Information extraction from clinical records: an example for breast cancer*

Ivan Lazic
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
ivan.lazic@uns.ac.rs

Niksa Jakovljevic
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
jakovnik@uns.ac.rs

Jasmina Boban
School of Medicine
University of Novi Sad
Novi Sad, Serbia
jasmina.boban@mf.uns.ac.rs

Igor Nosek
School of Medicine
University of Novi Sad
Novi Sad, Serbia
igor.nosek@mf.uns.ac.rs

Tatjana Loncar-Turukalo
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
turukalo@uns.ac.rs

2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON) | 978-1-6654-4280-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/MELECON53508.2022.9842995

***Abstract*—The extraction of relevant information from electronic health records (EHR) can facilitate large scale clinical studies related to certain diseases to uncover diversity of their biological and clinical signatures, and patterns of treatment and prognosis. Variety of EHR formats and use of clinical narrative present significant challenges to this task. In this work we describe a process of an automated information extraction from an oncology hospital clinical reports related to 2966 subjects with suspected or confirmed breast cancer. The lack of open medical term dictionaries for the Serbian language and the variety of clinical data types required, imply the use of rule-based approaches with exact matches, regular expressions, hierarchical rules and customized mini dictionaries to analyze clinical text. The accuracy of the applied approach has been validated on manually extracted clinical data of 50 breast cancer patients. The accuracy varied, field dependent, between 71.3% to 100%, indicating that certain relevant fields can be successfully captured, yet implying the need for sophisticated natural language processing tools for accurate extraction of more descriptive features.**

Keywords*—*electronic health record, information extraction, breast cancer.

I. Introduction

According to the World Cancer Report 2020 [1] cancer remains the first or second leading cause of premature deaths (at age from 30 to 69 years) in 134 of 183 countries. In 2016 there were 4.5 million premature deaths worldwide due to cancer, which accounts for 29.8% of all premature deaths due to noncommunicable diseases [1]. For women, breast cancer remains the most frequently diagnosed cancer type, and the leading cause of cancer death in women worldwide. In 2018 there were 2.1 million new cases and 627,000 deaths [2].

Identification of major risk factors and treatment improvements over the years have resulted in stabilized incidence rates and declining mortality rates in countries with high levels of the human development index [3]. The potential for further contribution lies in the vast amounts of data stored in forms of electronic health records (EHR) which could provide more insight into demographic, biological, clinical and pharmacological patterns of cancer development, treatment and prognosis [4].

The automatic analysis of EHRs is hindered due to the way information is stored in EHRs, which include structured and coded data fields, as well as narrative text in the designated fields, and uploaded documents (reports) [5]. Even though the coded fields provide quality assurance and are used for billing purposes, they limit the possibility to express all medically relevant aspects of disease, thus resulting in information loss [5, 6]. For these reasons unconstrained natural language as a clinical narrative is a common part of EHRs, which hinders development of high throughput clinical applications. There are numerous factors that influence automatic clinical text analysis, such as: lack of clear formatting, conciseness, non-standard abbreviations, typos, information redundancy, and longitudinal health information [7].

Natural language processing (NLP) tools, knowledge management and machine learning (ML) can be used to support (semi-) automatic extraction of clinical information from clinical narratives and identify relevant/targeted medical information [8,9,10,11]. The recent review addressing the use of machine learning for clinical text narratives [11], highlights the common use of rule-based approaches for information extraction from clinical narrative. For the machine learning approach, annotation has been identified as the major obstacle, while dataset size, availability and provenance further hamper the performance of machine learning models [11].

The studies providing different solutions for the clinical text mining in English language are abundant and rely on the knowledge representation systems such as Unified Medical Language System (UMLS) [12], and specialized guidelines (e.g. Breast Imaging-Reporting and Data System, BIRADS [13]) and databases. Many studies used MetaMap [14] to identify the relevant, application specific UMLS terms in clinical text. There are several text mining systems covering different levels of NLP tasks, including information extraction, such as MedLEE system able to process radiology, pathology, electrocardiography, echocardiology, and hospital discharge reports [15]. Another NLP system developed at the Mayo Clinic cTAKES (clinical Text Analysis and Knowledge Extraction System) provides for linguistic and semantic annotations [16]. cTAKES includes the following NLP modules (with indicated accuracy): sentence boundary detector (95%), tokenizer (95%), normalizer, part of speech tagger (94%), shallow parser (F-measure 93%) and name entity resolution annotator (NER) (F-measure of 72% for exact matching and 82% for non-exact matching). Negation and status of named entities were as well identified with F-measure of 96% and 94%, respectively.

*This work has received funding from the EU's H2020 RIA programme under grant agreement No 952179. The authors would like to acknowledge the support of the Faculty of Technical Sciences, Department of Power, Electronics and Telecommunications (DEET) under the project "Development and implementation of state-of-the-art methods in teaching and research at DEET".

978-1-6654-4280-0/22/$31.00 ©2022 IEEE

Numerous studies focus on more advanced NLP tasks, such as information extraction from clinical narrative. In [17] breast pathology reports were analyzed in order to extract pathologic diagnoses. The lexicons were created for each diagnose to include different terms used to denote them, and to provide for a great variability in the narrative with respect to both structure and terminology. The sensitivity of 99% and specificity of 96% was achieved using the dataset of 6711 pathology reports from three hospitals. Another example related to breast cancer aims at BIRADS extraction from 2159 radiology reports from 18 hospitals [18]. Both BIRADS value and laterality were extracted using supervised ML approaches with the best performance being achieved by conditional random fields for the BIRADS value (F measure 0.95), and for laterality with partial decision trees (F measure 0.91). The more comprehensive extraction of information from pathology reports into the predefined structured cancer representation was attempted by the open-source platform MedTAS/P [19]. The evaluation was done using manually annotated set of colon cancer patients. The best results were achieved for instantiating classes such as anatomical sites (F-measure 97–100%). The extraction of information on primary tumors or lymph nodes achieved lower F-measure (82–93%). Uncovering information of metastatic tumors was difficult due to low number of cases, thus it resulted in the F score of 65%.

While there are numerous examples of NLP usage for EHR's text analysis in English, the usage in other languages is modest [20]. In the Serbian language there are some efforts to produce NLP resources for medical terms, mainly devoted to specific tasks, such as marking diagnoses [21], or a medical lexicon in preparation for speech-to-text automated radiological report generation.

This work has resulted from efforts to contribute to the cancer imaging pan-European repository, where imaging data has to be accompanied by structured medical metadata. Specifically, the aim of this study is extraction of medical information related to demographic, medical history, diagnoses, treatment and follow-up of breast cancer patients from their EHRs. The efforts invested in clinical data collection have been part of the broader initiative within the 42-month project INCISIVE (https://incisiveproject.eu) addressing the cancer imaging and clinical data availability to facilitate development of the AI tools in health imaging.

The first data-related challenge in the INCISIVE project was harmonization of data coming from multiple participating healthcare institutions across Europe, which resulted in the common data schema [22]. The second step of data integration brought the solutions to: a) structural embedding to suit all types of information provided, b) data de-identification ensuring privacy, and c) selection of the encodings i.e. standards of medical terminology to be used [22].

Following the proposed data structure, in this work we present the challenges and methodology used to extract relevant mandatory information from EHRs of 2966 patients from a regional oncology hospital situated in the Republic of Serbia. The diverse type of medical terms required, and a need for longitudinal health data introduced an additional layer of complexity. The time constraints and the lack of domain specific open NLP resources for Serbian, have imposed use of the rule-based approaches in search for exact matches and regular expressions, and using hierarchical rules and small specific dictionaries based on features extracted from sample texts and domain knowledge.

II. Materials and Methods

A. Study

Clinical data collected in this study are part of the clinical records of the Oncology Institute of Vojvodina, Novi Sad, Serbia. The data have been collected in a retrospective study approved by the ethical committee of Oncology Institute of Vojvodina under No. 4/20/2-3489/2-3, and supported by the Institute's expert committee (No. 4/20/2-3823/2-9). Using the rigorous European privacy related guidelines, aligned with Serbian law on personal data protection, the medical team members have used INCISIVE standardized pipelines for data de-identification including: CTP DICOM anonymizer [23] for medical images, and customized scripts for removal of personal information and institutional headers from medical reports. The data have been pseudonymized at the current stage, meaning that the hospital preserves correspondence tables between the patient IDs and their internal record numbers for the validation and data curation purposes. At a later stage, this correspondence table will be deleted and data will remain anonymized.

The study included clinical records of breast cancer patients and those suspected of having breast cancer. The inclusion criteria for all subjects was age over 18 and existence of related medical imaging data. Data were collected irrespective of any demographic attribute. All breast cancer patients included in the study have histologically confirmed diagnoses of breast cancer as a primary cancer. A number of patient records corresponded to patients referred to the hospital for lesions that were histologically evaluated as benign, or high risk patients that had appointed breast imaging within the institution. The average age of the women whose data are included in this study is 54.5±10.2 years. As the menopausal status was not always clearly indicated, taking the average age as a threshold, there were altogether1333 subjects older than 54.5 years, and in the group of diagnosed patients 773 out of 1336.

B. Data collection and integration

The clinical data are stored and organized within the hospital information system (HIS), which comprises clinical data reports produced for each patient visit as well as for each service provided. Patients are identified using unique identifiers. In the case of breast cancer patients, this includes reports from radiology department (for each imaging modality), histology department, reports on surgery, laboratory analysis, oncology committee reports on prescribed treatment protocol, radio-therapy, surgery and oncologist appointments used for monitoring treatment and recovery processes. In order to avoid imposing further load to the HIS, the identified patient's data was exported and pseudonymized for further analysis. Laboratory data could be exported in the csv format, all other reports in pdf format, except for the surgery reports which could be exported only as an image (png format). The related medical imaging data was collected from the hospital PACS (Picture Archiving and Communication System) system. Extraction of relevant clinical information into the structured data sheets was done using de-identified clinical reports and imaging data.

The INCISIVE breast cancer clinical data template assumes longitudinal health information collected at diagnosis (baseline, B) and in available time points (TPs) during the patients' treatment (after first treatment, and after each follow-up visit). The clinical data at designated TPs should serve as metadata corresponding to at least one medical image from any of the imaging modalities: mammography (MMG), magnetic resonance images (MRI), computerized tomography (CT), positron emission tomography (PET), ultrasound (US), including histology images. The clinical information required is grouped into the following categories (the number of requested data fields in each is indicated in parenthesis): General information (14), Baseline at diagnosis (29), Follow-up TPs information (36 data fields for each follow up), Treatment (35), Histology (20), and Laboratory information (29). From these, 68 fields were indicated as mandatory fields, together with months from diagnoses where applicable.

These requirements had to be mapped into the corresponding points of the breast cancer healthcare pathway and protocols followed by the public hospitals in Serbia. As presented in Fig.1, the common imaging modalities used at diagnoses and follow-up are MMG and US. Unless otherwise indicated the imaging is performed at the time of diagnosis, and on a yearly basis. In between, patient treatment is defined and monitored by the following specialists: attending oncologist, radiologist, and surgeon. As TPs are aligned with imaging events, all treatment details on the patient journey between the imaging TPs have been collected, extracted and timestamped with months from diagnosis.

C. Challenges in the process of information extraction

In the preparatory analysis steps domain experts' knowledge was used to map the required data fields with the type of reports where this information should be searched. For example, *tumor staging* (using TNM - tumor, nodes and metastasis standard) has been collected based on two different histology reports, one referring to the findings in breast (T) and the other referring to the lymph node analysis (N). There were several challenges that impacted the selection of the methodological approach and performance: (1) the reports do not follow a predefined structure, even in the reports of the same type the level of descriptive details varies, as well as the way the final conclusions are presented; (2) the text is not always written in the form of full sentences, it is characterized with use of abbreviations, uncertainties and assumptions; (3) at some point sentences may be too long, with negation placed quite remotely to the corresponding term; (4) the data fields related to some descriptive characteristic, such as *mass shape* or *mass margin*, are prone to more subjective descriptions and might not fit any of the categories offered in the predefined data template; (5) the type of cancer is commonly written in Latin, but if malignancy is absent, the format of the report and the types of noted lesions can vary widely.

The reports containing the clinical narrative are in the unconstrained language format, including a mixture of the Serbian and Latin languages, commonly used in medical practice for diagnoses, types of interventions (e.g. in surgery), histological findings and treatment protocol. Moreover, there is prominent use of English words, which are not incorporated in the Serbian language (raw Anglicisms, i.e. using English words not adopted to the Serbian language). It is worth noting that the use of Cyrillic script is not common in medical reports.

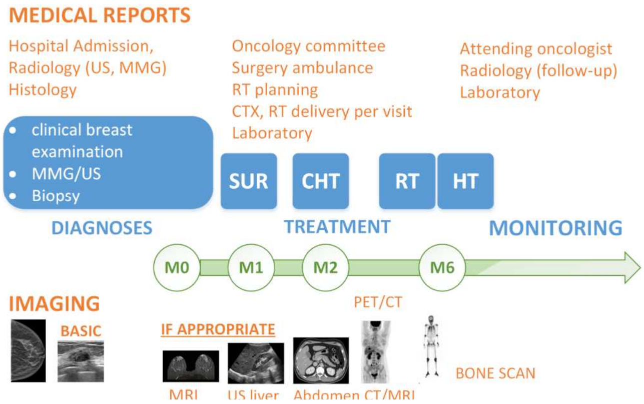

Fig. 1. The basic breast cancer healthcare pathway includes surgery (SUR), chemotherapy (CHT), radio (RT) and hormone therapy (HT). Different types of medical reports produced and different imaginbn g modalities used at each stage are indicated. The time points in months are indicative and case-dependant.

D. Rule-based keyword search

In order to develop the solution for information extraction each of the required data fields was approached separately, or in small groups if their reporting is tightly coupled. The rule based approach implemented in this work required interaction with clinical experts and examination of multiple reports followed by conversion of the findings into a set of explicit pattern-matching rules.

The essential step in the approach taken in this work was laying out the correct time intervals for each observed field. In order to capture the timeline of the patient journey, the dates of mammography images were tracked, and associated to TPs. In cases where several MMG exams preceded in the years prior to diagnosis, the MMG exam labeled as the baseline scan was the first one with the presence of disease. Identification of the presence of disease was based on histopathology reports where applicable, or the BIRADS classification of at least 5, for patients where histopathology reports were not available (i.e. HP was performed in another institution). If the patient had these conditions fulfilled multiple times (in situations of relapse), the earliest moment is selected as the baseline point. Afterwards, if there is a clear baseline position, the following MMG reports are added as the successive timepoints if they meet a condition that they reference themselves to the previous report which is a standard practice. For patients that do not have a set baseline, histology done in the hospital, or have no malignant findings, the first breast imaging report was taken as the starting point and the timeline was built from there onward.

According to the constructed timeline, the fields in the required table are potentially filled by scanning the related documents that fit within the time range. The current iteration of the procedure relies on rule-based approach by detecting exact matches to certain keywords or keyword patterns to fill in a true/false type of field (presence or absence). Due to the complexity of Serbian grammar which assumes changing of nouns by the case, the base of the word for each keyword was searched for (analogous to stemming). Additionally, for each keyword flat dictionaries were formed of all synonyms identified based on the sample text in both Serbian and Latin. language. In order to extract detailed information (such as any numerical values) additional searches were needed limited to paragraphs containing the keywords of interest. Certain values are easy to spot as they come in mostly predetermined patterns of letters. For example, metric measurements are easier to find as they're often followed by SI units.

Special care was also taken into account for use of negation terms. These are often found near if the narrative style is telegraphic. However, there are still persistent issues as the complexity of the report grows (such as in longer, dependent sentences most often). Without any sophisticated linguistic analysis, negation was only detected by its presence in a nearby surrounding of the observed keyword.

E. *Performance validation*

50 breast cancer patients diagnosed, treated and monitored within the institution were selected as a validation set. These patients had longitudinal health data for at least four years, including baseline diagnosis and at least three more time points of patient imaging and clinical data. It is worth noting that even with careful manual extraction, human errors are possible and were found and corrected in the second pass through the patient documents.

III. Results

The clinical reports of 2966 patients have been downloaded, pseudonymized and analyzed. As the hospital is a regional healthcare institution, patients can be referred to the hospital at a later stage, after receiving histologically verified diagnosis, for surgery and/or treatment. Based on the automated analyses we identified 600 patients that were diagnosed, treated and monitored within the hospital (denoted G1), 736 treated and/or monitored (G2), and 1630 patients referred to the hospital but not diagnosed with malignant disease (G3). The patients were classified in these groups based on the presence and content of one of the following: histology (Latin diagnoses of malignancy) or oncology committee reports, TNM staging or in G2 patients in the absence of other reports BIRADS ≥5 was used as criterion.

In order to evaluate the performance and analyze possible sources of confounding, for each data field the suggested value was output, together with the information on the report from which it was extracted, and the corresponding sentence or paragraph. Through multiple iterations of the described process, many synonym medical terms were identified and included in dictionaries.

The manual information extraction by human readers for each patient in the validation set took on average 2.5 hours. Additionally, from the total of 235 data fields required for each patient, the information was present for at most 152 data fields, due to the level of details required, absence of all imaging modalities and biological data. The data extracted using the proposed expert system were compared to the human reader data on the subset of 50 patients. The accuracy, measured as the fraction of correctly identified clinical data for some relevant data fields, is summarized in Table 1. The information on tumor size is sometimes confounded with some other measures in the examination report, while BIRADS and TNM if reported, are quite reliably extracted. The accuracy of chemo therapy protocol (CHT) and hormone therapy (HT) depends on the quality of the dictionaries comprising different types of protocols and medications. The information on radio therapy (RT) is often incomplete in the reports, without clear information on the delivered dose and the number of fractions. The type of surgery (SUR), provided in Latin, achieves accuracy of 83.3%, and confusion arises mainly due to abbreviations. The information on Laboratory data can be extracted with 100% accuracy if the analysis is done in the hospital, while if the patient brings externally obtained lab results, basic information can be found in oncology reports and are more difficult to follow due to inconsistent formatting. From the general data, age at diagnosis is the field captured correctly without exception. In the groups with cancer diagnoses the minimum age was 26 years, and maximum 91, with a normal distribution (57.36±10.63 years). The data related to menopause, familial cancer history and number of births were often missing. In G1 and G2 (1336 patients), where available, the records show that in 113 patients there was a cancer history from the mother's side, and in 101 patients from the father's side. In G3 from 1630 patients, the available information confirmed the cancer history in 17 cases from the mother's side, and in 14 cases from the father's side.

Data fields related to the time point of diagnosis: type of images taken, BIRADS classification, breast density and information related to biopsy results (tumor type, grade), and TNM staging whenever available, were reliably extracted. Fig. 2a presents the distribution of imaging modalities in the studied data set, used at the time of diagnoses and in the follow up visits. It is obvious that MMRs are the most common modalities, supported by MRI in diagnoses, while PET imaging relevant for metastasis is the least frequent. The distribution of BIRADS in patient with some malignant findings (G1, G2) and without (G3) is presented in Fig 2b. Related to the biopsy data, the information on estrogen (ER), progesterone (PR) and HER2+ receptors and tumor grade could not be extracted at all times. For example, in the G1 group there are 243 patients with surgery reports, in 216 the tumor grade has been detected, and in 243 ER, PR and HER2+ extracted.

For further follow-up the most relevant information was related to the applied treatment, changes in treatment, and patient status. Besides laboratory data, the relapse or spread of disease can be determined by analysis of radiology reports as described by BIRADS, description of changes, or the absence/presence of suspicious lesions in abdominal, chest, bone or PET scans. Whereas BIRADS classification is successfully extracted from the imaging reports, the reports for other body parts are very descriptive, utilizing complex medical domain terminology. The analysis of these reports

TABLE I. Performance Evaluation Against Human Reader

Tumor size	*BIRADS*	*TNM*	*Breast density*	*SUR*	*CHT*	*RT*	*HT*
71.4%	97.6%	90.5%	83,3%	83,3%	85,7%	83.3%	90.5%

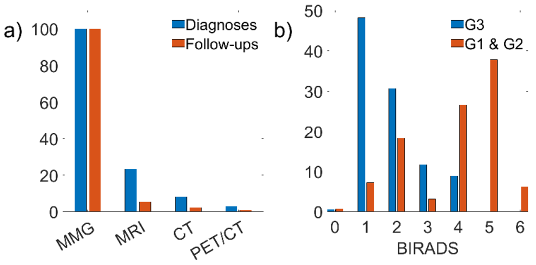

Fig. 2. a) The percentage of different type of medical images in breast cancer diagnosed patients (G1&G2) at the time of diagnoses and in the follow-up visits; b) the percentage of BIRADSs found between the diagnosed (G1&G2) and patients with no malignancy (G3).

would require more advanced NLP tools, unless information is summarized clearly in attending oncology reports by additional information on metastatic spread.

IV. Discussion and Conclusions

The paper describes an expert based system for an automatic extraction of relevant clinical metadata accompanying medical images based on medical reports of breast cancer patients. The initial efforts mainly rely on the rule-based approaches such as exact matching and regular expressions. The development and implementation of the automated analysis came upon several challenges that introduced error messages or prevented information extraction. The heterogeneity of the required data fields and a need to create small dictionaries for a large number of medical terms within unconstrained clinical narrative presented the main difficulty.

The extraction of timeline in longitudinal health data is challenging due to several factors. Missing reports due to patient transfer to the hospital at the later stage (or due to some other reasons) hinder the timeline reconstruction. In this specific scenario, where imaging data was required, the radiologist reports tracked in time are valuable for recovering the patient monitoring trajectory. As medical images of the same type are compared to the images in the previous visit, errors occur if radiologist mistakenly report the date or refers to an image non existing in the hospital archive. Once the time points of visits are extracted, all other events related to treatment and monitoring are extracted based on oncology/histology/surgery reports from the designated period.

Finding certain numerical information from the clinical text is usually resolved with finding the corresponding key word in the text, and looking at the adjacent text on both sides as it is not uniform throughout reports. For example, for the maximum diameter of tumor the extraction is more successful from the radiology reports, as histology reports usually describe in more detail all changes from the analyzed tissue. However, it might happen that additional numerical position-related descriptions in radiology reports lead to erroneous outputs ("*…whole central part of the breast is occupied with dusty micro calcifications with tendency of grouping in the area of 5cm with multifocal infiltrated shadows underneath with diameter 8x14mm….*"). In this example, by the means of automated analysis 14 mm is extracted as maximum diameter of the change, whereas the whole structure extends in the area of 5cm. Further improvements can be made by incorporating the basic NLP tools available for the Serbian language, like stemmers and negation identification. Even at this stage, the developed tool offers efficient solution to cumbersome time-consuming manual data extraction and provides for fast identification of relevant patient categories.

Further improvement of this work should include the basic NLP operations, such as linguistic and semantic annotations including: sentence boundary detector, tokenizer, normalizer, part of speech tagger, shallow parser and name entity resolution annotator, such as those provided for English by [15, 16]. Negation is very relevant as it can be expressed in multiple ways, which should be comprehensively summarized. Taking in consideration time consuming manual information extraction, the proposed solution facilitates fast automated generation of the structured reports, offering possibility for fast manual validation if needed.

This paper presents complex information extraction from EHR into structured time-resolved breast cancer representation templates, including multiple data fields from several types of clinical reports collected over time. The results presented in this paper can not be directly compared to the similar results from the clinical text mining literature for English language, due to differences in task complexity, data set used and maturity of NLP tools for Serbian language and this specific application domain. When compared at the level of specific data fields, such as BIRADS or tumor type or grade, comparable performance is largely due to available adherence to certain guidelines, knowledge ontologies, international classification of disease, or tumor grading conventions. The drop in performance is associated with more demanding fields, such as those that require identification of some relationship, such as information on disease spread into lymph nodes [19].

Efforts should be invested on training of the medical personal and even automating data inputs to EHR to minimize linguistic and structural variability. The medical reports should be consistent, reporting uniformly and using the same guidelines at least within an institution. Harnessing the full potential of information present in EHR thus requires awareness and efforts from both clinical and ML communities.

References

[1] C.P. Wild, E. Weiderpass, and B.W.Stewart, editors, *World Cancer Report: Cancer Research for Cancer Prevention*. Lyon, France: International Agency for Research on Cancer, 2020.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J Clin*., vol. 68(6), pp.394–424, 2018.

[3] M. Arnold, H.E. Karim-Kos, J.W.Coebergh, G. Byrnes, A. Antilla, J. Ferlay, et al., "Recent trends in incidence of five common cancers in 26 European countries since 1988: analysis of the European Cancer Observatory," *Eur J Cancer.* 51(9), pp.1164–87, 2015.

[4] C. Friedman, "Semantic text parsing for patient records," *In Medical informatics*, Springer, Boston MA, pp. 423-448, 2005.

[5] A. Pomares-Quimbaya, M. Kreuzthaler, and S. Schulz, "Current approaches to identify sections within clinical narratives from electronic health records: a systematic review," *BMC medical research methodology*, vol. 19 (1), pp.1-20, 2019.

[6] S.B. Johnson, S. Bakken, D. Dine, S. Hyun, E. Mendonça, F. Morrison, T. Bright,T. Van Vleck, J. Wrenn, and P. Stetson P. "An electronic health record based on structured narrative," *JAMIA*, vol. 15(1), pp.54–64, 2008.

[7] N.G. Weiskopf, G. Hripcsak, S. Swaminathan, and C. Weng, "Defining and measuring completeness of electronic health records for secondary use," *J Biomed Inform*, vol. 46(5), pp. 830-836, 2013.

[8] Y. Wang, L.Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, et al., "Clinical information extraction applications: A literature review", *J Biomed Inform*, vol. 77, pp.34–49, 2018.

[9] R. Altman, "Artificial intelligence (AI) systems for interpreting complex medical data sets", Clinical Pharmacology & Therapeutics, vol. 101(5), pp.585-586, 2017.

[10] Spasić, I., Livsey, J., Keane, J. A., & Nenadić, G. (2014). Text mining of cancer-related information: review of current status and future directions. International journal of medical informatics, 83(9), 605-623.

[11] Spasic I, Nenadic GClinical Text Data in Machine Learning: Systematic Review JMIR Med Inform 2020;8(3):e17984 doi: 10.2196/17984

[12] O. Bodenreider The Unified Medical Language System (UMLS): integrating biomedical terminology Nucleic Acids Res., 32 (2004), pp. D267-D270

[13] E.S. Burnside, E.A. Sickles, L.W. Bassett, D.L. Rubin, C.H. Lee, D.M. Ikeda, et al. The ACR BI-RADS® experience: learning from history J. Am. Coll. Radiol., 6 (2009), pp. 851-860

[14] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium (p. 17). American Medical Informatics Association.

[15] C. Friedman A broad-coverage natural language processing system AMIA Symposium, Los Angeles, CA, USA (2000), pp. 270-27

[16] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kippe r-Schuler, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications J. Am. Med. Inform. Assoc., 17 (2010), pp. 507-513

[17] J.M. Buckley, S.B. Coopey, J. Sharko, F. Polubriaginof, B. Drohan, A .K. Belli, *et al.* The feasibility of using natural language processing to extract clinical information from breast pathology reports J. Pathol. Inform., 3 (2012), p. 23

[18] Castro, S. M., Tseytlin, E., Medvedeva, O., Mitchell, K., Visweswaran, S., Bekhuis, T., & Jacobson, R. S. (2017). Automated annotation and classification of BI-RADS assessment from radiology reports. Journal of biomedical informatics, 69, 177-187.

[19] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K.S.J. Cooper, *et al.*Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model, J. Biomed. Inform., 42 (2009), pp. 937-949

[20] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than english: opportunities and challenges," *Journal of biomedical semantics*, vol. 9(1), pp.1-13, 2018.

[21] U. Marovac, A. Avdić, D. Janković, and S. Marovac, "Creating Resources for Marking Diagnoses in Electronic Health Reports in Serbian," *International Journal of Electrical Engineering and Computing*, vol. 4(1), pp. 18-23, 2020.

[22] A. Kosvyra, D. Filos, D. Fotopoulos, T. Olga, and I. Chouvarda, "Towards Data Integration for AI in Cancer Research," *In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2054-2057, 2021.

[23] RSNA, "CTP-The RSNA Clinical Trial Processor." [Online]. Available: http://mircwiki.rsna.org/index.php?title=CTP-The_RSNA_Clinical_Trial_Proces