The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Prediction of the Onset of Lifestyle-related Diseases Using Regular Health Checkup Data

Mitsuru Tsunekawa[*1] Natsuki Oka[*1] Masahiro Araki[*1] Motoshi Shintani[*2]
Masataka Yoshikawa[*3] Takeshi Tanigawa[*4]

[*1] Kyoto Institute of Techonology [*2] SG Holdings Group Health Insurance Association
[*3] Japan System Techniques Co.,Ltd. [*4] Juntendo University

This study proposes a method for predicting the onset of lifestyle-related diseases using periodical health checkup data. We carefully examined insurance claims data to identify the onsets of the diseases and used them as correct answers for supervised learning. We adopted the undersampling and bagging approach to address the class imbalance problem. We aimed to predict whether lifestyle-related diseases, other than cancer, will develop within one year. The precision and recall of the proposed method were 0.32 and 0.89, respectively. Compared with a baseline that sets thresholds for each examination item and considers their logical sum, it was found that much higher precision could be obtained while maintaining recall, which is meaningful as it allows for the suppression of the number of targets for health guidance, without increasing the negligence of those that are likely to become severely ill.

1. Introduction

Many people have recently begun using Internet mail-order sales, greatly increasing the number of deliveries. Consequently, the social interest in the work environment and health management of courier drivers is increasing. If appropriate health guidance can reduce the occurrence of drivers' lifestyle-related diseases and prevent severe idiopathic illnesses during driving, medical expenses and traffic accidents can be decreased. We, therefore, aimed to use drivers' regular health checkup data to predict the onset of lifestyle-related diseases as accurately as and ensure provision of appropriate health guidance.

Many studies have used machine learning and data mining techniques to predict disease onset from medical data. For example, [Weng 17] highlighted the superiority of machine learning techniques to predict cardiovascular risk from routine clinical data; [Yatsuya 16] predicted the occurrence probability of myocardial infarction or cerebral infarction using health examination results. Morevoer, [Uematsu 17] proposed a model that predicts pneumonia hospitalization using the Lasso logistic regression of regular health checkup data, which is similar to our research model in that it tries to predict using periodical health checkup data of healthy people. Our study objective was to predict whether lifestyle-related diseases, other than cancer, will develop within one year, using regular medical examination data of Courier drivers.

2. Target Data

2.1 Data overview

We used insurance claims and regular health checkup data of employees from the SG Holdings Group Health Insurance Association. Health insurance claims data is created when a person is injured or ill and visits a medical institution, whereas health checkup data is taken regularly (typically once a year). The two datasets were anonymized and linked with a hash code for uniquely identifying patients. In this study, disease onsets were extracted from the insurance claims data and used as correct answers for onset prediction; the health checkup data is used as input for onset prediction.

Health insurance claims data includes disease name codes and medical examination dates etc. Details of the health checkup data are provided later.

In this study, we analyzed the insurance claims data from 1996 to 2017 and the health checkup data from 2006 to 2018, of individuals aged 15-74 years; the total health checkup data was 961,906 sheets for 156,145 people, and the total insurance claims data was 1,617,078 sheets for 108,581 people.

2.2 Disease names as prediction targets

An individual's diagnosed diseases names were obtained by searching through the disease name codes included in the health insurance claims data and corresponded with those in the ICD-10. Table 1 presents the ICD-10 codes and the corresponding disease names to be predicted (hereinafter, referred to as "severe disease names").

2.3 Feature values used for prediction

The following examination items of health checkup data were used as feature values for prediction: Health examination data included not only the numerical data of inspection results, but also the results of a questionnaire on lifestyle habits, and the judgment results of six levels derived from the examination data by medical institutions. Items of abdominal girth and visual acuity judgment, heart rate, visual acuity judgment, fundus judgment, and metabolic judgment were removed, since the ratio of missing values was

Contact: Mitsuru Tsunekawa, Kyoto Institute of Technology, and E-mail: m-tsune@ii.is.kit.ac.jp

Table 1: ICD-10 codes of severe disease names.

ICD-10	disease name
E10	Insulin dependent diabetes mellitus
E11	Non-insulin dependent diabetes mellitus
E14	Diabetes mellitus other than the above
I20	Angina pectoris
I21, I22	Acute myocardial infarction
I42	Cardiomyopathy
I44 I49	Arrhythmia, Conduction defects
I60, I690	Subarachnoid hemorrhage
I61, I691	Intracerebral hemorrhage
I63, I693	Cerebral infarction

50% or more. The health examination data also included findings freely described by doctors; however, we excluded them because natural language understanding is necessary for use of free description.

Health examination data items used as input features are as follows:
Sex; Age; Height; Weight; Body fat percentage; Systolic blood pressure; Diastolic blood pressure; Number of red blood cells; Hemoglobin; Hematocrit; Platelet count; GOT; GPT; γ-GTP; Total cholesterol; HDL cholesterol; LDL cholesterol; Neutral fat; Uric acid; Creatinine; eGFR; HbA1c; Questions about medicine to lower blood pressure, insulin injection, or medicine to lower blood sugar, medicine to ameliorate dyslipidemia, stroke, chronic renal failure, anemia, smoking habits, weight change from the age of 20 years, exercise habits, walking habits, walking speed, weight change over the past year, eating speed, meal just before going to bed, after dinner snacks, skipping breakfast, drinking habits, drinking alcohol amount, sleeping time, willingness to improve lifestyle habits, and willingness to receive health guidance; Judgments on urinary protein and urine sugar; Representative judgment; Judgments on physical measurements, hearing ability, blood pressures, anemia, liver function, renal function, uric acid and gout, blood sugar, sugar metabolism, and urinalysis; Examination judgment.

2.4 Characteristics of data

The data typically have two characteristics. First, considerable imbalance: For example, in 2017, the proportion of people diagnosed with severe diseases was only 4.5%. The learning of such unbalanced data may be greatly affected by the properties of a large number of negative examples, persons who are not diagnosed with severe diseases. Therefore, a method that can successfully learn this imbalanced data must be adopted.

Second, classifying data as positive or negative is not easy. In this study, we aimed to predict whether a person who is healthy at the time of a regular health checkup will be diagnosed with one of the severe diseases within a year of the checkup. Consequently, data was positive if the person will fall sick within a year, and negative if not. Therefore, it was necessary to accurately judge the presence or absence of illness at the time of a health checkup.

The point at which the target disease name first appeared in an employee's insurance claims data was not necessarily the point when he/she first developed the disease. It is not uncommon for individuals with previously diagnosed diseases to join a health insurance association in an industry with large personnel flow. However, because the data used in this study belonged to a health insurance association, it was only available for the period of joining the association; the insurance claims data before entering the association could not be confirmed. The next section describes how we addressed this problem.

3. Data selection and machine learning method

3.1 Data selection

We addressed the classification problem of predicting whether individuals will suffer severe illness within one year, by using medical examination data. The data selection method for positive and negative data was as follows.

First, to address the previously mentioned data availability problem, we used the following method and determined whether the insurance claims data of an individual's first-time diagnosis of a severe disease was actually first. We first calculated the hospital visit interval for the same disease after receiving the diagnosis of a disease. If the hospital visit interval was shorter than the interval between the day of joining the health insurance association and the day of first-time diagnosis of a severe disease, the diagnosis was judged to be actually first; alternatively, if the visit interval was greater, it was judged not to be first. The visit interval was calculated using three interval data, which we regarded as sufficient. The specific procedure was as follows (see also Figure 1):
1) Extract the oldest data (*) with a severe disease name from an individual's insurance claims data. 2) Select three consecutive data with the same disease name that are newer than the extracted data and calculate the hospital visit intervals. 3) Retrieve the individual's oldest insurance claims data (**) and calculate the interval between data (**) and data(*). 4) If the maximum of the three values calculated in "2)" is smaller than that calculated in "3)," regard data (*) as the first-diagnosis data of the disease.

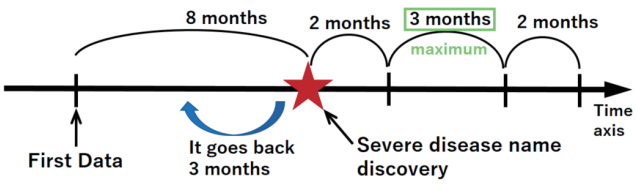

Figure 1: Method of judging the genuineness of the first diagnosis.

Next, the positive health checkup data was chosen from the data included within the range of one year or less before the first appearance of the severe disease. When there were multiple data in the range, we adopted the oldest one. We also considered the amount of changes in the health checkup

data, calculated the differences between the chosen data and the previous data and between the chosen data and the two previous data, and added them to the feature set (Figure 2). Since some items in the health checkup data would have changed with the development of the disease, we considered that the discrimination accuracy improved by explicitly adding the change amount to the feature set.

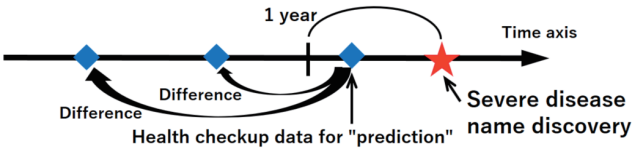

Figure 2: Selection of positive data.

For negative data, however, we excluded data of individuals who had been diagnosed with a severe disease even once and used only the remaining data. Furthermore, if there is no insurance claims data after more than one year from when the health checkup data was extracted, the possibility is that the individual may have been diagnosed with a severe disease within one year of data extraction. This is possible if the person leaves the job. Therefore, to eliminate this possibility, we excluded the data of individuals with no insurance claims data after one year or more from the extracted health examination data. As with the positive data, we calculated the differences using three consecutive health examination data and added them to the feature set (Figure 3).

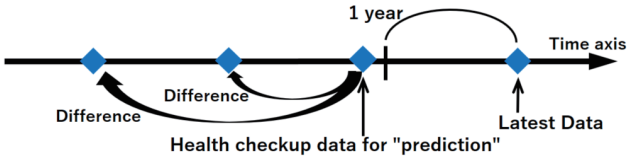

Figure 3: Selection of negative data.

There were cases in which an individual had multiple medical examination data that fit the selection criteria. This was common in both positive and negative data. However, we used only one data per individual to prevent data imbalance. If we did not select the oldest data for positive data, we used data after being diagnosed with one of the severe diseases; however,for negative example, we can select data at any point in time.

Thus, we obtained 1255 positive data and 37664 negative data, with 133 features. The missing values were filled with the median values.

3.2 Machine learning method

In this study, undersampling and bagging [Wallace 11] was adopted as an effective learning method for imbalanced data. Bagging is a method of improving classification accuracy by combining classifiers which are called weak learners. We used decision trees without pruning as classifiers because they were unstable and resulted in higher performance. Undersampling created balanced data.

4. Results and discussion

There were 500 weak learners. Even if the number of weak learners was changed to range between 100 and 500, there was little change in recall and precision; however, if the number was less than 100, the precision decreased. Since the decision tree used for the weak learners was an algorithm not affected by the scale, data scaling was not performed. We used 70% of the dataset for learning and 30% for evaluation. Table 2 presents the confusion matrix of the proposed method. The positive precision and recall were 0.32 and 0.89, respectively.

Table 2: Confusion matrix of the proposed method.

		Predicted class	
		Positive	Negative
Actual class	Positive	334	43
	Negative	700	10600

We used a judgment category table[*1] that was officially released by the Japan Society of Ningen Dock as the baseline method. A threshold value for each item was set and discrimination was carried out by logical OR operation on each item. We only used the items that were common to the input features of this study. The total number of items used was 13. The table classified health checkup data into four categories: No abnormality, Mild abnormality, Follow-up required, and Medical treatment required.

The precision-recall curves of the proposed method and the baseline method are shown in Figure 4. The thresholds of three categories, excluding "No abnormality," were used to plot the precision-recall curve of the baseline method. For the proposed method, we changed the ratio of the positive and negative examples in undersampling as follows: 1:0.25, 1:0.5, 1:1, 1:2, 1:4, 1:8, and 1:16. We added another precision-recall curve of the proposed method in which the number of features was reduced to 13 in order to compare with the baseline method using the same features.

Since the graph of the proposed method lies clearly above that of the baseline method, the proposed method can be considered superior to the baseline method. The results demonstrated that much higher precision could be obtained by the proposed method when the recall was about the same degree as the baseline method; increasing precision while maintaining recall is meaningful as it allows for the suppression of the number of targets for health guidance, without increasing the negligence of those who are likely to become severely ill.

We identified the features that were important for classification. The top 3 were HbA1c, metabolism judgement, and the question about taking insulin injection or a drug that lowers blood glucose. Metabolic judgement refers to the judgment of the danger of metabolic syndrome in six levels from the health checkup data. HbA1c is one of the indicators used to judge diabetes, prescribing insulin injection and medication to lower blood glucose is a treatment

*1 https://www.ningen-dock.jp/wp/wp-content/uploads/2013/09/Dock-Hantei2018-20181214.pdf

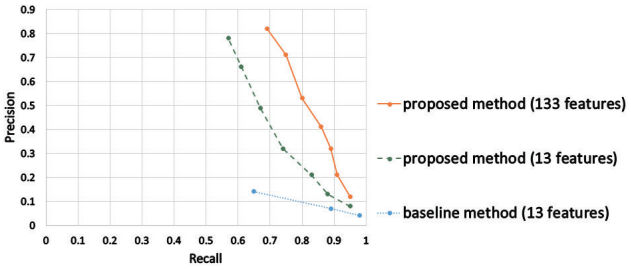

Figure 4: Precision-recall curves of the proposed method and the baseline method.

related to diabetes. Thus, diabetes can be easily distinguished using health checkup data. Among the positive data, the number of diabetes mellitus was as high as 74%; therefore, if identify diabetes, the result would have a high overall accuracy. To confirm this assumption, we focused solely on diabetes and its prediction. Positive cases were defined as persons diagnosed with diabetes mellitus; negative cases comprised patients with severe diseases other than diabetes and healthy people. The created dataset comprised 921 positive and 37998 negative cases. As a result of classification, the positive precision and recall were 0.34 and 0.92, respectively. The confusion matrix is shown in Table 3.

Table 3: Confusion matrix for diabetes prediction using the proposed method.

		Predicted class	
		Positive	Negative
Actual class	Positive	254	23
	Negative	494	10906

Next, to compare with diabetes, we tried to predict the second most frequent angina using the proposed method. As before, positive cases were defined as persons diagnosed with angina pectoris, and negative cases consisted of patients with severe diseases other than angina pectoris and healthy people.

The created dataset comprised 229 positive and 38690 negative cases. As a result of prediction, the positive precision and recall were 0.03 and 0.90, respectively. Table 4 shows the confusion matrix. As evident, the precision reduced and angina pectoris was difficult to discriminate.

Table 4: Confusion matrix for angina pectoris prediction uising the proposed method.

		Predicted class	
		Positive	Negative
Actual class	Positive	62	7
	Negative	1755	9852

5. Conclusion

5.1 Summary

In this study, we proposed a method to predict the onset of lifestyle-related diseases other than cancer, using periodical health checkup data, and a method to select learning data based on the insurance claims data. When all target disease names were identified as positive cases, we obtained positive precision and recall values, 0.32 and 0.89, respectively. Compared to the judgment category table, which the Japan Society of Ningen Dock used as a baseline, it was found that much higher precision could be obtained when the recall was about the same degree.

5.2 Future tasks

In this study, doctors' findings from the health examination data were excluded; however, applying natural language processing to this part will allow the data to be used. As another method of coping with imbalanced data, we plan to use an anomaly detection method that constructs a model using data of healthy people as normal data and detecting data that does not fit the model. We also plan to predict the onset of lifestyle-related diseases after more than one year of a regular medical examination.

A major limitation of this study was that although insulin injection and drugs that lower blood glucose should not be prescribed before receiving a diagnosis of diabetes, the item "Do you take insulin injection or a drug that lowers blood glucose" was used as one of the main items to predict the onset of diabetes. This means that the selection process of positive and negative data need to be reconsidered.

References

[Weng 17] Weng, F. S., Reps, J.,Kai, J., Garibaldi, M. J., and Qureshi, N.: Can Machine-learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?, PLoS One, 12(4), doi:10.1371/journal.pone.0174944 (2017).

[Yatsuya 16] Yatsuya, H., Iso, H., Li, Y., Yamagishi, K., Kokubo, Y., Saito, I., Sawada, N., Inoue, M., and Tsugane, S.: Development of a Risk Equation for the Incidence of Coronary Artery Disease and Ischemic Stroke for Middle-aged Japanese Japan Public Health Center-Based Prospective Study. Circulation Journal, 80(60), 1386-1395 (2016).

[Uematsu 17] Uematsu, H., Yamashita, K., Kunisawa, S., Otsubo, T., and Imanaka, Y.:Prediction of Pneumonia Hospitalization in Adults Using Health Checkup Data, PLoS One, 12(6), doi:10.1371/journal.pone.0180159 (2017).

[Wallace 11] Wallace, C. B.,Small, K., Brodley, E. C., and Trikalinos, A. T.: Class Imbalance, Redux, IEEE 11th International Conference on Data Mining, IEEE Xplore, doi:10.1109/ICDM.2011.33 (2011).