug usages. Same as above we use a time-window approach to summarize such values.

For the CVD explanatory model, we constructed time-series features based on time windows just before CVD onset. Figure 3 depicts such feature formulation, where we use a six-month-long time window just begore CVD onset and compute statistics in it. For the large category in Table 1, except for anemia drugs, we counted prescriptions of drugs in each category in the window, forming 4 explanatory variables. For the small category, except for ESA formulation, we also counted prescriptions of drugs in each category in the window. We included combination products so we also counted insulin resistance improving drugs, biguanide formulations, small intestinal cholesterol transporter inhibitors, and Ca antagonists. Note that they are counted only when used as combination products as listed in Table 1. In total they come to 15 explanatory variables. Lastly, we used two ESA related explanatory variables. One is ESA dosage amount in the window, while the other is the disease duration of anemia before taking ESA, forming two other variables.

3.4 Model Construction

By assigning positive labels to samples in the CVD group and negative labels to samples in the non-CVD group, we constructed predictive and explanatory models using machine learning algorithms. For explanatory variables, the missing values corresponding to each laboratory test were imputed via the last observation carried forward method. If no data were available for a test, the mean value of the corresponding training data was used instead. Additionally, all the values were standardized.

In the following step, by using the aforementioned covariates and training labels, we applied the Random Forest (RF), Logistic Regression (LR), and Decision Tree using Python scikit-learn library (https://scikit-learn.org/) to create classification models. We optimized the models by fine-tuning the

Figure 4: Important factors of CVD prediction model

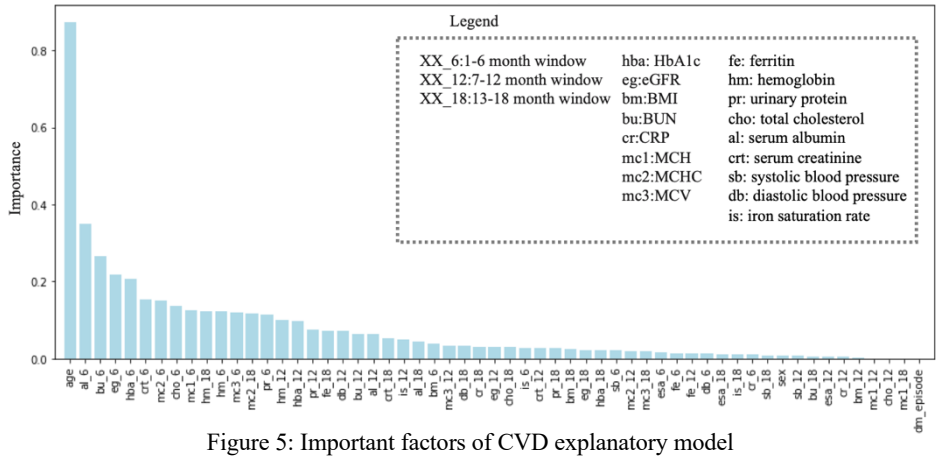
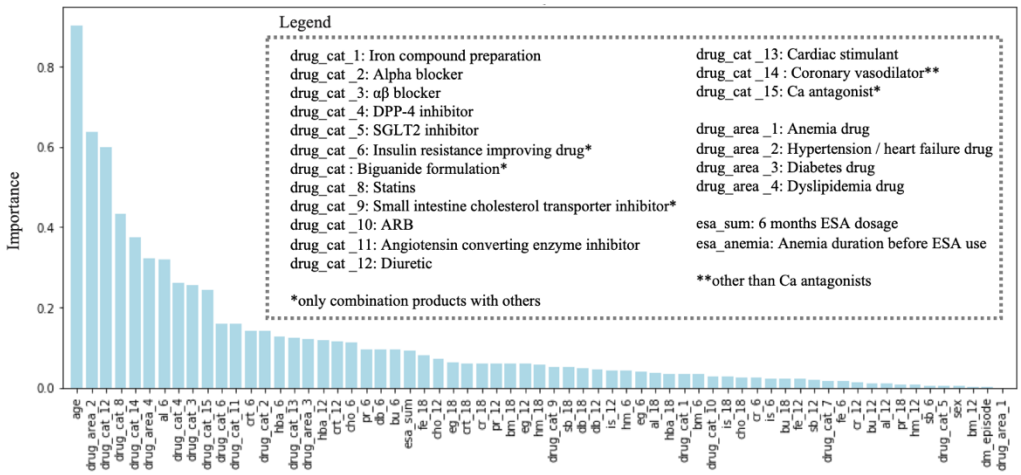


Figure 5: Important factors of CVD explanatory model



hyperparameters of the algorithms. After identifying the optimal parameters via inner four-fold cross validation, we evaluated these models using outer five-fold cross validation.

4 Experimental Results

Using learned models we evaluated those performance. Table 2 shows the classification results when using the best-performing algorithm. Both remarked good enough scores as predictive and explanatory models. Figures 4 and 5 show the resultant important factors of predictive and explanatory models. Note that we used coefficients of the Logistic Regression result taking feature interactions into account. In both, age is marked as the best influential factor.

In Figure 4, other than age, kidney function related factors including BUN, serum albumin, and eGFR in the nearest window marked high for the predictive model. This means at the point of CKD onset, where eGFR marks not so low, the kidney functionality could affect CVD in 5 years. CRP and blood pressure are also marked as risk factors so we could find CVD signs in kidney and heart even at the CKD onset point. As for far windows to CKD onset, we did not observe conspicuous risk factors. We would expect a use of another statistics such as standard deviation, or smaller windows could show additional interesting observations.

	Predictive Model	Explanatory Model
AUC	0.77	0.90
Algorithm	RF	LR

Table 2: Best Prediction Performance

For the explanatory model, the large category of hypertension or heart failure drug marks second important in Figure 5. It is expected that usage of cardiac stimulants and diuretics of getting higher as the risk of CVD is growing. It is remarkable that we observed statins as an important risk factor, which thought to be beneficial to keeping healthy status of coronary vessels.

5 Conclusions

This paper discussed modeling approaches when we consider illustrative results in the presence of time-series variables. In the light of obtaining good performance, our approach is simple and promising to examine temporal structures of time-series data. Our approach also shows good practice to illustrate and explain temporal structures features, especially in lab tests. We can conclude that our results have high interpretability and definitely solicit medical and pharmacological interests. It is also thought that possible variation of window size and statistics metrics would give more intriguing results.

References

[Brenner, 2001] Brenner BM, Cooper ME, de Zeeuw D, Keane WF, Mitch WE, Parving HH, et al. Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. The New England journal of medicine. 2001;345(12):861-9.

[De Nicolca, 2015] De Nicola L, Provenzano M, Chiodini P, Borrelli S, Garofalo C, Pacilio M, et al. Independent Role of Underlying Kidney Disease on Renal Prognosis of Patients with Chronic Kidney Disease under Nephrology Care. PLoS ONE. 2015;10(5):e0127071.

[Du, 2016] Du N, Dai H, Trivedi R, Upadhyay U, Gomez-Rodriguez M, and Song L. Recurrent marked temporal point processes: Embedding event history to vectore. In KDD, 2016.

[Hamet, 2017] Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism: clinical and experimental. 2017;69s:S36-s40.

[Inaguma, 2017] Inaguma D, Imai E, Takeuchi A, Ohashi Y, Watanabe T, Nitta K, et al. Risk factors for CKD progression in Japanese patients: findings from the Chronic Kidney Disease Japan Cohort (CKD-JAC) study. Clinical and experimental nephrology. 2017;21(3):446-56.

[Inaguma, 2020] Inaguma D, Kitagawa A, Yanagiya R, Koseki A, Iwamori T, Kudo M, et al. Increasing tendency of urine protein is a risk factor for rapid eGFR decline in patients with CKD: A machine learning-based prediction model by using a big database. PLoS ONE. 2020;15(9): e0239262.

[Johnson, 2018] Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial Intelligence in Cardiology. Journal of the American College of Cardiology. 2018;71(23):2668-79.

[JMHLW, 2018] Japanese Ministry of Health, Labor and Welfare. Overview of DPC/PDPS. https://www.mhlw.go.jp/file/06-Seisakujouhou-12400000-Hokenkyoku/0000197983.pdf

[JAPIC, 2020] Japan Pharmaceutical Information Center, YJ code. https://www.japic.or.jp/service/iyaku/iyaku_name.html

[Lewis, 2001] Lewis EJ, Hunsicker LG, Clarke WR, Berl T, Pohl MA, Lewis JB, et al. Renoprotective effect of the angiotensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. The New England journal of medicine. 2001;345(12):851-60.

[Liu, 2018] Liu Y, Zhang Y, Liu D, Tan X, Tang X, Zhang F, et al. Prediction of ESRD in IgA Nephropathy Patients from an Asian Cohort: A Random Forest Model. Kidney & blood pressure research. 2018;43(6):1852-64.

[Niel, 2019] Niel O, Bastard P. Artificial Intelligence in Nephrology: Core Concepts, Clinical Applications, and Perspectives. American journal of kidney diseases : the official journal of the National Kidney Foundation. 2019;74(6):803-10.

[Perkovic, 2019] Perkovic V, Jardine MJ, Neal B, Bompoint S, Heerspink HJL, Charytan DM, et al. Canagliflozin and Renal Outcomes in Type 2 Diabetes and Nephropathy. The New England journal of medicine. 2019;380(24):2295-306.

[Toto, 2010] Toto RD, Greene T, Hebert LA, Hiremath L, Lea JP, Lewis JB, et al. Relationship between body mass index and proteinuria in hypertensive nephrosclerosis: results from the African American Study of Kidney Disease and Hypertension (AASK) cohort. American journal of kidney diseases : the official journal of the National Kidney Foundation. 2010;56(5):896-906.

[Yang, 2014] Yang W, Xie D, Anderson AH, Joffe MM, Greene T, Teal V, et al. Association of kidney disease outcomes with risk factors for CKD: findings from the Chronic Renal Insufficiency Cohort (CRIC) study. American journal of kidney diseases : the official journal of the 358 National Kidney Foundation. 2014;63(2):236-43.

[Wanner, 2018] Wanner C, Heerspink HJL, Zinman B, Inzucchi SE, Koitka-Weber A, Mattheus M, et al. Empagliflozin and Kidney Function Decline in Patients with Type 2 Diabetes: A Slope Analysis from the EMPA-REG OUTCOME Trial. Journal of the American Society of Nephrology : JASN. 2018;29(11):2755-69.

[Xiao, 2019] Xiao J, Ding R, Xu X, Guan H, Feng X, Sun T, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. Journal of translational medicine. 2019;17(1):119.

[Yu, 2018] Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature biomedical engineering. 2018;2(10):719-31.

[Xiao, 2017] Xiao S, Farajtabar M, Ye X, Yan J, Song L, and Zha H. Wasserstein learning of deep generative point process models. In NIPS, 2017.