International Journal for Modern Trends in Science and Technology, 7(12): 20-22, 2021
Copyright © 2021 International Journal for Modern Trends in Science and Technology
ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0712004
Available online at: http://www.ijmtst.com/vol7issue12.html

Extraction of Clinical Data from Electronic Health Records using Regular Expression

Anjali Sharma[1] | Divya Sharma[2] | Shagun Sharma[3] | Gulshan K. Sharma[4] | Rakesh Sharma[5] | Vandana Nunia[6]

[1]Department of Pathology, Bhagwan Mahavir Cancer Hospital and Research Centre, Jaipur, Rajasthan, INDIA
[2] JK Lakshmipat University, Mahapura, Rajasthan, INDIA
[3]Department of Botany, University of Rajasthan, Jaipur, Rajasthan, INDIA
[4,5] Bioinformatics Infrastructure Facility DBT-BIF, University of Rajasthan, Jaipur, Rajasthan, INDIA
[6] Department of Zoology, University of Rajasthan, Jaipur, Rajasthan, INDIA

To Cite this Article
Anjali Sharma, Divya Sharma, Shagun Sharma, Gulshan K. Sharma, Rakesh Sharma and Vandana Nunia. Extraction of Clinical Data from Electronic Health Records using Regular Expression. *International Journal for Modern Trends in Science and Technology* 2021, *7*, pp. 20-22. https://doi.org/10.46501/IJMTST0712004.

Article Info
Received: 17 October 2021; Accepted: 27 November 2021; Published: 29 November 2021

ABSTRACT

Implementation of electronic health record (EHR) systems continues to expand in hospitals to store huge amount of data. Transforming this huge clinical data or 'big data' into knowledge to improve patient care has been the goal of biomedical informatics professionals for many decades. These extracted data sets can be modelled and used for many big data applications. We have used big data, in the context of EHR systems and extracted this data for secondary applications like prediction and prognosis of diseases. Medical reports or unstructured data were parsed from the PDF, DOC and image files. Here, we have tried to develop an automated data extraction method, which depends on python programming language and uses regular expressions to grab patterns in the data.

KEYWORDS: Electronic health records, Data mining, Regular expression, Python.

I. INTRODUCTION

With the use of health management systems (HMS) and electronic health records (EHR) in the biomedical field, enormous amount of informative and unstructured data sets are generated. Biomedical scientists are facing the challenges of managing massive amounts of this clinical data. This huge size of electronic data or 'big data' cannot be managed efficiently by traditional methods. Various technologies are being developed for data extraction from EHR, using simple pattern matching to complete processing methods based on symbolic information and rules based on statistical methods and advanced machine learning [1-5]. Based on these techniques various data Information extraction models have been developed, explored and analyzed further [6-11]. Major drawbacks of these developed systems consist of cater to one specific type of clinical or biomedical data. Therefore, there is a need to design and develop a comprehensive system that caters to diverse kinds of users through different kinds of access modes, while maintaining patients' data consistency and privacy.

The technology of information extraction has advanced significantly and applied widely in the biomedical field. Thus, coupled with the application of big data analytic techniques, there are many ways in which this goal can be achieved. Recently, advanced machine learning techniques are being used to manage this data and support automated systems to enable secondary use of EHRs for clinical and translational research. One critical step involved in this is the information extraction (IE) task, which automatically extracts and encodes clinical information from text to tables or data frames. Various techniques based on some levels of text extraction, text mining or natural language processing (NLP) includes basic python processing to convert the character stream into a sequence of lexical items (words, phrases, and syntactic markers). Like NLP, regular expression or Regex can be effectively used to describe strings of characters (words or phrases or any arbitrary text). Here, with the application of regex, we have designed a script using python regex which extracts clinical data from EHR by using specific words/patterns/functions. Further, we also tried to build a simple universal model for extraction of clinical data from EHR with all desired information.

II. METHODOLOGY

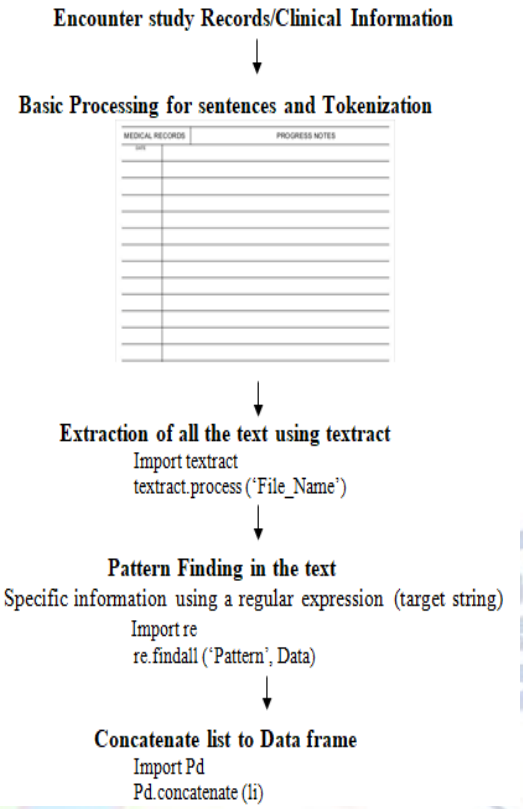
Study Records

Cancer patient's electronic health record files were obtained from Bhagwan Mahavir Cancer Hospital and Research Centre, Jaipur, Rajasthan, India. Identifying informations of the patient were removed from each record and informations like case number, age, sex, nature of material, gross and microscopic features were extracted.

Work Description

In the initial step we input the target medical file containing 15 cases. A parser was written to extract specific information by using regular expressions. A textract module was used in reading the docx file. It was then routed the filename provided into the appropriate parser and returned the whole file into a byte-string encoded.

We have used a regular expression for extracting information from raw data. Regular expressions provided a flexible and concise means to match strings of text. Afterwards extraction of information like case number, age, sex, nature of study material, gross and microscopic clinical features of cancer patients was carried out. Patterns/matched objects were searched by findall functions returning a list of matched patterns as output. This whole information was then split according to the words or patterns provided by the return list of matched patterns. To substitute the pattern provided in the particular data as data mining tasks were classified to descriptive and predictive data mining. The workflow demonstrates the structured illustration of the work done (Fig 1).


Encounter study Records/Clinical Information
Basic Processing for sentences and Tokenization
MEDICAL RECORDS
PROGRESS NOTES
DATE
Extraction of all the text using textract
Pattern Finding in the text
Specific information using a regular expression (target string)
Concatenate list to Data frame

```
Import textract
textract.process ('File_Name')
```

```
Import re
re.findall ('Pattern', Data)
```

```
Import Pd
Pd.concatenate (li)
```

Figure: 1 Structural representation of the work done (a step by step guide).

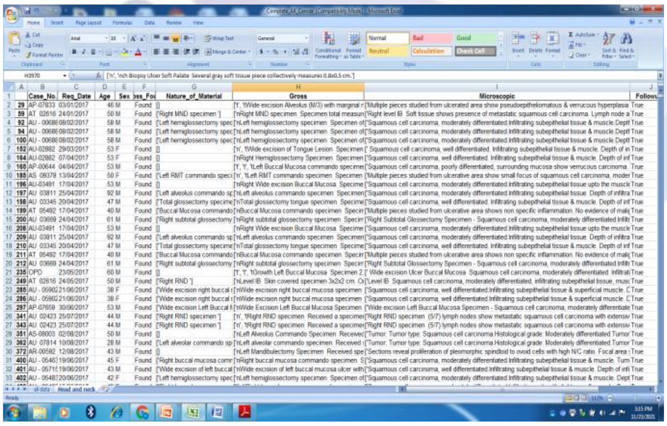
III. RESULT AND DISCUSSION

In the present study, we have introduced an automated information extraction and presentation system that is designed to extract structured clinical information of the patient from EHR. The Implementation of regex expression algorithms under a new approach of extracting out clinical data/information from EHR took a lesser time period than traditional methods supporting it as an efficient way of time saving approach. The huge amount of clinical data and matched patterns of the patients from EHR was successfully extracted and converted to excel as shown in Figure-2.

Present study broadly worked on structured data normalizer, unstructured data extractor, and unstructured data classifier. Sensitivity of the algorithm was found to be 95% with 96% of specificity with structured data. However, the precision value comes out to be 86%. Whereas, with unstructured data sensitivity was 85%. Another similar big data extraction study was carried out by Boytcheva, S et. al., [12] from outpatient care information of 100 million patients to evaluate diabetic compensation. The extracted information was stored in a structured format. Total number of two cycles were involved during the extraction and the extracted outcome were about 38,300,000 with a precise value of 92% and recall value of 98%. Jonnagaddala, J., et al., [13] reported the utility of extracted data as an alternative for identification of specific heart disease patterns and risk assessment. Result showed an overall score of 83% precision which was found to be less than our structured data extraction approach. Additionally, the time attribute was found to evaluate diabetes and hypertension efficiently at 92% and 93% respectively.

IV. CONCLUSION

Considering all the applications of text mining and extraction of critical information from EHRs further significant studies can be performed and health related risks can be assessed. But, lack of easy and efficient IE methods, information stored in EHR are not being utilized for secondary use. Methods based on NLP and Machine Learning tend to perform better in this area but more experience is required to use them. In the present study we have developed an easy to use machine learning algorithm to extract clinical data from EHR and tried to provide potential solutions to bridge this gap.



H3970 ['n', 'nch Biopsy Ulcer Soft Palate Several gray soft tissue piece collectively measures 0.8x0.5 cm.']

A	B	C	D	E	F	G	H	I	J
	Case_No	Req_Date	Age	Sex	oss_Fo	Nature_of_Material	Gross	Microscopic	Follow
29	AP-07833	03/01/2017	46	M	Found	[]	['', 'tWide excision Alveolus (M/3) with marginal r	['Multiple pieces studied from ulcerated area show pseudoepitheliomatous & verrucous hyperplasia	True
59	AT 02616	24/01/2017	50	M	Found	['Right MND specimen ']	['nRight MND specimen Specimen total measur	['Right level IB Soft tissue shows presence of metastatic squamous cell carcinoma. Lymph node a	True
92	AU - 0068E	08/02/2017	58	M	Found	['Left hemiglossectomy spec	['nLeft hemiglossectomy specimen Specimen of	['Squamous cell carcinoma, moderately differentiated Infiltrating subepithelial tissue & muscle.Deptl	True
94	AU - 0068E	08/02/2017	58	M	Found	['Left hemiglossectomy spec	['nLeft hemiglossectomy specimen Specimen of	['Squamous cell carcinoma, moderately differentiated Infiltrating subepithelial tissue & muscle.Deptl	True
100	AU - 0068E	08/02/2017	58	M	Found	['Left hemiglossectomy spec	['nLeft hemiglossectomy specimen Specimen of	['Squamous cell carcinoma, moderately differentiated Infiltrating subepithelial tissue & muscle.Deptl	True
152	AU-02882	29/03/2017	53	F	Found	[]	['n', 'tWide excision of Tongue Lesion Specimen	[' Squamous cell carcinoma, well differentiated. Infiltrating subepithelial tissue & muscle. Depth of in	True
164	AU-02882	07/04/2017	53	F	Found	[]	['nRight Hemiglossectomy Specimen Specimen	['Squamous cell carcinoma, well differentiated. Infiltrating subepithelial tissue & muscle. Depth of inf	True
165	AP-00644	04/04/2017	53	M	Found	[]	['', 't', 'tLeft Buccal Mucosa commando specime	['Squamous cell carcinoma, poorly differentiated, surrounding mucosa show verrucous carcinoma.	True
185	AS 09378	13/04/2017	50	F	Found	['Left RMT commando speci	['n', 'tLeft RMT commando specimen Specimen	['Multiple pieces studied from ulcerative area show small focus of squamous cell carcinoma, moder	True
196	AU-03491	17/04/2017	53	M	Found	[]	['nRight Wide excision Buccal Mucosa Specime	['Squamous cell carcinoma, moderately differentiated Infiltrating subepithelial tissue upto the muscle	True
197	AU 03811	25/04/2017	92	M	Found	['Left alveolus commando sp	['nLeft alveolus commando specimen Specimen	['Squamous cell carcinoma, moderately differentiated Infiltrating subepithelial tissue. Depth of infiltra	True
198	AU 03345	20/04/2017	47	M	Found	['Total glossectomy specime	['nTotal glossectomy tongue specimen Specime	['Squamous cell carcinoma, well differentiated. Infiltrating subepithelial tissue & muscle. Depth of inf	True
199	AT 05492	17/04/2017	40	M	Found	['Buccal Mucosa commando	['nBuccal Mucosa commando specimen Specin	['Multiple pieces studied from ulcerative area shows non specific inflammation. No evidence of malig	True
200	AU 03669	24/04/2017	61	M	Found	['Right subtotal glossectomy	['nRight subtotal glossectomy specimen Specim	['Right Subtotal Glossectomy Specimen - Squamous cell carcinoma, moderately differentiated Infiltr	True
208	AU-03491	17/04/2017	53	M	Found	[]	['nRight Wide excision Buccal Mucosa Specime	['Squamous cell carcinoma, moderately differentiated Infiltrating subepithelial tissue upto the muscle	True
209	AU 03811	25/04/2017	92	M	Found	['Left alveolus commando sp	['nLeft alveolus commando specimen Specimen	['Squamous cell carcinoma, moderately differentiated Infiltrating subepithelial tissue. Depth of infiltra	True
210	AU 03345	20/04/2017	47	M	Found	['Total glossectomy specime	['nTotal glossectomy tongue specimen Specime	['Squamous cell carcinoma, well differentiated. Infiltrating subepithelial tissue & muscle. Depth of inf	True
211	AT 05492	17/04/2017	40	M	Found	['Buccal Mucosa commando	['nBuccal Mucosa commando specimen Specin	['Multiple pieces studied from ulcerative area shows non specific inflammation. No evidence of malig	True
212	AU 03669	24/04/2017	61	M	Found	['Right subtotal glossectomy	['nRight subtotal glossectomy specimen Specim	['Right Subtotal Glossectomy Specimen - Squamous cell carcinoma, moderately differentiated Infiltr	True
235	OPD	23/05/2017	60	M	Found	[]	['', 't', 'tGrowth Left Buccal Mucosa Specimen 2.	[' Wide excision Ulcer Buccal Mucosa Squamous cell carcinoma, moderately differentiated. Infiltrati	True
249	AT 02616	24/05/2017	50	M	Found	['Right RND ']	['nLevel IB Skin covered specimen 3x2x2 cm. Oi	['Level IB Squamous cell carcinoma, moderately differentiated, infiltrating subepithelial tissue, musc	True
285	AU - 05902	21/06/2017	38	F	Found	['Wide excision right buccal r	['nWide excision right buccal mucosa specimen	['Squamous cell carcinoma, well differentiated.Infiltrating subepithelial tissue & superficial muscle. C	True
286	AU - 05902	21/06/2017	38	F	Found	['Wide excision right buccal r	['nWide excision right buccal mucosa specimen	['Squamous cell carcinoma, well differentiated.Infiltrating subepithelial tissue & superficial muscle. C	True
297	AP-07659	30/06/2017	53	M	Found	['Wide excision Left Buccal M	['nWide Excision Left Buccal Mucosa Specimen	['Wide excision Left Buccal Mucosa Specimen - Squamous cell carcinoma, moderately differentiate	True
341	AU 02423	25/07/2017	44	M	Found	['Right RND specimen ']	['n', 'tRight RND specimen Received a specime	['Right RND specimen (5/7) lymph nodes show metastatic squamous cell carcinoma with extensiv	True
343	AU 02423	25/07/2017	44	M	Found	['Right RND specimen ']	['n', 'tRight RND specimen Received a specime	['Right RND specimen (5/7) lymph nodes show metastatic squamous cell carcinoma with extensiv	True
351	AS-08003	02/08/2017	50	M	Found	[]	['nLeft Alveolus Commando Specimen Receivec	['Tumor: Tumor type: Squamous cell carcinoma.Histological grade: Moderately differentiated Tumor	True
362	AU 07814	10/08/2017	28	M	Found	['Left alveolar commando sp	['nLeft alveolar commando specimen. Received :	['Tumor: Tumor type: Squamous cell carcinoma Histological grade: Moderately differentiated Tumor	True
372	AR-00592	12/08/2017	43	M	Found	[]	['nLeft Mandibulectomy Specimen Received spe	['Sections reveal proliferation of pleomorphic spindloid to ovoid cells with high N/C ratio. Focal area :	True
400	AU - 0546:	19/06/2017	45	F	Found	['Right buccal mucosa comr	['nRight buccal mucosa commando specimen S	['Squamous cell carcinoma, moderately differentiated Infiltrating subepithelial tissue & muscle. Tum	True
401	AU - 0571E	19/06/2017	43	M	Found	['Wide excision of left buccal	['nWide excision of left buccal mucosa ulcer with	['Squamous cell carcinoma, well differentiated Infiltrating subepithelial tissue & muscle. Depth of infi	True
402	AU - 0548:	20/06/2017	42	F	Found	['Left hemiglossectomy spec	['nLeft hemiglossectomy specimen Specimen of	['Squamous cell carcinoma, moderately differentiated Infiltrating subepithelial tissue & muscle. Dept	True

all data | Head and neck

Figure: 2 Extracted information on excel sheet: collective data of more than 38,000 patients into an accessible data format.

REFERENCES

[1] Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., & Tyson, M. FASTUS: A finite-state processor for information extraction from real-world text. In IJCAI 1993, 93, 1172-1178.

[2] Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., & Weischedel, R.. BBN: Description of the SIFT system as used for MUC-7. In Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998.

[3] Sridevi, M., & Arunkumar, B. R. Information Extraction from Clinical Text using NLP and Machine Learning: Issues and Opportunities. In National Conference on "Recent Trends in Information Technology"(NCRTIT), International Journal of Computer Applications (0975-8887) 2016.

[4] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 17(5), 507-513.

[5] Liu, H., Bielinski, S. J., Sohn, S., Murphy, S., Wagholikar, K. B., Jonnalagadda, S. R.,& Chute, C. G. (2013). An information extraction framework for cohort identification using electronic health records. AMIA Summits on Translational Science Proceedings, 2013, 149.

[6] Sondhi, P., Gupta, M., Zhai, C., & Hockenmaier, J. (2010, August). Shallow information extraction from medical forum data. In Coling 2010: Posters (pp. 1158-1166).

[7] Uzuner, Ö., Solti, I., & Cadag, E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association, 2010:17(5), 514-518.

[8] Bae I, Kim JS. A refinement system for medical information extraction from text-based bilingual electronic medical records. J Korean Soc Med Inform. 2008;14(3):267–274.

[9] Park YT, Lee YT, Jo EC. Constructing a real-time prescription drug monitoring system. Healthc Inform Res. 2016;22(3):178–185

[10] Glavaš, G. TAKELAB: medical information extraction and linking with MINERAL. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. 389-393).

[11] Kraus, S., Blake, C., & West, S. L. Information extraction from medical notes. Medinfo, 2007: 1-2.

[12] Boytcheva, S., Angelova, G., Angelov, Z., & Tcharaktchiev, D. (2015). Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care. Cybernetics and Information Technologies, 15(4), 58-77.

[13] Jonnagaddala J, Liaw ST, Ray P, Kumar M, Dai HJ, Hsu CY. Identification and Progression of Heart Disease Risk Factors in Diabetic Patients from Longitudinal Electronic Health Records. *Biomed Res Int*. 2015;2015:636371.