Structured and Unstructured Data Integration with Electronic Medical Records

Diogo Baptista, João C. Ferreira, Ruben Pereira and Márcia Baptista

***Abstract*— Medicine is a field with high volatility of changes. Everyday new discoveries and procedures are tested with the sole goal of providing a better-quality life to patients. With the evolution of computer science, multiple fields saw an increase of productivity and solutions that could be implemented. More specifically, in medicine new techniques started being tested in order to under-stand how the systems and practices used can reach higher performances, while maintaining the predefined high standards of quality. For many years data generated in hospitals was collected and stored yet few tools were implemented to extract knowledge or any type of advantage. One of the areas that successfully implemented in medicine was the usage of data processing tools and techniques to further extract information regarding the high abundance of data generated in a daily basis, in this field of work. This data can be stored in different ways which leads to multiple approaches on how to deals with it. The sole purpose of this paper is to give an overview of some case studies where structured and un-structured data was used, joint and separately and the value of it.**

***Index Terms*—Structured Data, Unstructured Data, Natural Language Processing, Data Integration, Electronic Medical Records.**

I. Introduction

With the evolution of the medical field, the life expectancy has been increasing. This evolution leads to good things, such as the end of some diseases (smallpox, plague, etc.) but, on the other hand, new problems arise with the appearance of new ones such as dementia, cancer, etc. But, as medicine grew and developed, so did many other areas, specially computer science. Alongside they have a direct impact on fighting new medical challenges.

Nowadays, healthcare providers store loads of data, medical and non-medical. This data can regard drug prescription, treatment records, general check-up information, physician's notes, surgical information or financial and administrative documents and are either stored in legacy systems or electronic medical records (EMR) [1]. The EMR are computerized medical information systems that collect, store and display patient information [2].

The usage of EMR accommodates multiple advantages. Reference [3] considers that those advantages can be summarized as "optimizing the documentation of patient encounters, improving communication of information to physicians, improving access to patient medical information, reduction of errors, optimizing billing and improving reimbursement for services, forming a data repository for research and quality improvement, and reduction of paper".

The format in which this information is recorded is very important due to its direct impact on how it can be moulded to provide greater insight.

There are three different types of data structures: structured, semi-structured and unstructured [4]. Although in the present work the authors mainly focus in the structured and unstructured. The specificities of both data types are explained further in this paper.

Associated with these different types of data structures, there are different data manipulation techniques, such as data mining (DM) and text mining (TM) (also known as Text Data Mining [5]). These techniques have been used separately, yet the author did not find many studies that combined both approaches, resulting in this work.

The body of this paper is composed of four main topics, the literature review, where the authors emphasize other projects developed in the area, a small chapter to present the data used in the project, the architecture projected and the last one regarding the obtained results.

To complete, a final chapter of conclusions where the author sums up several points approached during the paper and reflects on them.

II. Literature Review

This chapter intends to show and clarify the work that has already been done in the area of clinical information extraction from EMR and its integration with already structured clinical data. There are some case studies that already manage to integrate extracted information from EMR with structured data.

In 2018 a study developed a hybrid model that would involve the mining of texts integrated with the mining of structured data to allow prediction for early stages of dementia. This was achieved by coupling together TM and Natural Language Processing (NLP) with DM [6]. In this study two approaches were compared, the first only

Manuscript received March 06, 2019; revised March 26, 2019. This work has been partially supported by Portuguese National funds through FITEC - Programa Interface, with reference CIT "INOV - INESC Inovação - Financiamento Base".

Diogo Baptista is a Master's student from Instituto Universitário de Lisboa (ISCTE-IUL) (e-mail: diogo_veiga@iscte-iul.pt).

Ruben Pereira is with Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisbon, Portugal (e-mail: ruben.filipe.pereira@iscte-iul.pt).

Joao C. Ferreira is with INOV INESC Inovação – Instituto de Novas Tecnologias, Lisbon, Portugal (e-mail: joao.ferreira@inov.pt), Instituto Universitário de Lisboa (ISCTE-IUL) and Information Sciences, Technologies and Architecture Research Center (ISTAR-IUL), Lisbon, Portugal (e-mail: jcafa@iscte-iul.pt).

Marcia Baptista is with the INOV INESC Inovação – Instituto de Novas Tecnologias, Lisbon Portugal (corresponding author e-mail: marcia.baptista@inov.pt).

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

considering the already structured data, and the second considering said structured data plus a new variable that was the result of text clustering of the patient's previous pathological history. With the final results the authors concluded that the unstructured text that would probably be discarded, has significant value for these models since the best values of precision were obtained in the second approach.

Also in 2018, another study intended to show how different the results would be in a geriatric syndrome case classification, while using three data sources: data claims, structured information from Electronic Health Records (EHR) and unstructured text form the same EHR [7]. The results were somewhat aligned with the conclusions of the first study, since the authors of [7] concluded that the presence of the unstructured element used in the study allowed for a considerable number of individuals to be included that would, otherwise, be missed if only considering the structured information (claims or EHR).

The third case study is also from 2018, and the authors elaborated five iterations to predict the possibility of readmission *post* hospital discharge congestive heart failure.

The data used in this study included both structured (age, ethnicity, marital status, etc.) and unstructured (discharge summary) structures and the five iterations varied in which data was used [8]. The five iterations were:

- 1) Structured Data;
- 2) Unstructured Data;
- 3) Address Class Imbalance (Feature Selection);
- 4) Merge Data (1+2);
- 5) Merge Data (1+3).

And with the best results acquired from iteration five, the authors of [8] concluded that said iteration should be used for further research. Showing once again the advantage of the mixed approach.

More case studies regarding this mixed approach have been made but the most recent were presented in this work.

Although these recent case studies show positive results, they all sought to solve the problem of integrating structured data with unstructured data but for their specific problems, i.e., in all the presented scenarios new variables were created based on the unstructured text characteristics., appearing after a document clustering analysis or other analysis.

The approach differs from the ideas of these papers because the authors intend to develop a system for a more generic integration, making the most use of the written text.

III. Data

Before moving on to the architecture of the system itself, it is good to get an overview of the data in question, both structured and unstructured.

A. Structured Data

On the one hand the structured data used in this work contemplates a collection of records from the Emergencies Departments (ED) of a Portuguese hospital regarding the time between January 1st, 2015 and December 31st, 2017. Each record represents one interaction with hospitals ED.

The aforementioned data has been extracted from the hospitals data base in the form of an excel file and it contains 305.636 records and 15 attributes. The attributes are represented in the table I.

The "Episode Alert" consists of simple identifier of the specific encounter. The "Patient ID", "Nurse ID", "Doctor First Observation ID" are, as the name implies, identifiers of the patients, nurses and doctors. This identification is necessary as a privacy measure. The "ED Sub Department"

TABLE I
ED Data Attributes

Attribute #	Attribute Name	Data Type
1	Episode Alert	Integer
2	Patient ID	String
3	ED Sub Department	String
4	Admission Date	Timestamp
5	Triage Color	Integer
6	Nurse ID	String
7	Triage Date	Timestamp
8	Date First Medical Observation	Timestamp
9	Doctor First Observation ID	String
10	Discharge Date	Timestamp
11	Discharge Status	String
12	Discharge Destination	String
13	Readmission	Flag
14	Disease Code (ICD[a] Code)	Integer
15	Disease Code Description	String

[a] International Code of Diseases 9

represents the local where the event happened. The "Admission Date" is the data at which the patient was admitted. The "Triage Color" categorizes the patient in a scale of color that goes from blue (5) to red (1)[1], from non-urgent to imminent danger, respectively. The triage process is according to the Manchester Triage Protocol. The "Date First Medical Observation" represents the moment when the first observation was made to the patient from a doctor. The "Discharge Date" represents the date at which the patient got the discharge clearance. The "Discharge Destination" involves the destination to where the patient was sent. The "Readmission" is a flag that identifies which records represent a readmission (1) or a simply a new entry (0). Finally, the disease codes are represented as numeric codes the, "Disease Code (ICD Code)", and its description "Disease Code Description". All these descriptions and codes are from International Code of Diseases 9 (ICD 9) and is regulated by the World Health Organization (WHO) as a way to standardize disease description and "promote international comparability in the collection, processing, classification, and presentation of mortality statistics" [9].

B. Unstructured Data

On the other hand, the unstructured data used in this work consists of a set of EMR's from the same Portuguese hospital in an excel file, but this time with appointments. In other words, each line from the excel is related to an individual appointment. Since this dataset is related to the appointment's department, it will be treated by Appointment Department (AD) dataset.

This dataset contains data relative to the year of 2017 and to medical services that do not require the admission of the patient to the hospital or health facility, such as diagnosis,

[1] 1 – Red; 2 – Orange; 3 – Yellow; 4 – Green; 5 – Blue;

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

rehabilitation, treatment, etc. Also, the dataset contains information relative to several medical specialties: pulmonology, oncology, rheumatology, gastroenterology, hematology, nephrology, pediatrics, pediatric hematology and urology.

Next, present in the table II, it is represented the structure of a line of the dataset.

TABLE II
AD DATA ATTRIBUTES

Attribute #	Attribute Name	Data Type
1	Clinical Episode	Integer
2	Medical Specialty	String
3	Medical Specialty Code	Integer
4	Medical Specialty Description	String
5	Diagnosis Description	String
6	Diagnosis Code	Integer
7	Date	Date
8	Clinical Narrative Text	String

The "Clinical Episode" works exactly as an identifier for each appointment. The "Medical Specialty" infer to the medical specialty of the appointment. The "Medical Specialty Code" represents the code of the medical specialty and the "Medical Specialty Description" represents the correspondent description. The "Diagnosis Description" contains the description of the diagnosis given by the doctor. The "Diagnosis Code" represents the id of the disease in the internal system of the hospital. This code also is adapted from the ICD 9 code aforementioned. In other words, this code from this dataset, can be connected to the ICD 9 code from the structured dataset. The "Date" represents the date of the appointment. And finally, the "Clinical Narrative Text", maybe the most valuable variable of this dataset, contains a descriptive narrative of the appointment, written by the doctors themselves. This free text can contain a broad of different elements that can give valuable knowledge of what has been done, the drugs prescribed for each disease, its dosages, etc.

Now that the data used in this word is presented, the next chapter overviews the architecture of the system implemented.

IV. THE ARCHITECTURE OF THE SYSTEM

The system implemented for this paper demanded different steps of preparation for each dataset from its "raw form" to its "processed form" ready to be integrated. The overall process is represented in the Figure 1.

A. *Emergency Department Processing*

The process of preparing this dataset to the final integration involves four steps: removing rows that have the fields "Disease Code (ICD9 Code)" and/or "Disease Code Description" empty, changing all Timestamp data types to Date, creating a new column to accommodate de number of days each patient took in the ED and translating the column "Discharge Destination" since it is still in Portuguese.

Once those changes have ended, the excel file is passed through a python script developed by the authors to turn it into an XML file.

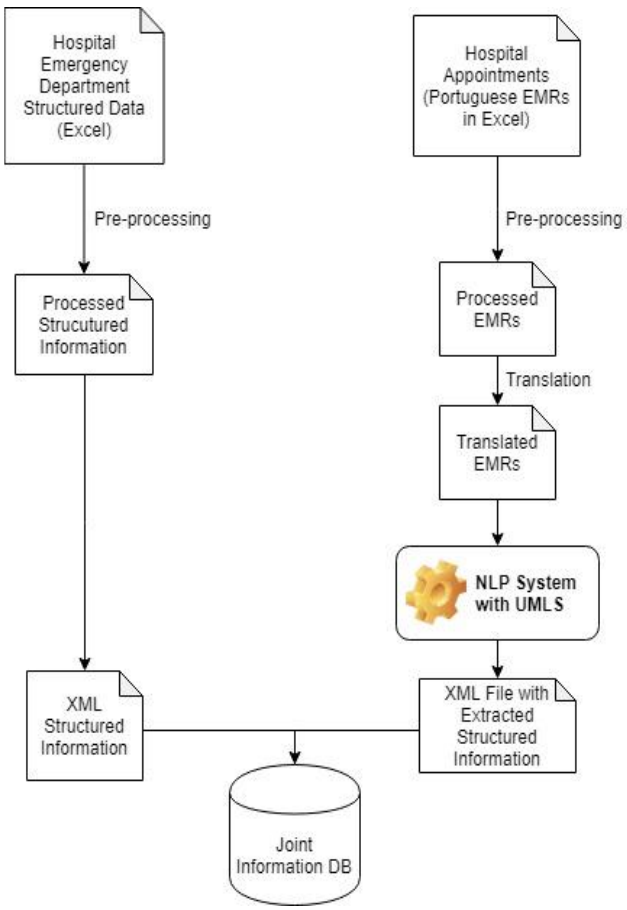

Fig. 1. Generic overview of the process of integration of both datasets, from the excel files to the final conjoint database.

B. *Appointments Department Processing*

The dataset regarding the appointments of the Portuguese hospital require more operations in order to prepare the data to integration.

Firstly, the rows with empty attributes "Diagnosis Code" and/or "Diagnosis Description" were eliminated in the first dataset, as well as in the second.

Secondly, since the attribute "Clinical Narrative Text" is free text, there is no common framework in the many entries of the excel file, which usually results in improper grammatical use, spelling errors, local dialects [5], short phrases and abbreviations [4]. Hence this second step that consists of misspelling correction and unfolding of abbreviations.

Once that text is pre-processed the data is passed through a translator to change it from Portuguese to English, standardizing the language in both datasets. For each row of the dataset new text file is created with the information of its respective row. The used translator consists of a python script developed by the authors that uses a free API from Google Translator[2].

After the dataset is translated, the clinical narratives are ready to be passed through cTAKES, the NLP system selected to be used in the project. The cTAKES system consists of an open-source clinical NLP system implemented in Java. It is "a modular system of pipelined components combining rule-based and machine learning

[2] https://py-googletrans.readthedocs.io/en/latest/

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

techniques aiming at information extraction from the clinical narrative"[10]. There are several other clinical NLP systems but, not only "cTAKES aims to provide best-of-breed NLP modules to the community and facilitates the translation of research into practice" [11], as it has already proven to be successful when applied to summarization [12], information extraction from EMRs regarding ulcerative colitis and Chron's disease [13] and successful identification of patients' smoking status from clinical texts [14].

To enable the identification, and consequently, the extraction of the clinical entities found in the texts, the NLP system database full of clinical terms from the Unified Medical Language System (UMLS). The UMLS is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems [15].

The final result from the NLP system is a set of XML files, one for each text file fed to the system. Each XML file contains the clinical terms found in the written texts, which can go from, medication, to diseases, passing though symptoms, anatomical regions, etc.

C. Database Storage of the Data

After all the processing is done, i.e., the creation of the XML files from both datasets, the hospital emergency department excel, and the hospital's appointment excel, these new files are used to generate a new centralized database that will contain information from two different departments of the hospital.

The database created to integrate the different data sources is an SQLite database considering that the datasets were not too big. As said, since there are two different data sources, the authors opted to develop the database with two different tables, one for each dataset.

A python script was developed to go through the XML files and insert its data into de SQLite database.

The linking point between the two tables is the ICD 9 attribute present in both sides. Therefore, the need to be sure that the attributes "Diagnosis Code" and "Disease Code" were both complacent with the ICD 9.

The two SQLite table attributes are represented in the tables III and IV, the ED SQLite table and the AD SQLite table respectively.

On the one hand, by looking at table III, most of its attributes did not change although some changes are visible.

TABLE III
ED SQLITE TABLE ATTRIBUTES

Attribute #	Attribute Name	Data Type
1	Episode Alert	Integer
2	Patient ID	String
3	ER Sub Department	String
4	Admission Date	Date
5	Triage Color	Integer
6	Nurse ID	String
7	Doctor First Observation ID	String
8	Discharge Date	Date
9	Discharge Status	String
10	Discharge Destination	String
11	Readmission	Flag
12	Disease Code (ICD 9[a] Code)	Integer
13	Disease Code Description	String
14	In Hospital Time	Integer

[a] International Code of Diseases 9

From the initial 15 attributes only 13 got to the final SQLite table.

A new variable was inserted: "In Hospital Time". This variable was calculated as the difference between "Admission Date" and "Discharge Date". This date represents the amount of days from the patient's entry in the ED to its leave.

Another variation regards the change of data types for some of the attributes "Admission Date" and "Discharge Date" that initially were Timestamp but are now Date. Meaning that the temporal data (hours, minutes and seconds) was not considered.

On the other hand, looking at table IV, many differences can be spotted when comparing it with the initial attributes of the excel file since the AD dataset received a lot more processing then the ED dataset.

TAB LE IV
AD SQLITE TABLE ATTRIBUTES

Attribute #	Attribute Name	Data Type
1	Entity Type	String
2	Entity Value	String
3	Medical Specialty Code	Integer
4	Medical Specialty Description	String
5	Diagnosis Description	String
6	Diagnosis Code	Integer
7	Date	Date
8	File Identification	String

The two new columns "Entity Type" and "Entity Value" are two new attributes and are directly related with the cTAKES and UMLS intervention in this project.

The first one can have one of the following values per entry: medication, disease, anatomic region, sign/symptom and clinical procedure. And represents clinical and medical types found in the medical texts of the excel file.

The "Entity Value" present the specific word related with the type, e.g., there can be an entry with "Entity Type" = Medication followed by "Entity Value" = Tocilizumab.

The column "Clinical Episode", which was the identifier of the specific episode on the excel disappeared since its purpose was solely to identify the specific entry of the excel, which is no longer needed.

The "File Identification" attribute was conceived to know which XML file is related with which rows of the table.

In the following chapter the results will be further discussed.

V. RESULTS

As explained in the previous chapters, the main purpose of the work is to create a centralized database with the information from two different hospital departments.

Presented in the figures 2 and 3 are screenshots of the AD SQLite table and ED SQLite table, respectively, also presented in table IV. Although some of the columns are not visible it is possible to see how the information is arranged in the table. How this table connects with the ED SQLite table is exactly as explained in the previous chapter.

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

VI. Conclusion

In this work the authors propound a system that would centralize the data from two different hospital departments (Emergency and Appointments), gathering the structured information from the emergencies, already in system, and EMRs from appointments which contained both structured data and unstructured texts.

The paper shows that the integration of these two different departments is possible, which can lead to future integrations with data from other departments.

Some limitations of this research are as follow: the translation and department difference, since most of the data is from a real Portuguese hospital where Portuguese medical narratives can be found. Since the translation is not perfect, some terms can be mistranslated or not well interpreted by the translator which can lead to loss of valuable information.

The second limitation regards the difference between departments which leads to only one variable linking two different realities.

In the future the authors intend, not only, to apply some python and R algorithms to extract clinical knowledge that allows the establishment of patterns and relations, but also to extend this research to the healthcare practitioners, by associating the clinical knowledge extracted from EMRs with who wrote them and aligning it with the ids present in the ED creating more bridge points between datasets. This way it would be possible to verify, for example, which medication is more prescribed by a given healthcare practitioner, which diagnostic is most common in the ED and the necessary medication and even relate that with specific periods of the year.

	Entity_Type	Entity_Value	Medical_S	Medical_Specialty_Des	Diagnosis_Description	Diagnosis
1	PROCEDURE	Administration Procedure	40695	Rheumatology	Erythema Multiforme	6951
2	ANATOMICALSITE	Administration Procedure	40695	Rheumatology	Erythema Multiforme	6951
3	SIGNSYMPTOM	Administration occupational activities	40695	Rheumatology	Erythema Multiforme	6951
4	ANATOMICALSITE	Blood	40695	Rheumatology	Erythema Multiforme	6951
5	SIGNSYMPTOM	Blood Pressure	40695	Rheumatology	Erythema Multiforme	6951
6	SIGNSYMPTOM	Chief complaint (finding)	40730	Immunohemotherapy	Latent Yaws	1028
7	SIGNSYMPTOM	Chief complaint (finding)	40695	Rheumatology	Erythema Multiforme	6951
8	DISEASE	Communicable Diseases	40691	Ifecciology	Infectious And Parasitic Diseases, Nop Or Not Specified	136
9	SIGNSYMPTOM	Complication	40695	Rheumatology	Erythema Multiforme	6951
10	DISEASE	Disease	40691	Ifecciology	Infectious And Parasitic Diseases, Nop Or Not Specified	136
11	DISEASE	Erythema	40695	Rheumatology	Erythema Multiforme	6951
12	DISEASE	Erythema	40695	Rheumatology	Erythema Multiforme	6951
13	DISEASE	Erythema Multiforme	40695	Rheumatology	Erythema Multiforme	6951
14	DISEASE	Erythema Multiforme	40695	Rheumatology	Erythema Multiforme	6951
15	ANATOMICALSITE	Heart	40695	Rheumatology	Erythema Multiforme	6951
16	PROCEDURE	Interventional procedure	40730	Immunohemotherapy	Latent Yaws	1028
17	MEDICATION	Ivermectin	40691	Ifecciology	Infectious And Parasitic Diseases, Nop Or Not Specified	136
18	DISEASE	Latent Yaws	40730	Immunohemotherapy	Latent Yaws	1028
19	ANATOMICALSITE	Oral Cavity	40691	Ifecciology	Infectious And Parasitic Diseases, Nop Or Not Specified	136
20	MEDICATION	Oral Dosage Form	40691	Ifecciology	Infectious And Parasitic Diseases, Nop Or Not Specified	136
21	DISEASE	Parasitic Diseases	40691	Ifecciology	Infectious And Parasitic Diseases, Nop Or Not Specified	136

Fig. 2. Screenshot partially showing AD_SQLite table of the database developed.

Doctor_First_Observation_ID	Discharge_Date	Discharge_Status	Discharge_Destination	Readmission	ICD_9_Code	ICD_9_Description
M4	201	A	Home - Medical Assistant	0	7804	Dizziness and Giddiness
M5	2001	A	Family Doctor - Health Center Not Specified	0	570	Erythema Infectiosum
M6	2001	A	Unrelated Home - Exterior	0	9309	Foreign Body of External Eye Not Otherwise Spec
M7	2001	A	Family Doctor - Health Center Not Specified	0	5990	Urinary tract infection, site not specified
M8	2001	A	Home - Medical Assistant	0	38611	Benign Paroxysmal Vertigo
M5	2001	A	Family Doctor-Health Center Not Specified	0	7242	Lumbago
M9	2001	A	Family Doctor-Usf	0	7840	Headache
M9	2001	A	Family Doctor-Usf	0	401	Essential Hypertension
M1O	2001	A	Family Doctor-Health Center Not Specified	0	9592	Shoulder and Upper Arm Injury Not Otherwise Sp
M11	2001	A	Family Doctor-Health Center Not Specified	0	78079	Other Malaise and Fatigue
M12	2001	A	Unrelated Home - Exterior	0	E819	Motor Vehicle Traffic Accident of Unspecified Na
M13	2001	A	Internship -Traumatology	0	8120	Closed Fracture of Upper End of Humerus
M4	2001	A	Home-Medical Assistant	0	461	Acute Sinusitis
M14	2001	A	Internment -UIMC	0	42731	Atrial Fibrillation
M15	2001	A	Family Doctor-Health Center Not Specified	0	4660	Acute Bronchitis
M9	2000	A	Family Doctor-Health Center Not Specified	0	2859	Anemia Not Otherwise Specifie
M16	2000	A	Home - Medical Assistant	0	7802	Syncope and Collapse
M15	2001	A	Family Doctor-Health Center Not Specified	0	78052	Insomnia Unspecified
M17	2000	A	Family Doctor-Health Center Not Specified	0	78650	Chest Pain Not Otherwise Specified
M8	2001	A	Home - Medical Assistant	0	463	Acute Tonsillitis

Fig. 3. Screenshot partially showing ED_SQLite table of the database developed.

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

ACKNOWLEDGMENT

This work has been partially supported by Portuguese National funds through FITEC - Programa Interface, with reference CIT "INOV - INESC INOVAÇÃO - Financiamento Base.

REFERENCES

[1] L. Luo *et al.*, "A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system," *BMC Med. Inform. Decis. Mak.*, vol. 16, no. 1, pp. 1–15, 2016.

[2] S. McLane, "Designing an EMR planning process based on staff attitudes toward and opinions about computers in healthcare," *CIN - Comput. Informatics Nurs.*, vol. 23, no. 2, pp. 85–92, 2005.

[3] L. G. Yamamoto and A. N. G. A. Khan, "Challenges of Electronic Medical Record Implementation in the Emergency Department," *Pediatr. Emerg. Care*, vol. 22, no. 3, 2006.

[4] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, "Mining Electronic Health Records: A Survey," vol. 50, no. 6, pp. 1–41, 2017.

[5] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on electronic medical records: A review," *J. Healthc. Eng.*, vol. 2018, 2018.

[6] L. B. Moreira and A. A. Namen, "A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia," *Comput. Methods Programs Biomed.*, vol. 165, pp. 139–149, 2018.

[7] H. Kharrazi *et al.*, "The Value of Unstructured Electronic Health Record Data in Geriatric Syndrome Case Identification," *J. Am. Geriatr. Soc.*, vol. 66, no. 8, pp. 1499–1507, 2018.

[8] A. Sundararaman, S. Valady Ramanathan, and R. Thati, "Novel Approach to Predict Hospital Readmissions Using Feature Selection from Unstructured Data with Class Imbalance," *Big Data Res.*, vol. 13, pp. 65–75, 2018.

[9] "ICD - ICD-9 - International Classification of Diseases, Ninth Revision." [Online]. Available: https://www.cdc.gov/nchs/icd/icd9.htm. [Accessed: 18-Mar-2019].

[10] G. K. Savova *et al.*, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications," *J. Am. Med. Informatics Assoc.*, vol. 17, no. 5, pp. 507–513, 2010.

[11] J. Masanz, S. V Pakhomov, H. Xu, S. T. Wu, C. G. Chute, and H. Liu, "Open Source Clinical NLP - More than Any Single System.," *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, vol. 2014, pp. 76–82, 2014.

[12] S. Sohn and G. K. Savova, "Mayo clinic smoking status classification system: extensions and improvements.," *Mayo Clin. Smok. Status Classif. Syst. Extensions Improv.*, vol. 2009, pp. 619–23, 2009.

[13] A. N. Ananthakrishnan *et al.*, "Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: A novel informatics approach," *Inflamm. Bowel Dis.*, vol. 19, no. 7, pp. 1411–1420, 2013.

[14] G. K. Savova, P. V. Ogren, P. H. Duffy, J. D. Buntrock, and C. G. Chute, "Mayo Clinic NLP System for Patient Smoking Status Identification," *J. Am. Med. Informatics Assoc.*, vol. 15, no. 1, pp. 25–28, 2008.

[15] "Unified Medical Language System (UMLS)." [Online]. Available: https://www.nlm.nih.gov/research/umls/. [Accessed: 16-Mar-2019].

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)