Extracting Association Rules from Turkish Otorhinolaryngology Discharge Summaries

Başak OĞUZ YOLCULAR[1], Mehmet Kemal SAMUR[2], Uğur BİLGE[1]

[1]Biyoistatistik ve Tıbbi Bilişim, Akdeniz Üniversitesi, Antalya, Türkiye
[2]Departments of Biostatistics and Computational Biology, Harvard Medical School and Dana Farber Cancer Institute, Boston, USA
oguzbsk@gmail.com, ubilge@akdeniz.edu.tr, samur@jimmy.harvard.edu.tr
(Geliş/Received:07.06.2017; Kabul/Accepted:13.12.2017)
DOI: 10.17671/gazibtd.319690

***Abstract*—** The objectives of this study were to structure otorhinolaryngology discharge summaries with text mining methods and analyze structured data and extract relational rules using Association Rule Mining (ARM). In this study, we used otorhinolaryngology discharge notes. We first developed a dictionary-based information extraction (IE) module in order to annotate medical entities. Later we extracted the annotated entities, and transformed all documents into a data table. We applied ARM Apriori algorithm to the final dataset, and identified interesting patterns and relationships between the entities as association rules for predicting the treatment procedure for patients. The IE module's precision, recall, and f-measure were 95.1%, 84.5%, and 89.2%, respectively. A total of fifteen association rules were found by selecting the top ranking rules obtained from the ARM analysis. These fifteen rules were reviewed by a domain expert, and the validity of these rules was examined in the PubMed literature. The results showed that the association rules are mostly endorsed by the literature. Although our system focuses on the domain of otorhinolaryngology, we believe the same methodology can be applied to other medical domains and extracted rules can be used for clinical decision support systems and in patient care.

***Keywords*—** Association Rule Mining, Information Extraction, Otorhinolaryngology, Rule Extraction, Text Mining

1. INTRODUCTION

Physicians often have to use unstructured free-text data, including discharge summaries, patient reports, doctor's notes and hospital records for clinical research and in patient care. Since huge amounts of patient data are recorded and stored in unstructured, free-text clinical reports, it is impractical to use them in this form as it requires a lot of time and effort to read these files.

Text mining is a methodology that promotes an automatic analysis of a corpus of text documents for the extraction of meaningful information and knowledge from large amounts of text files [1]. Information Extraction (IE) is a critical component of the text mining approach that aims to extract entities and concepts from narratives, and transforms unstructured text into a structured form [1, 2]. By using IE methods, unstructured textual data in documents can be transformed into a structured form for further processing with data mining tools. Several techniques can be used to extract information from text. These vary from simple pattern matching to complete processing methods based on statistical and machine learning algorithms. The extracted information can potentially be used for clinical decision support [3].

A variety of methods and systems has been developed in the clinical domain to extract information from clinical narratives [4-9]. In general, these systems have been designed to structure documents, extract named entities (such as diseases, drugs) and identify associations and links between entities (such as gene-gene, gene-disease, and disease-drug) [7, 10, 11]. Although early systems developed in the biomedical domain were promising, most of them were built for extracting information from clinical texts written in English [2, 12-18]. Therefore, for many other languages in particular for Turkish there is a need for research in this area.

In Turkey, a major portion of patients' clinical observations, including radiology reports, operative notes, and discharge summaries are recorded as narrative text. In some hospitals even laboratory and medication records are only available as part of the physician's notes. Hence, there is a growing need for IE systems, for automatic processing of clinical narrative for further data analysis, as well as new generation clinical decision support systems. To the best of our knowledge, there are only two IE projects [19, 20] which extract information from Turkish medical narratives. Both of these systems were designed for Turkish radiology reports and they propose to map clinical text to standardized concepts of medical ontologies automatically. Turkish medical narratives are

usually recorded in an unstandardized format and it is difficult to extract entities with a high performance using common ontologies or terminologies.

Association rule mining (ARM) is a method that is specifically designed for extracting associations amongst entities. The major advantages of ARM are that it provides understandable and easily interpretable outputs and it is capable of discovering potentially interesting associations amongst a large number of attributes. ARM generates association rules and these rules are suitable for embedding in decision support systems. Because of these advantages, ARM is used in many different areas including market research and is also rapidly gaining popularity in the medical informatics field. For example, ARM was applied to predict heart disease [21], hospital infections using public surveillance data [22], and malaria in South Korea [23], in uncovering dengue outbreaks using local and remote sensing data [24], in identifying help-seeking behavior of adolescents [25], and in exploring the Attention Deficit/Hyperactivity Disorder (ADHD) comorbidity [26].

Here we used otorhinolaryngology discharge summaries that were stored in electronically available text format. Overall this study has three main parts. The first part is the extraction of entities from unstructured discharge summaries by mapping them into standard terminologies (IE module) and storing them in structured data tables. The second part identifies association rules between the extracted concepts (using ARM). Finally the last part deals with assessing the validity of the extracted rules by substantiating them in the existing literature.

2. METHODS

This section consists of four components: (1) an overview of otorhinolaryngology discharge summaries, the input of the system; (2) the text mining process, extraction of medical entities; (3) ARM analysis, rule extraction; (4) Rule filtering and Rule validation.

2.1. Dataset

The patient discharge summaries used in this work were obtained from the Otorhinolaryngology department of Akdeniz University Hospital in Turkey. The discharge summaries were manually written and recorded in an unstructured format (Microsoft Word) by clinicians or medical secretaries. The dataset includes 600 unstructured documents. An otorhinolaryngology discharge summary typically consists of five major sections including demographic information (date of birth, gender, name, surname, etc.), anamnesis (symptom and patient history), physical examination results, laboratory results and an operation report (operation name, operation date, pre-diagnosis, surgeon, etc.) for patients who underwent surgery. In the development of the IE module, we first randomly selected 400 documents out of the total 600, and left the remaining 200 documents for the test and evaluation of the system performance. We used the IE module on the test dataset and corrected the results with a manually created dataset by a domain expert. Later in the ARM analysis we used all 600 documents for the rule extraction process.

2.2. Extracting concepts from discharge summaries

First we transformed all Word documents into standard text format by applying a four level pre-processing method (Figure 1). We then structured each document using n-gram analysis and entity tagging and eventually all documents were converted into rows in a dataset table ready for processing by ARM rule extraction.

Pre-processing

IE typically requires some "pre-processing" such as spell checking, document structure analysis, sentence splitting, tokenization and word sense disambiguation [3]. The discharge summaries were first tokenized using whitespaces and punctuation marks and stop words such as “there”, “more”, “yet” were filtered out to achieve high system performance and reliability. In order to correct spelling errors and stem all the words, we used the Zemberek library [27] which is an open source project that has been developed to provide NLP (Natural Language Processing) solutions for the Turkish language for systems developers.

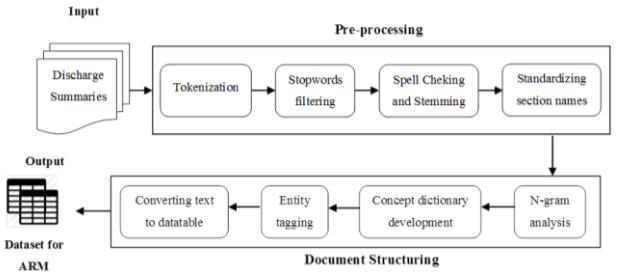

Figure 1. Document processing pipeline

Section segmentation

This module segments all discharge summaries into separate sections. All section headers were converted to capital letters to distinguish them from the textual content. We created a mapping list which included the non-standard section headers and their standard pairs, for example age, birth date, date of birth were all standardized as Age. These standard section headers were used as column headings in data tables.

Information Extraction

A fundamental requirement for IE is a dictionary of Otorhinolaryngology terms. Since no such dictionary is available in Turkish, we chose to construct a dictionary to

extract medical entities from the discharge summaries. We calculated co-occurrences of terms within each section by using the word level n-grams. N-gram method is a contiguous sequence of n items (syllables, letters, and words) from a given sequence of text and it helps to reduce the problems which arise from identifying entities presented by groups of words (Larynx Cancer, Mass of Neck) [28]. For each "n", where "n" is number of words in the entity, our algorithm passed through the data collection once and measured the frequencies of unigrams, 2-grams, 3-grams, and 4-grams). Table 1 shows an example of n-grams from the symptom and diagnosis sections. We considered a word or a group of consecutive words that occur frequently in the entire text collection as a candidate. Using these candidate entities, a dictionary which includes medical terms and their synonyms was compiled and validated by a domain expert.

Table 1. Frequency for Significant Concepts

Symptom			**Diagnosis**		
Words in Turkish	*Words in English*	*f*	*Words in Turkish*	*Words in English*	*f*
Ses Kısıklık	Hoarseness	124	Larenks Kanser	Larynx Cancer	107
Boyun Şişlik	Mass of Neck	75	Tümör	Tumor	78
İşitme Azlık	Hearing Loss	53	Otitis	Otitis	36
Yutma Güçlük	Dysphagia	39	Dil Kanser	Tongue Cancer	29
Burun Tıkanıklık	Nasal Obstruction	37	Septum Deviasyon	Septum Deviation	19
Baş Dönme	Vertigo	34	Hipertrofi	Hypertrophy	12
Kulak Akıntı	Otorrhea	31	Vejetasyon	Vegetation	10
Kulak Önü Şişlik	Preauricular Swelling	28	-	-	-
Dil Yara	Tongue lesion	25	-	-	-
Nefes Darlık	Dyspnea	24	-	-	-

f:The observed frequency

We used the dictionary to extract all medical entities except for the "Patient History" and "Age". The age of each patient was calculated by using the date of birth and the "Patient history" section was used to identify alcohol and smoking history for each patient. We searched for co-occurrences of words and word groups such as "no smoking", "no cigarette", "packet of cigarette", "glass of wine", "no alcohol", "wine", which were defined with the n-gram method we mentioned above, to get information about patients' smoking and alcohol consumption, and afterwards converted the text data into categorical variables as "Smoking" and "Alcohol".

Eventually, all the entities in the documents were extracted by using a dictionary lookup method based on string matching and transformed into a data table. A sample output is given in Figure 2.

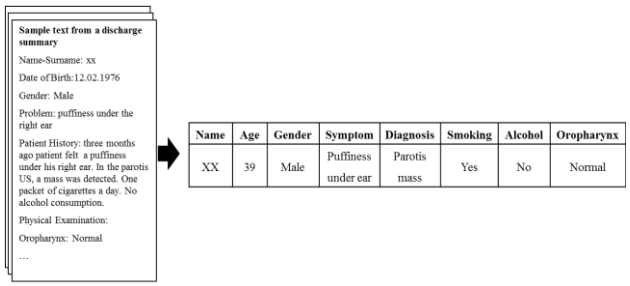

Figure 2. Sample output of data transformation

Evaluation of IE module

We used the 200 discharge summaries as a held-out dataset to evaluate the IE module. We compared the output produced by our system with the observations of a domain expert who reviewed the summaries. In this step, we counted (i) complete tagging (all entities in each section should be extracted) and (ii) non-complete tagging (some entities were incorrectly extracted or absent). We evaluated the performance of the IE module by calculating the precision, recall, and f-measure. In the literature, they are defined as:

$$precision = \frac{tp}{tp+fp}$$

$$recall = \frac{tp}{tp+fn}$$

$$f\text{-}measure = \frac{2xPrecisionxRecall}{Precision+Recall}$$

where tp is the number of true positives, fp is the number of false positives (incorrectly identified items) and fn is the number of false negatives [29]. We calculated precision as the number of correctly identified items divided by the total number of items identified by our system. Recall was computed as the number of correctly identified items divided by the total number of correct items. By combining these measures, we computed the f-measure.

2.3. Association rule mining

In this study we used the association rule mining approach to analyze the structured dataset. ARM is a major data mining technique and is most commonly used as a pattern discovery method [21]. It aims to retrieve frequent patterns and rules which are hidden in a dataset.

An association rule has the form A⇒B, where A and B are sets of items and the B set is likely to occur whenever the A set occurs. A simple example of an association rule for a medical application is the following:

IF (Temperature is Strong Fever) AND (Skin is Yellowish) AND (Loss of appetite is Profound) => (Hepatitis is Acute).

The rule states that if a person has a Strong Fever, Yellowish skin and Profound Loss of appetite, then the person has Acute Hepatitis [24].

In ARM, two measures are commonly used to help a researcher decide the usefulness of an association rule: support and confidence. The support of an association rule A ⇒ B is the percentage of records that contain both A and B and is defined as:

$$Support(A \Rightarrow B) = \frac{number\ of\ records\ with\ A\ and\ B}{total\ number\ of\ records}$$

The confidence of an association rule A ⇒ B is the ratio of the number of records that contain both A and B to the number of records that contain A, and is defined as:

$$Confidence(A \Rightarrow B) = \frac{number\ of\ records\ with\ A\ and\ B}{number\ of\ records\ with\ A}$$

Support measures how frequently an association rule occurs in the entire set of records, whereas confidence measures the strength and reliability of a rule [25]. The task in ARM involves finding all rules that satisfy user-defined constraints on minimum support and confidence with respect to a given dataset. In our study, we applied a third significance metric to the dataset called lift [21]. The lift of a rule (A⇒B) measures the deviation from independence of A and B:

$$Lift\ (A \Rightarrow B) = \frac{Confidence(A \Rightarrow B)}{Support(B)}$$

Lift quantifies the predictive power of A ⇒ B. It measures how much the presence of B depends on the presence or absence of A, and vice versa. The higher the lift is, the more likely that the existence of A and B together is not just a random occurrence, but is due to a relationship between them [24]. In general, a lift value greater than 1 provides strong evidence that A and B depend on each other. A lift value below 1 specifies that A depends on the absence of B, and vice versa. A lift value close to 1 indicates A and B are independent [30].

2.4. Rule filtering and validation

Association rules and the parameters used in them were calculated, analyzed and visualized by the Waikato Environment for Knowledge Analysis (WEKA) which is an open source data mining software tool developed in Java [31]. We applied the Apriori algorithm [32] to the structured dataset in order to extract associations between medical concepts. Confidence, combined with lift and support, was used to evaluate the significance of each rule. Minimum support, confidence and lift were used as the main filtering parameters. Since the dataset was very sparse, the minimum support value should be low enough to obtain a sufficient number of frequent item sets. We found the most meaningful set of rules, by setting the minimum support value to 2%, the confidence value to 90%, the lift value to 2, and getting a final confirmation from a domain expert. To determine high-confidence and non-redundant rules we used a two-step filtering process. In step 1, rules that had confidence, support and lift values less than a given minimum were removed. In step 2, the following filter was adopted to remove redundant rules: Consider two rules R1: (A⇒B), R2: (C⇒B) with the same consequence. We considered that rule R2 was redundant if C ⊂ A while both R1 and R2 had the same confidence, support and lift value. We removed C as it is already covered by A.

To establish that the extracted associations have validity in the real world, a domain expert reviewed the publications on PubMed and found research articles relating to each rule, and checked the validity of our automatically extracted rules.

3. RESULTS

3.1. Characteristics of Dataset

We started our study with 600 discharge summaries. But because of the large amount of missing data in these texts we had to discard many documents, and used only fields that were most regularly written by physicians. After eliminating irrelevant variables (such as patient ID, name, etc.) and missing values from the dataset we compiled a dataset consisting of 274 patient records with seven attributes that have less than 50% missing values for the final analysis. There were four attributes for risk factors in the dataset, namely; gender, age, smoking habits and alcohol consumption. The remaining three attributes were symptom, diagnosis and surgical procedure which were clinical data fields. Age attribute was discretized into four categories after consulting with the domain expert. The characteristics of the dataset are listed in Table 2 with their frequencies and percentages.

3.2. ARM analyses of dataset

Table 3 shows the association rules which could be used for prediction of Surgical Procedure (SP). We present the left side of the discovered rules as predictors (antecedent) and the right side of the rules as the predicted variable (consequent). We discovered 349 association patterns as potential rules for predicting the outcome of surgical procedure. After filtering of association rules that met the threshold, the number of rules was reduced to 15.

Table 2. Frequencies and percentages of attributes and their categories (n=274)

Attributes		n	%
Age	0-25	19	6.9
	26-40	37	13.5
	41-60	118	43.1
	> 60	75	27.4
	Missing	25	9.1
Smoking Habits	Yes (Y)	77	28.1
	No (N)	133	48.5
	Missing	64	23.4
Alcohol Consumption	Yes (Y)	20	7.3
	No (N)	190	69.3
	Missing	64	23.4
Gender	Female (F)	86	31.4
	Male (M)	188	68.6
Symptom	Hoarseness (HS)	58	21.2
	Mass of Neck (MN)	39	14.2
	Otorrhea & Hearing Loss (OH&HL)	18	6.6
	Preauricular Swelling (PS)	20	7.3
	Dyspnea & Hoarseness (DN&HS)	13	4.7
	Nasal Obstruction & Post-Nasal Drainage (NO&PND)	11	4.0
	Tongue lesion (TL)	29	10.6
	Vertigo & Hearing Loss (V&HL)	12	4.4
	Hoarseness & Dysphagia (HS&DP)	13	4.7
	Others	61	22.3
Diagnosis	Larynx Cancer (LC)	113	41.2
	Septum Deviation (SD)	8	2.9
	Tongue Cancer (TC)	29	10.6
	Otitis (OT)	48	17.5
	Neck Tumor (NT)	11	4.0
	Parotid Tumor (PT)	54	19.7
	Septum Deviation & Concha Hypertrophy (SD&CH)	11	4.0
Surgical Procedure	Neck Dissection (ND)	11	4.0
	Hemiglossectomy (HG)	29	10.6
	Laryngectomy (LG)	111	40.5
	Mastoidectomy (MD)	20	7.3
	Parotidectomy (PD)	54	19.7
	Septoconchaplasty (SCP)	11	4.0
	Septoplasty (SP)	8	2.9
	Tympanoplasty (TP)	28	10.2
	Others	2	0.7

A simple example of the description for Rule 3 is the following:

IF (Age is between 26 and 40) AND (Symptoms are Otorrhea & Hearing Loss) AND (Diagnosis is Otitis) => (Surgical Procedure is Tympanoplasty).

The rule states that if a person's age is between 26 and 40, has otorrhea and hearing loss and diagnosed as otitis, then the person undergoes a tympanoplasty operation.

In Azevedo et al.'s study [33], hearing loss was detected as a significant symptom for otitis and the mean age of patients was 26.3 years. In addition, Szaleniec et al.'s study [34] showed that the mean age of patients was 41 years. Both of these studies also indicated that tympanoplasty and mastoidectomy were the most common treatments for otitis. Three of our 15 rules (Rules 1, 2 and 3) produce similar results to these publications. In this study we found that tongue cancer patients who had tongue lesion (Rule 4) and were older than 60 years old underwent a hemiglossectomy operation. Similar age groups were previously reported in Kudoh [35] and Karadeniz et.al.'s studies [36] for operation type and diagnosis. Bussu et al.'s study [37] confirms Rule 7. Their study showed age range for both malignancies and benign group. In addition, Rules 6 and 8 are confirmed with Somefun et al.'s study [38] in which it was observed that preauricular swelling and mass of neck were common symptoms in patients who were diagnosed with a parotid tumor. Markou et al. [39] investigated the role of patient's age and other characteristics in larynx cancer. Parallel to Rule 15, they found the mean age of the patients with larynx cancer as 62 years. In addition, several other studies reported the mean age at diagnosis of larynx cancer as between 60 and 64 years [40-45]. We found that hoarseness, dysphagia and dyspnea were related to larynx cancer (Rules 9, 11 and 12). In the medical literature these symptoms were also reported as the most common symptoms for larynx cancer [43, 46]. Similar to our results, tobacco and alcohol consumption were declared as risk factors for the development of larynx cancer in Matos et al.'s study [47] (Rules 10, 13 and 14).

4. DISCUSSION

In this study, we performed the automatic extraction of rules from unstructured medical text data. We first structured the otorhinolaryngology discharge summaries with text mining methods and then extracted association rules from the structured data. The overall recall and precision of our IE module is 84.5% and 95.1% respectively, which are reasonably good results.

Table 3. The extracted rules from the dataset

	Antecedent						Consequent		Quality Measures		
Rule No	Gender	Age	Smoke	Alcohol	Symptom	Diagnosis	Surgical Procedure	Support (%)	Confidence (%)	Lift ratio
1	-	-	-	-	V&HL	OT	MD	4.4	100	13.7
2	-	-	-	-	OH&HL	OT	TP	6.6	100	9.8
3	-	26-40	-	-	OH&HL	OT	TP	3.7	100	9.8
4	-	>60	-	-	TL	TC	HG	5.1	100	9.4
5	M	-	-	-	-	PT	PD	10.6	100	5.1
6	-	-	-	-	MN	PT	PD	8.8	100	5.1
7	-	41-60	-	-	-	PT	PD	8	100	5.1
8	-	-	-	-	PS	PT	PD	7.3	100	5.1
9	M	41-60	-	-	HS	LC	LG	12.1	100	2.5
10	M	-	Y	Y	-	LC	LG	5.5	100	2.5
11	-	-	-	-	HS&DP	LC	LG	4.7	100	2.5
12	M	-	-	-	DN&HS	LC	LG	4.7	100	2.5
13	M	41-60	Y	-	-	LC	LG	10.6	100	2.5
14	M	-	Y	-	-	LC	LG	16.4	98	2.4
15	M	>60	-	-	-	LC	LG	15.3	95	2.4

Previously in the medical literature, there have been a vast number of medical text mining and information extraction systems reported [48]. The developed systems generally use existing terminologies such as SNOMED CT [49], MESH [50], UMLS [50-52] and are mostly developed for English language for identifying and extracting medical entities from text documents. Turkish medical narratives are usually written in an unstandardized format and it is difficult to extract entities with a high performance using standard type ontology or terminology. In order to overcome this challenge, we generated our own concept dictionary to identify significant entities. We used the n-gram method to set up the dictionary, and to improve the performance of the ARM analysis we only used the high frequency concepts; this may cause less frequent concepts not being represented in the dictionary, and as a result not appearing in rule associations. Another limitation of our work is the dictionary lookup method we used in the IE module; the simple string matching algorithm seemed to work well for detecting most entities, but in some cases it did not recognize some entities. The performance of this module could be improved further by using a more sophisticated string matching algorithm.

An important limitation in our work is the dataset, as the performance of the system strongly depends on the quality of data. In our dataset, there were many missing fields including laboratory results and physical examination results; because of this, we were only able to analyze a dataset with a small number of patients and attributes. Any further studies with an increased number of variables could produce more medically interesting rules.

As for the otorhinolaryngology discharge notes, we discovered 15 relational rules which were in agreement with the recent literature. The extracted rules are given in Table III. As mentioned before, to validate extracted associations, a domain expert reviewed the recent publications related to the extracted association rule in PubMed and checked the validity of our rules. By comparing our results with the literature, we can argue that there was good support for the automatically extracted rules from this study.

Our system is domain-dependent, and the current version can only be used for otorhinolaryngology discharge summaries. But a similar approach could be applied to develop systems for different clinical domains.

REFERENCES

[1] F. Zhu, P. Patumcharoenpol, C. Zhang et al., "Biomedical text mining and its applications in cancer research", *J Biomed Inform*, 46(2), 200-11, 2013.

[2] M. Khabsa, C.L. Giles, "Chemical entity extraction using CRF and an ensemble of extractors", *J Cheminform*, 7(1), 12, 2015.

[3] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research", *Yearb Med Inform*, 128-44, 2008.

[4] Y.K. Lin, H. Chen, R.A. Brown, "MedTime: a temporal information extraction system for clinical narratives", *J Biomed Inform*, 46(20), 8, 2013.

[5] B. Tang , Y. Wu, M. Jiang, Y. Chen, J.C. Denny, H. Xu, "A hybrid system for temporal information extraction from clinical text", *J Am Med Inform Assoc*, 20(5), 828-35, 2013.

[6] J. D. Patrick, D. H. Nguyen, Y. Wang, M. Li, "A knowledge discovery and reuse pipeline for information extraction in clinical notes", *J Am Med Inform Assoc*, 18(5), 574-9, 2011.

[7] M. Jiang, Y. Chen, M. Liu, et al., "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries", *J Am Med Inform Assoc*, 18(5), 601-6, 2011.

[8] M. Sevenster, R. van Ommering, Y. Qian, "Automatically correlating clinical findings and body locations in radiology reports using MedLEE", *J Digit Imaging*, 25(2), 240-9, 2012.

[9] J. H. Chiang, J. W. Lin, C. W. Yang, "Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE)", *J Am Med Inform Assoc*, 17(3), 245-52, 2010.

[10] G. Schadow, C. J. McDonald, "Extracting structured information from free text pathology reports", *AMIA Annu Symp Proc*, 584-8, 2003.

[11] G. K. Savova, J. Fan, Z. Ye, et al., "Discovering peripheral arterial disease cases from radiology notes using natural language processing", *AMIA Annu Symp Proc*, 722-6, 2010.

[12] L. Cui, S. S. Sahoo, S. D. Lhatoo, et al., "Complex epilepsy phenotype extraction from narrative clinical discharge summaries", *J Biomed Inform*, 51, 272-9, 2014.

[13] S. Pradhan, N. Elhadad , B. R. South, et al., "Evaluating the state of the art in disorder recognition and normalization of the clinical narrative", *J Am Med Inform Assoc*, 22(1), 143-54, 2015.

[14] G. Divita, Q. T. Zeng, A. V. Gundlapalli, S. Duvall, J. Nebeker, M. H. Samore, "Sophia: A Expedient UMLS Concept Extraction Annotator", *AMIA Annu Symp Proc*, 467-76, 2014.

[15] V. I. Petkov, L. T. Penberthy, B. A. Dahman, A. Poklepovic, C. W. Gillam, J. H. McDermott, "Automated determination of metastases in unstructured radiology reports for eligibility screening in oncology clinical trials", *Exp Biol Med (Maywood)*, 238(12), 1370-8, 2013.

[16]T. Botsis, E. J. Woo, R. Ball, "The contribution of the vaccine adverse event text mining system to the classification of possible Guillain-Barre syndrome reports", *Appl Clin Inform*, 4(1), 88-99, 2013.

[17] E. Chen, M. Garcia-Webb, "An analysis of free-text alcohol use documentation in the electronic health record: early findings and implications", *Appl Clin Inform*, 5(2) 402-15, 2014.

[18] S. Bozkurt, J. A. Lipson, U. Senol, D. L. Rubin, "Automatic abstraction of imaging observations with their characteristics from mammography reports", *J Am Med Inform Assoc*, 22(1), e81-92, 2015.

[19]D. Onur, **Ontology Based Text Mining in Turkish Radiology Reports**, Computer Engineering Department, Middle East Technical University, 2012.

[20] E. Soysal, I. Cicekli, N. Baykal, "Design and evaluation of an ontology based information extraction system for radiological reports", *Comput Biol Med*, 40(11-12), 900-11, 2010.

[21] C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction", *IEEE Trans Inf Technol Biomed*, 10(2), 334-43, 2006.

[22] S. E. Brossette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, S. A. Moser, Association rules and data mining in hospital infection control and public health surveillance, J Am Med Inform Assoc, 5(4), 373-81, 1998.

[23] A. L. Buczak, B. Baugher, E. Guven, et al., "Fuzzy association rule mining and classification for the prediction of malaria in South Korea", *BMC Med Inform Decis Mak*, 15, 47, 2015.

[24] A. L. Buczak, P. T. Koshute, S. M. Babin, B. H. Feighner, S. H. Lewis, "A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data", *BMC Med Inform Decis Mak*, 12, 124, 2012.

[25] D. H. Goh, R. P. Ang, "An introduction to association rule mining: an application in counseling and help-seeking behavior of adolescents", *Behav Res Methods*, 39(2), 259-66, 2007.

[26] Y. M. Tai, H. W. Chiu, "Comorbidity study of ADHD: applying association rule mining (ARM) to National Health Insurance Database of Taiwan", *Int J Med Inform*, 78(12), e75-83, 2009.

[27] R. A. Erhardt, R. Schneider, C. Blaschke, "Status of text-mining techniques applied to biomedical text", *Drug Discov Today*, 11(7-8), 315-25, 2006.

[28] A. Eftekhar, W. Juffali, J. El-Imad, T. G. Constandinou, C. Toumazou, "Ngram-derived pattern recognition for the detection and prediction of epileptic seizures", *PLoS One*, 9(6) e96235, 2014.

[29] C. Cano, A. Blanco, L. Peshkin, "Automated identification of diagnosis and co-morbidity in clinical records", *Methods Inf Med*, 48(6), 546-51, 2009.

[30] C. Ordonez, N. Ezquerra, C. A. Santana, "Constraining and summarizing association rules in medical data", *Knowl Inf Syst*, 9(3), 259–83, 2006.

[31] E. Frank, M. Hall, L. Trigg, G. Holmes, I. H. Witten, "Data mining in bioinformatics using Weka", *Bioinformatics*, 20(15), 2479-81, 2004.

[32] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules", **20th The VLDB Conference**, Santiago, Chile, 1994.

[33] A. F. Azevedo, D. C. Pinto, N. J. Souza, D. B. Greco, D. U. Goncalves, "Sensorineural hearing loss in chronic suppurative otitis media with and without cholesteatoma", *Braz J Otorhinolaryngol*, 73(5), 671-4, 2007.

[34] J. Szaleniec, M. Wiatr, M. Szaleniec, et al., "Artificial neural network modelling of the results of tympanoplasty in chronic suppurative otitis media patients", *Comput Biol Med*, 43(1), 16-22, 2013.

[35] M. Kudoh, "Longitudinal assessment of articulatory and masticatory functions following glossectomy for tongue carcinoma", *Kokubyo Gakkai Zasshi*, 77(1), 27-34, 2010.

[36] A. Karadeniz, M. Saynak, Z. Kadehci, et al., "The results of combined treatment (surgery and postoperative radiotherapy) for tongue cancer and prognostic factors", *Kulak Burun Bogaz Ihtis Derg*, 17(1), 1-6, 2007.

[37] F. Bussu, C. Parrilla, D. Rizzo, G. Almadori, G. Paludetti, J. Galli, "Clinical approach and treatment of benign and malignant parotid masses, personal experience", *Acta Otorhinolaryngol Ital*, 31(3), 135-43, 2011.

[38] O. A. Somefun, J. O. Oyeneyin, F. B. Abdulkarrem, O. B. da Lilly-Tariah, L. T. Nimkur, O. O. Esan, "Surgery of parotid gland tumours in lagos: a 12 year review", *Niger Postgrad Med J*, 14(1), 72-5, 2007.

[39] K. Markou, J. Goudakos, S. Triaridis, J. Konstantinidis, V. Vital, A. Nikolaou, "The role of tumor size and patient's age as prognostic factors in laryngeal cancer", *Hippokratia*, 15(1), 75-80, 2011.

[40] F. T. Aires, R. A. Dedivitis, M. A. Castro, D. A. Ribeiro, C. R. Cernea, L. G. Brandao, "Pharyngocutaneous fistula following total laryngectomy", *Braz J Otorhinolaryngol*, 78(6), 94-8, 2012.

[41] L. D. Thompson, "Chondrosarcoma of the larynx", *Ear Nose Throat J*, 83(9), 609, 2004.

[42] J. T. Cohen, G. N. Postma, S. Gupta, J. A. Koufman, "Hemicricoidectomy as the primary diagnosis and treatment for cricoid chondrosarcomas", *Laryngoscope*, 113(10), 1817-9, 2003.

[43] L. D. Thompson, F. H. Gannon, "Chondrosarcoma of the larynx: a clinicopathologic study of 111 cases with a review of the literature", *Am J Surg Pathol*, 26(7), 836-51, 2002.

[44] M. Brandwein, S. Moore, P. Som, H. Biller, "Laryngeal chondrosarcomas: a clinicopathologic study of 11 cases, including two "dedifferentiated" chondrosarcomas", *Laryngoscope*, 102(8), 858-67, 1992.

[45] O. Casiraghi, F. Martinez-Madrigal, K. Pineda-Daboin, G. Mamelle, L. Resta, M. A. Luna, "Chondroid tumors of the larynx: a clinicopathologic study of 19 cases, including two dedifferentiated chondrosarcomas", *Ann Diagn Pathol*, 8(4), 189-97, 2004.

[46] I. Buda, R. Hod, R. Feinmesser, J. Shvero, "Chondrosarcoma of the larynx", *Isr Med Assoc J*, 14(11), 681-4, 2012.

[47] J. P. Matos, J. C. Silva, E. Monteiro, "Causes of death in patients with laryngeal cancer in stages I and II", *Acta Med Port*, 25(5), 317-22, 2012.

[48]S. Velupillai, D. Mowery, B. R. South, M. Kvist, H. Dalianis, "Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis", *Yearb Med Inform*, 10(1), 183-93, 2015.

[49] A. N. Nguyen, M. J. Lawley, D. P. Hansen DP, et al., "Symbolic rule-based classification of lung cancer stages from free-text pathology reports", *J Am Med Inform Assoc*, 17(4), 440-5, 2010.

[50] M. Stevenson, E. Agirre, A. Soroa, "Exploiting domain information for Word Sense Disambiguation of medical documents", *J Am Med Inform Assoc*, 19(2), 235-40, 2012.

[51] I. McCowan , D. Moore, M. J. Fry, "Classification of cancer stage from free-text histology reports", **28th Annual International Conference of the IEEE**, New York, NY, USA, September, 2006.

[52] A. Nguyen, D. Moore, I. McCowan, M. J. Courage, "Multi-class classification of cancer stages from free-text histology reports using support vector machine", **29th Annual International Conference of the IEEE**, Lyon, France, August, 2007.