

# Prediction of Key Patient Outcome from Sentence and Word of Medical Text Records

Takanori Yamashita<sup>1</sup>, Yoshifumi Wakata<sup>1</sup>, Hidehisa Soejima<sup>2</sup>, Naoki Nakashima<sup>1</sup>, Sachio Hirokawa<sup>3</sup>

<sup>1</sup>Medical Information Center, Kyushu University Hospital, Fukuoka, Japan

<sup>2</sup>Saiseikai Kumamoto Hospital, Kumamoto, Japan

<sup>3</sup>Research Institute for Information Technology, Kyushu University, Fukuoka, Japan

{t-yama, wakata, nnaoki}@med.kyushu-u.ac.jp

hidehisa-soejima@saiseikaikumamoto.jp, hirokawa@cc.kyushu-u.ac.jp

## Abstract

The number of unstructured medical records kept in hospital information systems is increasing. The conditions of patients are formulated as outcomes in clinical pathway. A variance of an outcome describes deviations from standards of care like a patient's bad condition. The present paper applied text mining to extract feature words and phrases of the variance from admission records. We report the cases the variances of "pain control" and "no neuropathy worsening" in cerebral infarction.

## 1 Introduction

### 1.1 background

Many medical institutes have been accumulating large amounts of medical data. Medical data include structured numerical data and unstructured text data. Unstructured text data is a wide variety of expressions. However, those data are essential, since those free texts are written by medical staff who actually take care of the patients. Therefore, analyzing medical text is expected to improve medical process and the clinical decision support (Meystre, 2008; Zhua, 2013).

There is previous text mining research on medical records. (Mowery, 2012) applied SVM (Support Vector Machine) to partition the emergency reports into SOAP (Weed, 1969) segments. The prediction of the disease or a cancer classification to the discharge summaries was studied in (Suzuki, 2008; Nguyen, 2010). (Codon, 2009) construct the model that automatically populates pertinent parts of a structured cancer representation from text pathology reports. These are mainly classification and performance evaluation. On the other hand, there are not many contents to which specific sentence and word appeared the symptom and the condition are provided.

### 1.2 Clinical pathway

A clinical pathway determines standard medical procedures for an inpatient with respect to each disease and to each medical treatment. This is also expected to improve medical management by advancing standardization. The Japanese Society for Clinical Pathway<sup>1</sup> promotes the construction of a standard electronic clinical pathway aiming at the standardization of medical treatment and improvement in medical processes.

"All variance outcome oriented clinical pathway" is a series of medical treatment units which consist of three layers of (a) outcome, (b) assessment and (c) task (Figure 1). Doctors or nurses in medical practice keep records of their tasks and assessments of patients' conditions. The variance is recorded in an outcome layer when a patient's condition doesn't achieve the criteria of an assessment layer. Thus, we can grasp abnormal condition of the patient and the change of medical intervention plan (Nakashima, 2007).

The present paper applied text mining and machine learning to admission records to extract the words that represent outcome variance (patient condition) and evaluated the prediction performance. Furthermore, we considered the patient condition related to the outcome variance from extracted feature words and sentences.

<sup>1</sup><http://www.jscp.gr.jp>

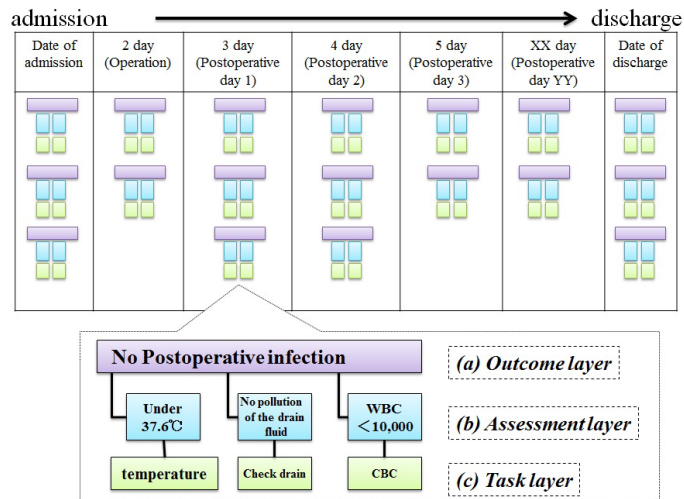


Figure 1: The 3-layer structure of Clinical Pathways

## 2 Data and Method

### 2.1 Admission Text Records

In this paper, we analyzed the admission records of 1,222 patients to whom clinical pathway of cerebral infarction was applied in Kumamoto-Saiseikai Hospital in April 2014 – January 2016.

The clinical pathway of cerebral infarction has set 14 outcomes (Table 1). “no paralysis” and “no depressed level of consciousness” cover the large part of variances. However, we focus in the present paper on “pain control” and “no neuropathy worsening”, since they are considered clinically important. In order to analyze the 1,222 admission records, we constructed a search engine of the textual records. We used GETA<sup>2</sup> system available at NII GETA project. Using this search engine, we tried extraction of the words that may serve as a determinant of outcome variance.

Outcome	Variance count
no Paralysis	1026
no depressed level of Consciousness	734
Dietary intake	522
Vital stable	513
Pain control *	456
no Neuropathy worsening *	356
Circulatory dynamics stable	157
no Urination disorder	133
Respiratory status stable	122
no Chest Infection symptom	14
no Side effect symptom	12
keep Rest	6
no Dyscoria symptom	4
no Imbalance syndrome symptom	1

Table 1: Outcome in clinical pathway of Cerebral Infarction (\*: target in this study)

### 2.2 Classification by Support Vector Machine and Feature Selection

We applied SVM to predict if an admission record is in the outcome variance. The specific procedure is as follows. All admission records are vectorized after morphological analysis using medical dictionary

<sup>2</sup><http://geta.ex.nii.ac.jp/geta.html>

(about 80,000 words). If a patient’s record contains the outcome, it is labeled as positive example. In contrast, the cases which have no mark are used as negative data. Then the classification model is constructed using SVM (SVM-light (Joachims, 1999)).

we applied the model to the imaginary sentence that consists of a single word  $w_i$ . We used the predicted score of the sentence as the  $score(w_i)$  of the word.

The  $score(w_i)$  denotes the SVM score of a word  $w_i$  obtained by applying the model to the imaginary document that contains only the word. In (Sakai, 2012), the  $score(w_i)$  was used for the feature selection. In the present paper, we propose another two measures to evaluate the importance of each word. The first measure  $score(w_i) * df(w_i)$  is obtained as the product of the document frequency  $df(w_i)$  of the word. The second measure  $log(score(w_i) * df(w_i))$  is product of the log of the document frequency of the word and the score. Those measures are defined as “w.o, d.o, l.o”. Furthermore, the measure for which the absolute value was used “w.a, d.a, l.a” was established and 6 measures were used because there was also score of negative in SVM.

Then, we applied the model to all sentences to evaluate the score of each sentences. The top scored sentences were chosen as typical sentences of the outcome variance. We highlighted the feature words in those sentences to help interpreting the meaning of the sentence with focused feature words.

### 3 Result

#### 3.1 Feature Words and Feature Sentences

Table 2 lists the top 30 positive words as feature words for the outcomes of ”pain” and ”neuropathy worsening”. Table 2 shows feature sentences that contain such feature words. There are many sentences of “pain” that contain “dizzy”, “headache”, “nausea”, “fibroid” and “aneurysm”. The sentences of “neuropathy worsening” often shows “paralysis”, “right face” and “difficulty talking”.

Outcome	Feature words	Feature sentences
Pain	dizzy(132), hypalgesia(14), aneurysm(181)*, headache(81)*, nifedipine(67), nausea(61), fibroid(27), right angular(53), calcification(68), pravastatin(23), hemianopsia(94)	dizzy when body move, severe nausea. Feeling badness, a headache and dizziness appear suddenly. found aneurysm in cavernous sinus. anamnesis: fibroid, gallstone(postoperation), high blood pressure, irregular pulse.
Neuropathy worsening	paralysis(205), right face(106), renal failure(60), right knee(10), hypalgesia(14), sick sinus syndrome(15), Right facial paralysis(88), difficulty talking(165), flexion(65)*	When getting up, the paralysis of a right hand finger appeared and was also felt by the right face again. With the paralysis senses in mandibular nerve area of right face. The paralysis sense of the right face, the right forearm and the right thigh back side. Difficulty talking appeared. Forgetfulness and slow talking appeared.

Table 2: Feature words and sentences by SVM (\* possibility or impossibility, presence or absence)

#### 3.2 Feature Selection

The top N of positive words and negative words (or the top 2N of the absolute value) were selected to construct a model, and then we evaluated the prediction performance. We varied the number of words N (N=1,2,⋯,10,20,⋯, 100,200,⋯). We used 5-fold cross validation in the evaluation experiment. The prediction performance was evaluated by Accuracy and F-measure.

The baseline Accuracy of “pain” that uses all words is 0.58. The Accuracy is obtained 0.64 at N=9 (l.o, l.a), and then the best of Accuracy is attained 0.77 at N=700 (w.o, w.a) as shown in Figure 2. The baseline F-measure of “pain” that uses all words is 0.39. The F-measure is obtained 0.55 at N=30, and is attained 0.65 at N=100 (w.o, w.a) and 0.81 at N=700 (w.o, w.a) as shown in Figure 3.

The baseline Accuracy of “neuropathy worsening” that uses all words is 0.61. The Accuracy obtained 0.70 at N=3 (d.o, d.a) and around 0.75 at N=100 (6 measures), and then is attained 0.85 at N=700 (w.o, w.a) as shown in Figure 4. The baseline F-measure of “neuropathy worsening” that uses all words is 0.33. The best of F-measure is attained 0.60 at N=100 (w.o) and 0.75 at N=700 (w.o, w.a) as shown in Figure 5. The measure of Score and Score\*Df made high performance.

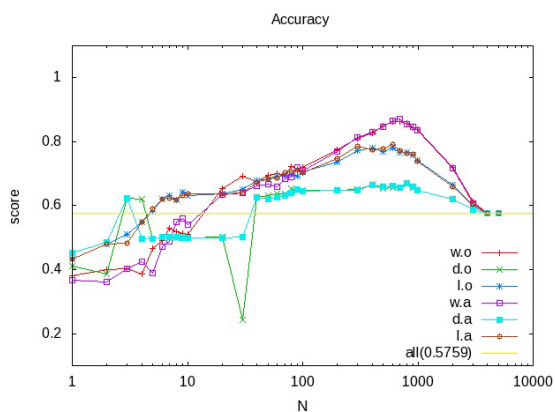


Figure 2: Accuracy(Pain)

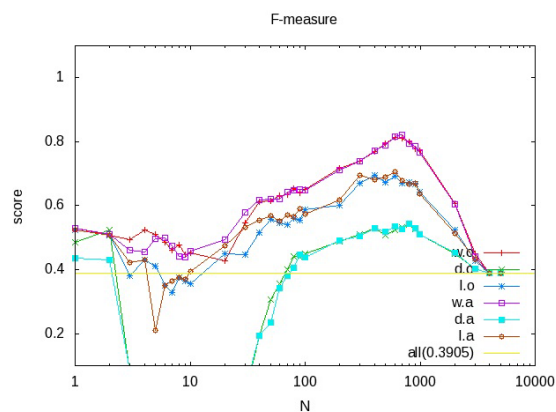


Figure 3: F-measure(Pain)

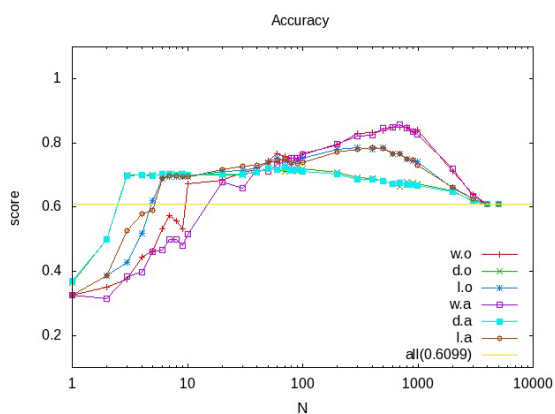


Figure 4: Accuracy(Neuropathy worsening)

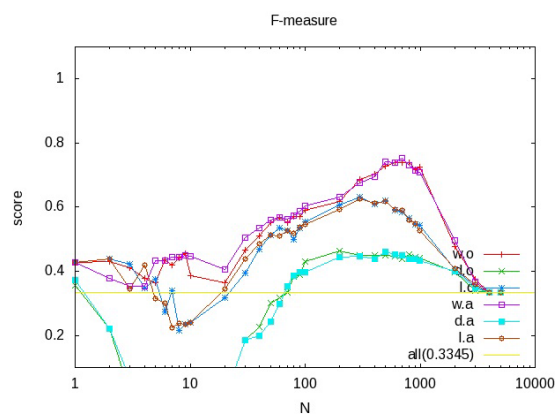


Figure 5: F-measure(Neuropathy worsening)

## 4 Conclusion

This present paper reported the extraction of feature words and typical sentences that describe the patient condition from the free texts. “dizzy”, “headache” and “nausea” are extracted as feature words of “pain”. “paralysis”, “difficulty talking” and “right face” are extracted as feature words of “neuropathy worsening”. These words make sense from a clinical viewpoint. Furthermore, the Accuracy with less than 10 words was better for the prediction performance than F-measure with it by feature selection in both the cases.

In the present paper, we considered feature sentences that contain those feature word and then interpreted context of the sentence. As the result, we succeeded in extracting the part of the patient’s site and the typical condition of the patient from feature words and feature sentences. Then, it will enable early care to critical indicator. We plan to analyze other outcomes and other cases. We aim to establish a method of medical text mining that can perform clinical evaluation for the improvement medical processes.

## Acknowledgements

This research is partially supported by JSPS KAKENHI Grand Number 15H02778

## References

- Meystre SM, Savova GK, Kipper-Schuler KC and Hurdle JF. 2008. *Extracting information from textual documents in the electronic health record: a review of recent research* Yearbook of medical informatics, 128–144.
- Zhua F, Patumcharoenpolc P, Zhanga C, Yanga Y, Chang J, Meechai A, Vongsangnak W and Shen B. 2013. *Biomedical text mining and its applications in cancer research* Journal of Biomedical Informatics, 46(2):200–211.
- Nakashima N, Okada H. and Gouchi A. 2007. *Hospital Information System and Clinical Pathways* Japan journal of medical informatics, 27(1):21–28 (in Japanese).
- Sakai T. and Hirokawa S. 2012. *Feature Words that Classify Problem Sentence in Scientific Article* Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, 360–367.
- Yamashita T, Wakata Y, Hamai S, Nakashima Y, Iwamoto Y, Flanagan B, Nakashima N and Hirokawa S. 2015. *Presumption Model for Postoperative Hospital Days from Operation Records* International Journal of Computer & Information Science, 1(16):50–59.
- Mowery D, Wiebe J, Visweswaran S, Harkema H and Chapman WW 2012. *Building an automated SOAP classifier for emergency department reports* Journal of Biomedical Informatics, 45(1):71–81.
- Suzuki T, Yokoi H, Fujita S and Takabayashi K 2008. *Automatic DPC code selection from electronic medical records: Text mining trial of discharge summary* Methods of Information in Medicine, 47(6):541–548.
- Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE and Colquist S 2010. *Symbolic rule-based classification of lung cancer stages from free-text pathology reports* Journal of the American Medical Informatics Association, 17(4):440–445.
- Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W and de Groen PC 2009. *Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model* Journal of Biomedical informatics, 42(5):937–949.
- Weed LL 1969. *Medical Records, Medical Education, and Patient Care*. Cleaveland Western Reserve University
- Joachims T, 1999. *Making large-scale support vector machine learning practical* Advances in kernel methods, 169–184