Text Mining of Medical Records for Radiodiagnostic Decision-Making

William Claster
Ritsumeikan Asia Pacific University, Beppu, Japan
wclaster@apu.ac.jp

Subana Shanmuganathan and Nader Ghotbi
Ritsumeikan Asia Pacific University, Beppu, Japan
Email: { s5subana, nader }@apu.ac.jp

Abstract-**The rapid growth of digitalized medical records presents new opportunities for mining terra bytes of data that may provide new information & knowledge. The knowledge discovered as such could assist medical practitioners in a myriad of ways, for example in selecting the optimal diagnostic tool from among numerous possible choices. We analyzed the radiology department records of children who had undergone a CT scan procedure at Nagasaki University Hospital in the year 2004. We employed Self Organizing Maps (SOM), an unsupervised neural network based text-mining technique for the analysis. This approach led to the identification of keywords with a significance value within the narratives of the medical records that could predict & thereby lower the number of unnecessary CT requests by clinicians. This is important because, in spite of the valuable diagnostic capacity of such procedures, the overuse of medical radiation does pose significant health risks and staggering cost especially with regard to children.**

Index Terms—**medical informatics, text-mining, data-mining, SOM, Kohonen Networks, Neural Networks**

I. INTRODUCTION

The increased use of medical radiation, especially diagnostic computed tomography (CT) scanning, has raised many concerns regarding the risk of adverse effects. CT procedures conducted without a careful consideration of the possible risk/benefit ratio has come under severe scrutiny, particularly in children [1]. Overuse of diagnostic CT radiation can lead to an increased risk of cancer. Additionally, it may lead to an unnecessary rise in health care costs [2][3][4].

The medical/diagnostic use of radiation in Japan is found to generate the largest share of the annual collective dose of exposure to ionizing radiation [5][6]. Significantly, most of the recent increase in medical radiation exposure has been ascribed to the increased frequency with which physicians request CT scans. Furthermore modern CT scans, with their high resolution capabilities, expose patients to higher doses on a per scan basis [7].

Previous work by researchers at Nagasaki University Hospital examined the justifications for CT scans in the diagnostic work-up of children suspected of either acute appendicitis or possible injuries after minor head trauma [8]. Based on the results of that study two recommendations were developed;

i) guidelines and/or algorithms should be developed in order to limit the usage of CT scanning and
ii)the use of CT scans should be reserved for patients who are reasonably expected to benefit from them.

However, such a stepwise approach to the request of CT diagnostic procedures can hardly be adopted in many other clinical conditions where CT scans are typically requested solely on the basis of the personal experience and judgment of the clinician. In this study we present text mining procedures based on a neural network framework in order to identify indicative factors built on significant keywords extracted from within the medical record narratives. These keywords and their weight/value suggest an innovative way for justifying a CT scan request. Our purpose is to develop an artificial intelligence strategy that can assist with the clinical decision-making process beyond traditional methods, based on an efficient utilization of the narratives in medical records.

Even though much concern is focused on the lack of digitalized narratives [9], we were able to access digitalized clinician records for this study. Radiological records of CT scans performed on children in Nagasaki Medical University Hospital during 2004 were used in the study. From these records groups of keywords were

extracted and analyzed to provide for a predictive weight/value based on the Self Organizing Map (SOM) clustering procedure applied to the words in the complete radiological record.

In section 2.1 the text mining methods employed are outlined and in section 2.2 the data and the processing techniques used are described. In section 3, the initial results of the analysis are presented along with interpretations based on medical expertise. Finally, conclusions and future research directions are outlined.

II. CT DATA AND TEXT MINING

A thorough understanding of the indicators for a CT scan request would require the analysis of large volumes of text data in the narratives of medical records. These records are generally in unstructured formats. The unstructured nature of the data hinders the use of traditional statistical methods. Such traditional methods show limited promise in isolating factors that could accurately predict patterns of CT scan usage. Hence, we chose SOM based text mining techniques to overcome these issues. Kohonen's SOM is a feed forward artificial neural network that uses unsupervised algorithmic training. It can be used to classify multidimensional data by means of clustering whereby similar input vectors are grouped together. Two inputs are considered to be similar, if under a chosen distance measure, the two inputs are close to each other. This type of clustering has led to considerable success in discovering relationships within datasets in the absence of any direction during the learning process. A package called Viscovery SOMine was employed to model the data. It provides a visualization tool that maps high dimensional inputs onto a two-dimensional map. Data features unknown through inspection may then become visible.

A. Processing of data in radiological records

Our study dataset was extracted from the Nagasaki University Hospital Radiology Department's CT scanning database. The individuals are children who received CT scans to aid in their clinical diagnosis (referred to as clinical information herein). The following are the data extracted from the main Nagasaki Hospital database:

1. Exam Title (an anatomical description of the CT scan exposure area, such as head, chest, abdomen, etc)
2. ID number (a unique code for each patient)
3. Age
4. Sex
5. Department
6. In/Out patient status
7. Clinical information *(as the reason to request a CT)*
8. Findings *(as part of CT report by radiologists)*
9. Impression 1, 2, and 3 *(as part of CT report by radiologists)*
10. Result *(as part of CT report by radiologists)*

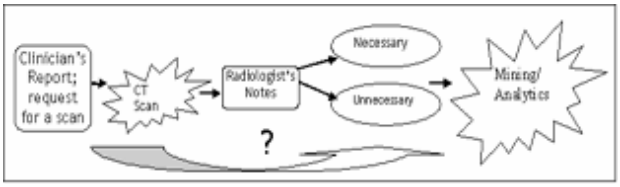

Figure 1. Search for significance in clinician's records

Based on our methodology, we assumed that the narrative texts in "clinical information" would include some factors that clinicians regarded as grounds for the CT requests; we analyzed these narratives using SOM based text mining methodologies. This analysis was performed as a search for any correlations between particular words within the narratives (physicians' notes) and a positive/negative outcome. The positive or negative outcome was determined according to an independent analysis of items 8 ~ 10 by a physician. A positive outcome, in this paper, means the decision to request a CT scan was judged to be useful in reaching a diagnosis/management for the patient, without some cases no abnormality would be found. A negative outcome was any consideration of the possible harmful exposure related to radiation.

Initially, the analyzed records were broken into words. With these words a doc x word matrix was created. The (sparse) matrix consists of the patient records as rows and (the total set of documents dictionary of) words as columns. Each record had its words and word weight in the matrix. If a record did not have a particular word it was marked with 0.

Then by applying formula (1), a weight was calculated for the words in the matrix.

$$W = tf \times idf \quad (1)$$

A 200 node SOM (Fig. 3) was created using the word weight matrix of the records. The SOM was grouped into two clusters based on their outcome as either positive or negative. Additional preference was given to facilitate the separation of the clusters into positive and negative nodes. The two cluster details, the words and their relevant weights can be seen in Fig. 2. Further details related to word groupings and their weights are listed in table 1. The word groupings from the two clusters (C1 negative and C2 positive) are then studied to see how they could be related to their outcome.

III. RESULTS AND DISCUSSION

The 200 node SOM created using the matrix and the two clusters (positive and negative outcome) were analyzed to see the word groupings within them and how they could be related to their outcome.

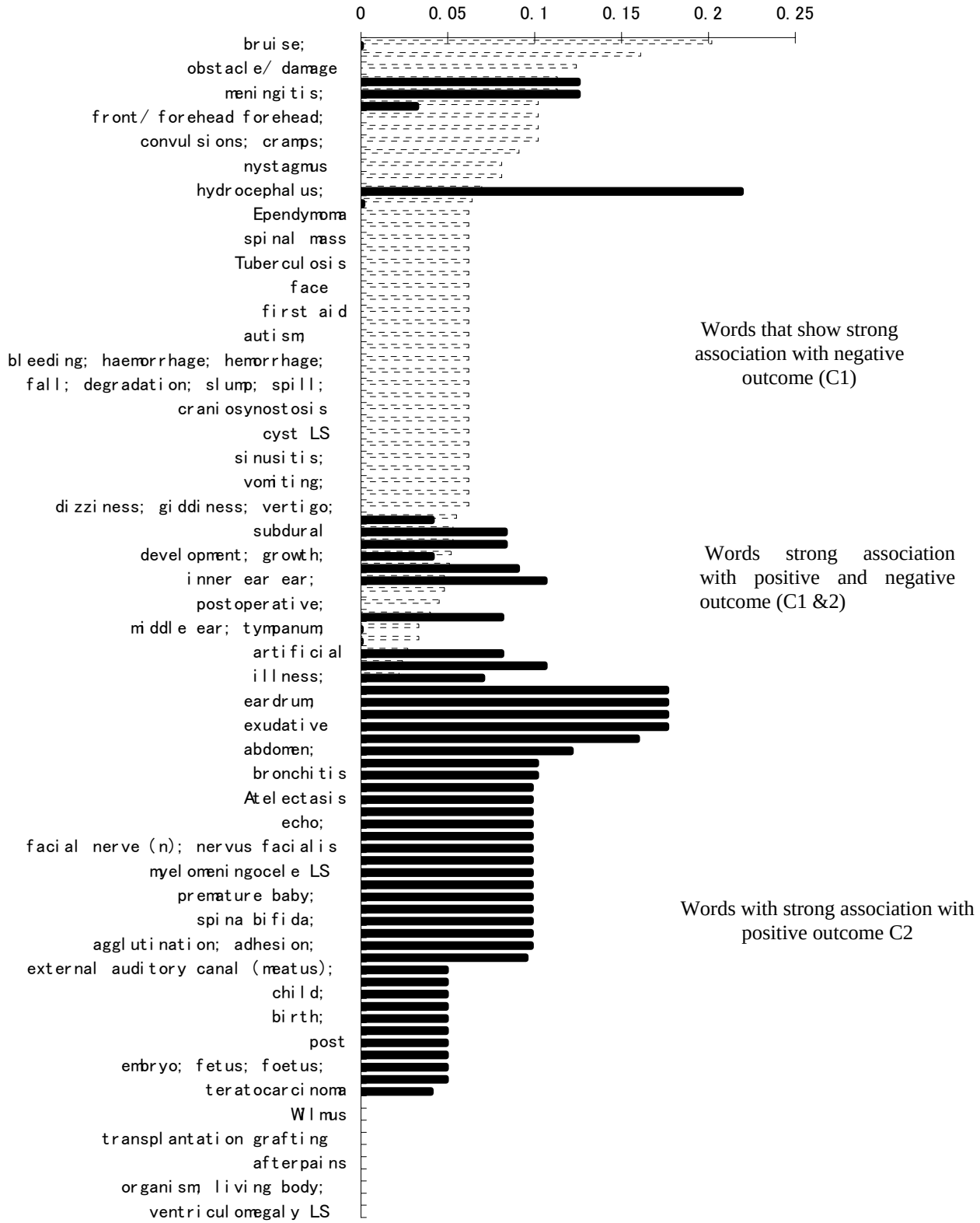

Figure 2. All words analyzed (C2 positive outcome / necessary scan and C1 negative outcome / unnecessary scan) word weights (mean).

Categories of keywords	Examples of keywords	Negative weight (C1)	Positive weight (C2)
Anatomical terms	Subdural	0.053	0.083
	Inner ear	0.048	0.106
Symptoms/signs	Nausea	0.062	0
	Deafness;	0.024	0.106
Diseases/syndromes/disorders/injuries	Tuberculosis	0.062	0
	Hydrocephalus	0.0695	0.2188
Diagnostic procedures	Mass screening	0.062	0
	Reflex test	0.2019	0.0002
Therapeutic modalities/procedures	Surgical operation	0.062	0
	First-aid	0.062	0
Indications of time	Postoperative	0.045	0
	Acute	0.062	0
Daily life events	Fall	0.051	0.09
	Birth	0.055	0.041
Others	Growth	0.052	0.041
	Accident	0.102	0

Table 1. List of valuable keywords (and their weights) from the narratives that show a statistically significant correlation with the CT results (C1 and C2 for negative and positive respectively). These keywords are grouped into different categories: anatomical terms, symptoms/ signs, diseases/ syndromes/ disorders/ injuries, diagnostic procedures, therapeutic modalities/ procedures, indications of time, daily life events, and finally others. The weights of these keywords is also shown in the table.

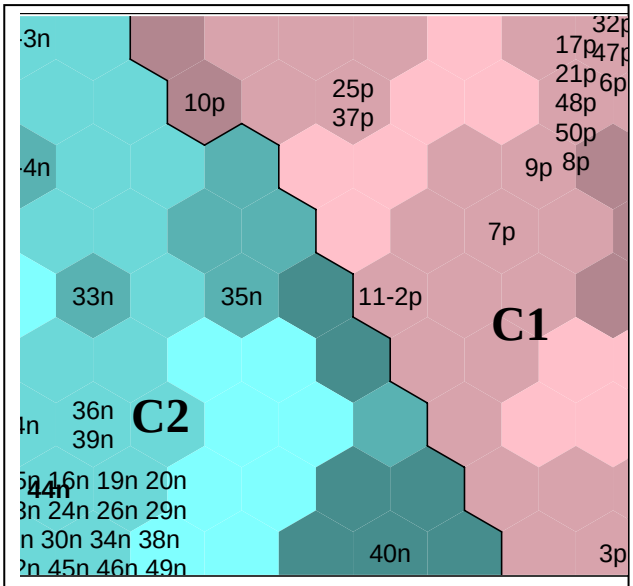

Figure 3. SOM of physician's CT referral notes (text being mined) grouped into positive (C2) versus negative (C2) outcome clusters.

Through the methods of text mining explained earlier, we identified a series of keywords within the CT scan referral rationale. The statistical strength assigned to the keywords (see Fig 2, cluster C1 & C2 mean), led to their separation into three sets; one that had strong associations with a positive finding, one with strong associations with a negative finding and a third group with weak associations with both. Note: these findings were about a positive or negative finding and/or usefulness of the CT request in reaching a diagnosis.

Accordingly, we were able to list and focus on two different subsets of keywords that were strongly correlated with either a positive, or a negative result. We repeated statistical tests to check for the strength of association and the statistical test results were in the acceptable range (a one tailed p-value of 6.47E-20). On the other hand, as mentioned above, there were another set of keywords which were not related strongly to any of the above-mentioned results and we could not find any statistically significant association between them and a positive or a negative result

We also observed that valuable keywords in the narratives could be grouped into different categories, such as, anatomical terms, symptoms or signs, a disease/ syndrome/ disorder or an injury, a diagnostic procedure,

a therapeutic modality (medicines or procedures), an indication of time (post or pre operative, acute or chronic, months, days or weeks), daily life events (birth, traffic accident, fall, fight, and so on) and another less defined but still significant category; “others”.

These different categories of words are generally found to occur together with a few other common words (such as a, the, above and low, referred to as stop words) in the form of a narrative. These narratives are the notes that clinicians prepare explaining the justification of a CT scan to the radiologist. Therefore, the occurrence of these words within the narratives can be used to categorize the clinical reasoning behind the justification of the CT scan. They provide us with data that can be mined to identify the keywords and combinations of them that denote a higher chance of having a useful result through a CT examination.

With this approach, an association was identified between the CT scan request narratives and a positive or negative outcome by the physician. We could determine the keywords and combinations of them that were more likely to help with the decision-making of the physician through an analysis of the respective weights /values of all the words used in the matrix. Accordingly, it could be hypothesized that CT scans would be a better diagnostic tool if doctors employed an artificial intelligence tool which could assign a significance weight for keywords present in the narrative text of medical records.

A C5.0 decision tree was also created (figure 4.). C5.0 is a commercial classification algorithm used to generate a decision tree using the idea of information entropy. It is not restricted to producing binary trees. The same weighting (calculated using the formula, tf x idf), was used as the input into the algorithm to generate a decision tree. We assumed uniform misclassification costs and performed a 3-fold validation resulting in a mean of 53.7 and standard error of 6.5 The decision tree was used to identify rules that may be used to categorize the input.

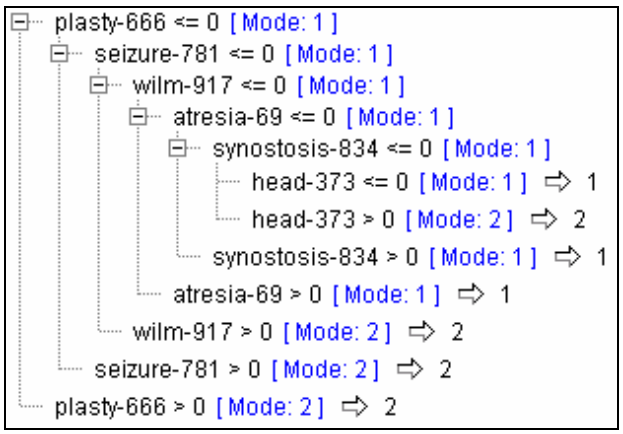

Figure 4. Tabbed version of C5.0 decision tree .1 indicates the predication of a necessary scan and 2 indicates the prediction of an unnecessary scan.

CONCLUSION AND FUTURE WORK

This shows that in the future medical doctors may be able to consider a pre-test weight for the justification of a test, prior to requesting it. The narratives in radiological records may provide clues - reported as probability weights, as to whether a CT scan will be beneficial in the diagnosis/management of children. This is very important because it could lead to a decrease in the number of unnecessary CT scan requests, particularly in children, when the chances of a significant impact on the management of the child are remote. By avoiding such requests, the average total radiation dose that children receive can be reduced.

Our analysis showed that medical narratives include non-medical terms, terms that are generally unnoticed and not recorded in a structured database, but which could have significant implications as to the need for CT scanning. Currently, further research is underway to include a semantic analysis of the relationships among words in each narrative, and to find further relationships among possible word groupings for assigning a collective weight/value for positive or negative outcomes.

We believe it is possible to design a text mining system that could help with such decision making when a clinician is considering whether a CT scan could be helpful in reaching a diagnosis or not. This text mining system can be fed with the hospital’s own data so that local patterns of association between clinical information and radiological findings are revealed, and updated regularly to improve decision-making.

IV. ACKNOWLEDGMENTS

The authors wish to express their sincere gratitude to Professor Monte Cassim who has been a tremendous inspiration for all of us here at Discovery Laboratory, especially, into text mining clinical data. Nagasaki Medical University Hospital clinical staff members are acknowledged for permission to use their data in this study.

V. REFERENCES

[1] Nickoloff EL, Alderson PO. Radiation exposure to patients from CT: reality, public perception, and policy. *Am J Roentgenol*,177: 285-287, 2001.

[2] Frush DP, Donnelly LF, Rosen NS. Computed tomography and radiation risks: what pediatric health care providers should know. *Pediatrics*, 112: 951-957, 2003.

[3] Brenner DJ, Ellison CD, Hall EJ, Berdon WE. Estimated risks of radiation induced fatal cancer from pediatric CT. *AJR*, 176: 289-296, 2001.

[4] Roebuck DJ. Risk and benefit in pediatric radiology. *Pediatr Radiol*, 29: 637-640, 1999.

[5] Ghotbi N, Morishita M, Ohtsuru A and Yamashita S. *Evidence-based guidelines needed on the use of CT scanning in Japan.* Japan Medical Association Journal. 2005, 48:451-457.

[6] Berrington de González A, Darby S. Risk of cancer from diagnostic X-rays: estimates for the UK and 14 other countries. *Lancet*, 363: 345-351, 2004.

[7] Shiralkar S, Rennie A, Snow M, Galland RB, et al. Doctors' knowledge of radiation exposure: questionnaire study. *BMJ*, 327: 371-372, 2005.

[8] Ghotbi N, Ohtsuru A, Ogawa Y, Morishita M, Norimatsu N, Namba H, Moriuchi H, Uetani M & Yamashita S. *Pediatric CT scan usage in Japan: results of a hospital survey.* Radiation Medicine. 2006, 24:560-567.

[9] Walsh S H. The Clinician's Perspective on Electronic Health Records. *BMJ* 2004; 328;1184-1187. doi:10.1136/bmj.328.7449.1184