ポスター | 知識工学／医療データ解析
ポスター1
知識工学／医療データ解析
2019年11月22日(金) 14:50 ～ 16:20 ポスター会場2 (国際展示場 展示ホール8)

# [2-P2-1-04] Predicting risk of complication in T2DM: a temporal phenotyping approach to detect risk of Diabetic Nephropathy

○Andri Malfian Labiro[1], Kazuya Okamoto[1,2], Purnomo Husnul Khotimah[2], Shusuke Hiragi[1,2], Osamu Sugiyama[2], Goshiro Yamamoto[2], Tomohiro Kuroda[1,2] （1. Graduate School of Informatics, Kyoto University, 2. Kyoto University Hospital）

キーワード：Phenotyping, Machine Learning, Electronic Medical Records, Type 2 Daibetes Mellitus (T2DM)

Electronic Medical Records (EMRs) contain a lot of information about patients’ medication history. Such information can be used for clinical research to discover the risk of complications from diseases, especially for long-term treatment of chronic diseases such as Type 2 Diabetes Mellitus (T2DM). In T2DM, if the symptoms are detected earlier, these complications can be prevented or delayed. Screening process by manual chart review takes a lot of time, some patients might be overlooked and patients’ characteristics might change along with age. To improve the screening process, we propose a method by temporal analysis of patients’ phenotypes to predict risk of complication using EMRs data. The purpose of this research is to identify the feasibility of predicting risk of diabetic nephropathy from long-term treatment of T2DM by using temporal phenotyping. In this approach, we use structured EMRs data of T2DM patients from Kyoto University Hospital. To observe the change in the phenotypes, we divide patients' medication history from prescription data into episodes by reconstructing medication history according to the notion of stable period. We apply rule-based algorithm to each medication episode for classifying patients, according to phenotypes associated with diabetic nephropathy and apply machine learning method to predict the possible outcome of future phenotypes.

# Predicting risk of complication in T2DM: A temporal phenotyping approach to detect risk of diabetic nephropathy

Andri Malfian Labiro[1] Kazuya Okamoto[1,2] Purnomo Husnul Khotimah[2,3] Shusuke Hiragi[1,2]
Osamu Sugiyama[2] Goshiro Yamamoto[2] Tomohiro Kuroda[1,2]
[1]Graduate School of Informatics, Kyoto University, Japan
[2]Kyoto University Hospital, Japan
[3]Indonesian Institute of Sciences, Indonesia

Abstract: In this study, we proposed a method for predicting patients' future condition using patients' phenotype sequence and machine learning. We generated phenotypes from patients EMRs by using a rule-based phenotyping algorithm. Long-term prescription data were used and reconstructed according to the notion of stable period to determine the timestamps where the phenotyping algorithm will be applied. The phenotype sequence was then fed into a CNN-LSTM based machine learning model for one-step prediction of the phenotype. We conducted two experiments, using a single parameter and multiple parameters including phenotypes and medication type. We obtained prediction accuracy of 91.3% and 92.1% respectively. The results show the feasibility of predicting risk of diabetes complication by compressing multiple features from raw data into a single feature (EMR phenotype). Using EMR phenotypes as a feature reduced the feature space and it is possible for the model to be reproduced across different EMRs.

Keywords: Phenotyping, Machine Learning, Electronic Medical Records, Type 2 Diabetes Mellitus (T2DM)

---

## 1. INTRODUCTION

Electronic Medical Records (EMRs) have been widely used for the purpose of storing patient treatment information into electronic database. This information can be used for research in healthcare for cohort study, predictive analytic and evident-based studies[1]. This is especially useful for a patient that needs long time treatment, for example, a diabetes patient.

Diabetes is a chronic disease marked by high level of glucose in blood plasma, also known as hyperglycemia[2]. It is an uncured condition that lasts for lifelong. However, it can be controlled into remission by suppressing the glucose level using medication or therapy. Among T2MD patients, 20%-40% are likely developed into Diabetic Nephropathy, which makes it one of the major diabetes complications. It is important to grasp and predict the patients' future condition so that medical doctors can plan personalized treatment for the patient to prevent a worse condition.

To grasp patients' condition, EMR phenotyping is effective because it can effectively describe patient characteristics from cohort[3]. Due to the heterogenous trait of diabetes[4], it is hard to describe a patient characteristics using standard definition. Patients may react differently to medications thus create different outcomes. Other risk factors can also contribute to the heterogenity, for example, age, sex, ethnicity, life style, etc. Using EMR phenotype definition, we are able to accurately classify patients according to the disease subtypes based on combination of data elements, events, value sets, etc. The EMR phenotyping method shows a superior results in detecting disease compared to conventional models[5].

This study is aimed to identify the feasibility of predicting risk of diabetic nephropathy by using EMR phenotyping approach and machine learning prediction.

We believe that EMR phenotyping can be used to predict a patients' future condition.

## 2. METHODOLOGY

The framework of our method is divided into four steps. First, we prepare datasets from long-term medications of T2DM patients. Second, we reconstruct the patients' prescription data according to the notion of medication episode construction framework[6][7]. Third, we build a phenotyping algorithm based on diabetes classification from Japan Diabetes Society[8]. The phenotypes are extracted from each of the new episodes and transformed into sequences. Fourth, we develop our prediction model using EMR phenotypes as input and the output is one-step phenotype forecast.

### 2.1 Dataset

In this study, we use T2DM patients' data from Kyoto University Hospital spans from year 2000 to 2019. We identified 15,314 unique patients and 574,594 prescription records in our dataset. The prescription records are selected based on diabetes medication name listed in [7]. In addition, Urinary Albumin Excretion (UAE) and estimated Glucose Filtration Rate (eGFR) are also used for phenotyping.

### 2.2. Episode reconstruction

Firstly, we need to decide timestamps at which we should capture the phenotypes. According to [6], a stable period (*SP*) is a period when the medication is not change for at least a certain threshold, which a medical doctor decides the effectiveness of the medication. Based on this fact, we believe that the phenotypes should be captured for each patients' *SPs*. In order to reconstruct the prescription data into *SP*, we use the framework in [7]. However, an *SP* of chronic patients maybe in long duration. Hence, we divide it into shorter *SPs* (*SP's*). Next, we combine the *SP's* with UAE and eGFR. Fig. 1 illustrates the strategy.

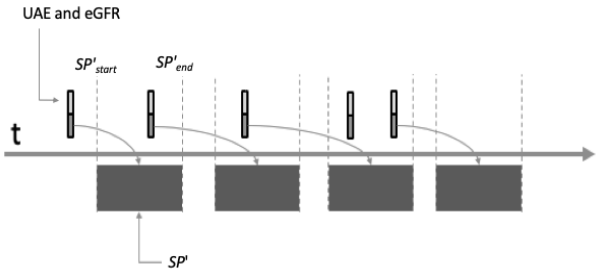


Fig. 1. Combining *SP'* with UAE and eGFR

The values of UAE and eGFR are taken from the date before the start of the current *SP'*. If there are more than one values present, we take the one closest to the current *SP'*. This consideration is taken because the previous patients' conditions (UAE and eGFR) affect the current *SP'*. Using these two features, we capture the phenotypes in each of the *SP'*.

In the implementation, we found that some patients have missing either or both UAE and eGFR in some periods. To tackle this, we fill in the missing value with value from previous period. Patients that do not have both UAE and eGFR test results are excluded from the dataset.

### 2.2 Phenotyping

We classify patients into five phenotypes (P), following the diabetes nephropathy classification rules from Japan Diabetes Society[8] as shown in Table 1.

Table 1. Diabetic nephropathy classification[8]

| Stage/Phenotype | Urinary albumin (mg/g Cr) or urinary protein (g/g Cr) | eGFR (mL/min/1.73 m2) |
|---|---|---|
| Stage 1 (pre-nephropathy) | Normoalbuminuria <30 | ≥ 30 |
| Stage 2 (incipient nephropathy) | Microalbuminuria 30-299 | ≥30 |
| Stage 3 (overt nephropathy) | Macroalbuminuria ≥300 or persistent proteinuria ≥0.5 | ≥30 |
| Stage 4 (kidney failure) | Any albuminuria/proteinuria status | <30 |
| Stage 5 (dialysis therapy) | Any status on continued dialysis therapy | |

Diabetic Nephropathy is divided into five stages of progression, marked by the change of eGFR and urinary albumin level in urine. In our algorithm, we describe the phenotypes as follows: P1 = *pre-nephropathy*, P2 = *incipient nephropathy*, P3 = *overt nephropathy*, P4 = *kidney failure*. We describe patient within normal eGFR range ($60 \leq x < 120$) as P0. We extract phenotypes from each *SP′* and transform it into sequence [$P_1, P_2, P_3, P_4, P_5, P_6, P_7, \ldots P_n$].

2.4. Prediction modeling

Our model is built using Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). The CNN part takes a subsequence as input and split it onto two subsamples, wrapped in *TimeDistributed* layer and feed into LSTM part as input. In this part, the input data is reshaped according to the required structure [X_train, X_train length, timesteps, features]. There are 128 neurons in the model, with kernel size = 1 and activation mode = “relu”. The output is then flattened before we feed it into the LSTM part of the model. Next, we create LSTM part with 100 hidden layers and activation mode = “relu”. We tried dropout rate 20%, 50% and 80%, which resulted in best dropout rate = 20% with best accuracy. We use 1 layer for the fully connected layers as output(Dense).

For the input, we divide patient episodes sequence into six-step subsequences as input and the next step as output. The next subsequence starts from second step and so on until the end of the sequence. The transformation from sequence into subsequences and output is illustrated in Fig. 2.

2.5 Evaluation

For training and testing, we divide the dataset into train, test = 80%, 20%. To avoid the sequences from a patient being split up, we construct the subsequences after we split the dataset into train and test data. To evaluate the model, we use the test dataset that has not been feed to

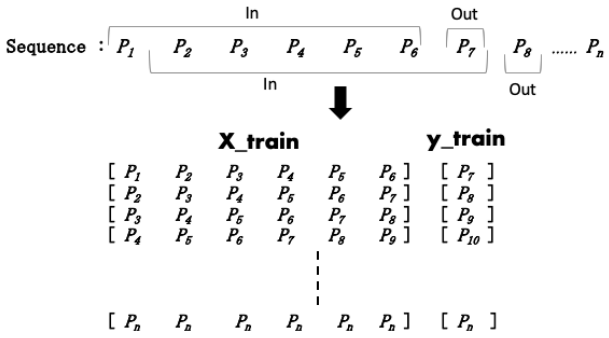


Fig. 2. Transforming phenotypes sequence into subsequences and output

the model, then we compare the prediction values with the true values. We further test the effect of adding more parameters into the prediction model, by using patients’ medication type as parameter, then we compare it with the result of using only a single parameter.

3. RESULTS

We divided raw prescription data into 84,638 *SP′*. We mined phenotypes from each of the *SP′* as follows: P0 = 29,624; P1 = 16,473; P2 = 24,986; P3 = 7,511; P4 = 6,044. From the *SP′*, we constructed 2,113 phenotypes sequences, which we divided further into 74,375 six-steps subsequences. We used 59,188 phenotype subsequences as training data and 15,187 as test data. The cross-validation result during training shows prediction accuracy of 91.32%. Then, using 15,187 sequences as test data, we evaluated the model. It shows accuracy of 91.3%. We compared the predicted value and the true value. The result is shown in Table 2.

Table 2. Predicted values vs true values

| | | Predicted values (P') | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | | 1 | | 2 | | 3 | | 4 | | |
| | | n | % | n | % | n | % | n | % | n | % | SUM |
| True values (P) | 0 | 4,919 | 92.93% | 272 | 5.14% | 102 | 1.93% | 0 | 0.00% | 0 | 0.00% | 5,293 |
| | 1 | 254 | 8.08% | 2,757 | 87.75% | 76 | 2.42% | 47 | 1.50% | 8 | 0.25% | 3,142 |
| | 2 | 39 | 0.90% | 43 | 1.00% | 4,072 | 94.24% | 124 | 2.87% | 43 | 1.00% | 4,321 |
| | 3 | 0 | 0.00% | 1 | 0.08% | 55 | 4.25% | 1,225 | 94.74% | 12 | 0.93% | 1,293 |
| | 4 | 1 | 0.09% | 21 | 1.85% | 32 | 2.81% | 133 | 11.69% | 951 | 83.57% | 1,138 |

The comparison shows high accuracy in predicting the phenotypes (P'). For example, P0 with the total values of 5,293, the total correct predicted values are 4,297(93.94%), while the rest are incorrectly predicted as P'1 and P'2 with total values of 275(5.18%) and 47(0.89%) respectively. The largest value of incorrect predictions is ± 1 phenotype, then the values became smaller. Next evaluation result using medication type as parameter is shown in Table 3.

Table 3. Predicted values vs true values (with multiple parameters)

| | | Predicted values (P') | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | | 1 | | 2 | | 3 | | 4 | | |
| | | n | % | n | % | n | % | n | % | n | % | SUM |
| True values (P) | 0 | 4,972 | 93.94% | 274 | 5.18% | 47 | 0.89% | 0 | 0.00% | 0 | 0.00% | 5,293 |
| | 1 | 240 | 7.64% | 2,784 | 88.61% | 101 | 3.21% | 14 | 0.45% | 3 | 0.10% | 3,142 |
| | 2 | 39 | 0.90% | 94 | 2.18% | 4,078 | 94.38% | 84 | 1.94% | 26 | 0.60% | 4,321 |
| | 3 | 0 | 0.00% | 1 | 0.08% | 90 | 6.96% | 1,196 | 92.50% | 6 | 0.46% | 1,293 |
| | 4 | 0 | 0.00% | 22 | 1.93% | 44 | 3.87% | 114 | 10.02% | 958 | 84.18% | 1,138 |

There were ten types of medication, with each *SP'* consisted one or more types. The result shows a slight increase in the prediction accuracy for all P' except for P'3, where the accuracy is lower when compared to using single feature. The overall obtained prediction accuracy was 92.1%.

## 4. DISCUSSIONS AND CONCLUSIONS

Our model is able to predict patients' future condition by predicting the phenotype one-step ahead from six-steps input. Adding medication type as prediction features slightly increase the accuracy. This result shows that adding multiple features did not give significant impact to the prediction result. However, we have not exhaustively conducted investigation on this matter.

As for our phenotyping algorithm, currently, it only considers two urine test results (UAE and eGFR). In the future, more factors will be incorporated to make the algorithm reflects the patients' condition more precise. We would like to investigate our hypothesis that even though more phenotyping parameters were added, the prediction model does not have to be changed. Hence, it would lead to a re-usable model. Using this method, we believe the prediction model could be applied across different EMRs[9].

## 5. ACKNOWLEDGMENT

The first author is supported by Indonesia Endowment Fund for Education (LPDP).

## 6. REFERENCES

[1] Cowie, Martin R., et al. "Electronic health records to facilitate clinical research." *Clinical Research in Cardiology* 106.1 (2017): 1-9.

[2] Committee of the Japan Diabetes Society on the Diagnostic Criteria of Diabetes Mellitus, et al. "Report of the committee on the classification and diagnostic criteria of diabetes mellitus." *Journal of diabetes investigation* 1.5 (2010): 212-228.

[3] Richesson, Rachel, and Michelle Smerek. "Electronic health records-based phenotyping." *Rethinking clinical trials: A living textbook of pragmatic clinical trials* 2016 (2014).

[4] Tuomi, Tiinamaija, et al. "The many faces of diabetes: a disease with increasing heterogeneity." *The Lancet* 383.9922 (2014): 1084-1094.

[5] Anderson, Ariana E., et al. "Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study." *Journal of biomedical informatics* 60 (2016): 162-168.

[6] Khotimah, Purnomo Husnul, et al. "Revealing oral medication patterns from reconstructed long-term medication history of type 2 diabetes." *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016.

[7] Khotimah, Purnomo Husnul, et al. "Medication Episode Construction Framework for Retrospective Database Analyses of Patients With Chronic Diseases." *IEEE journal of biomedical and health informatics* 22.6 (2017): 1949-1959.

[8] Haneda, Masakazu, et al. "A new classification of diabetic nephropathy 2014: a report from joint committee on diabetic nephropathy." *Journal of diabetes investigation* 6.2 (2015): 242-246.

[9] Chen, You, et al. "Building bridges across electronic health record systems through inferred phenotypic topics." *Journal of biomedical informatics* 55 (2015): 82-93.