Analysis of Electronic Health Records to Identify the Patient's Treatment Lines: Challenges and Opportunities

Marjan Najafabadipour[1(✉)], Juan Manuel Tuñas[1], Alejandro Rodríguez-González[1,2], and Ernestina Menasalvas[1,2]

[1] Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Madrid, Spain
{m.najafabadipour, alejandro.rg, ernestina.menasalvas}@upm.es, juan.tunas@ctb.upm.es
[2] ETS de Ingenieros Informáticos, Universidad Politécnica de Madrid, Madrid, Spain

Abstract. The automatic reconstruction of the patient's treatment lines from their Electronic Health Records (EHRs) is a significant step towards improving the quality and the safety of the healthcare deliveries. With the recent rapid increase in the adaption of EHRs and the rapid development of computational science, we can discover new insights from the information stored in EHRs. However, this is still a challenging task, being unstructured data analysis one of them. In this paper, we focus on the most common challenges for reconstructing the patient's treatment lines, which are the Named Entity Recognition (NER), temporal relation identification and the integration of structured results. We introduce our Natural Language Processing (NLP) framework, which deals with the aforementioned challenges. In addition, we focus on a real use case of patients, suffering from lung cancer to extract patterns associated with the treatment of the disease that can help clinicians to analyze toxicities and patterns depending on the lines of treatments given to the patient.

Keywords: Electronic Health Records · Natural Language Processing · Named Entity Recognition · Temporal relation identification

1 Introduction

Treatments target the symptoms, the disease, the impairments in physical and psychosocial functioning, disabilities, comorbidities, and the trajectory of the disorder [1]. This makes the detection of treatment lines from the clinical texts a fundamental task in the clinical information extraction, where a treatment line is a collection of drugs with their dosage and its starting and ending time points. The detection of treatment lines has several applications in the medical research such as assessing the healthcare quality [2], understanding the patient's treatment course [3] and improving the detection of adverse drug reaction [4].

Towards the digitization of medical data, clinicians chronologically record the details of the patient-clinician encounters in the computerized documents, known as

M. Bramer and M. Petridis (Eds.): SGAI-AI 2019, LNAI 11927, pp. 437–442, 2019.
https://doi.org/10.1007/978-3-030-34885-4_33

"Electronic Health Records (EHRs)" [5]. EHRs are therefore textual (unstructured) clinical documents containing several medical events related to the patient's treatment and the corresponding chronological sequence, which allows the reconstruction of the treatment lines. A typical treatment line can be composed of many EHRs, including tens of thousands of words. This makes the manual analysis of EHRs for identification of such a line a time-consuming and costly task. For this reason, the automatic discovery of treatment lines from clinical texts requires a great attention.

Identification of treatment lines includes several challenges: (1) NER, a paramount step of NLP to extract drug concepts, dosage metrics and time expressions from clinical texts; (2) temporal relation identification, to link treatment concepts to time expressions; and (3) the integration of structured results, to deal with the redundant information and to reconstruct the treatment lines.

Although, several NLP systems have been developed for extraction of information from clinical texts such as Apache cTAKES [6], MEDLEE [7], MetaMap [8], H2A [9], C-liKES [10], to name a few, the problem of the discovery of treatment lines from all the patient's EHRs still remains unsolved.

In this paper, we deal with the challenges associated with NER, temporal relation identification and the integration of structured results for a specific use case related to lung cancer domain. The EHRs of lung cancer patients used in our studies are available in Spanish. The main objective of this research is to contribute to the existing solutions by providing a prototype, indicating that analyzing EHRs enables the reconstruction of the patient's treatment lines. To do so, an NLP framework together with built in modules to extract concepts, detect temporal relations and build treatment lines is being designed.

The rest of paper is organized as follows: Sect. 2 explains the challenges associated with reconstructing the patient's treatment lines from their EHRs. Afterwards, Sect. 3 is dedicated to explaining our framework and its application to lung cancer domain. Finally, Sect. 4 describes the conclusions and future works.

2 Challenge for Reconstructing the Patient's Treatment Lines

Reconstruction of the patient's treatment lines from EHRs entails three main challenges: (1) NER; (2) temporal relation identification; and (3) the integration of structured results. These challenges are discussed in detail in the following Sub-sections.

2.1 Named Entity Recognition Challenges

EHRs contain a vast amount of valuable information written in narrative form, which lacks structure or have a structure depending on the hospital, service or the clinician generating them. Thus, the extraction of information from clinical texts is difficult.

Annotation of treatment events is highly dependent on the dosage metrics. Within clinical texts, recognition of these metrics introduces three main challenges. First of all, although the NER process relies on ontologies such as SNOMED [11] and UMLS [12],

these ontologies are limited to completely provide dosage metrics. Secondly, abbreviations are integral part of dosage metrics; it is thus difficult to assign semantics to them. Finally, the dosage metrics can be mentioned as simple as including only one variable or as complex as including several variables, which are very common, and yet are difficult to decode in an exact way.

Another interesting challenge is related to the recognition of time expressions from EHRs due to the limitation of ontologies to provide them, various formats, styles, categories (i.e., relative and absolute) a time expression can be written in, and the difficulty to interpret relative time expressions.

2.2 Temporal Relation Identification Challenges

Clinical texts include complex, diverse and sometimes, non-standard linguistics mechanisms to mention temporal relations. In addition, in some cases, the time point associated to the medical event is not even explicitly mentioned in EHRs. These make the automatic detection of temporal relations a very challenging task.

2.3 The Integration of Structured Results Challenges

The problem of Information redundancy is a fundamental concept associated with EHRs due to the interest of clinicians to "cut and paste" texts from past EHRs for summarizing past information in the newly generated EHRs. This creates another layer of complexity to reconstruct the patient's treatment lines from EHRs as many redundancies and references to the past treatments can appear with the current treatment lines. In addition, it can happen that a treatment line has to be discontinued due to no effect, toxicities or the side effects. This challenge also should be tackled in order to find the lines of treatments of the patients.

3 Solution

The main goal of this research is to be able to reconstruct the patient's treatment lines. As mentioned in Sect. 1, we are working specifically on a use case related to the patients suffering from lung cancer. Therefore, we present our framework to analyze EHRs in order to reconstruct the lung cancer patient's treatment lines (Fig. 1). Our framework is responsible for analyzing EHRs to annotate concepts from clinical texts, identify temporal relations and build the treatment lines.

We will describe in what follows, the annotators used and developed to be able to recognize drug concepts, dosage metrics and time expressions from clinical texts. In NER, the first step is to annotate concepts using standard ontologies. To recognize drug concepts from clinical texts, we use the UMLS annotator of C-liKES [10], which is built upon the Unstructured Information Management Architecture (UIMA) framework. The UMLS annotator can identify noun and noun phrases concepts that have relevant matches in the UMLS ontology. We here focus on recognizing the specific treatments concepts of tyrosine kinase inhibitor, chemotherapy, radiotherapy, immunotherapy, and antiangiogenic for our implementation.

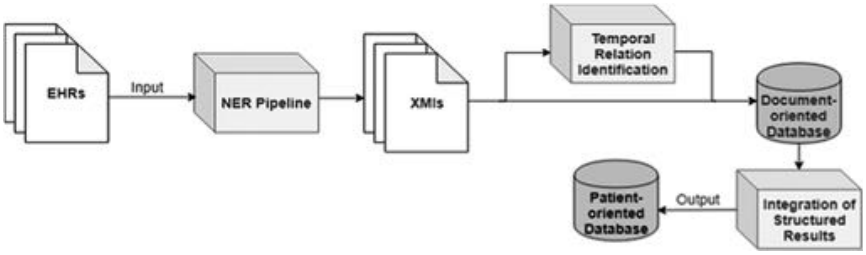

Fig. 1. Architecture of our framework

Apart from the treatment concepts, new annotators should be developed for annotation of dosage metrics and time expressions as they are not provided in the UMLS. To recognize dosage metrics associated with drugs for the aforementioned lung cancer treatments, a rule-based NLP annotator is being developed. In addition, to extract and normalize time variables appeared in the clinical texts, we use a rule-based NLP annotator built over UIMA framework, named Temporal Tagger that is presented in a previous work of the authors [13].

To process EHRs using these annotators, we have implemented them under a single NER pipeline. Once the clinical texts are ingested, the outcomes of annotation process are stored in a set of XML Metadata Interchange (XMI) files.

Then, a temporal relation identification process is implemented using a rule-based approach to link annotated time expressions to treatment concepts in clinical texts. Although once the information in EHRs has been annotated, one could use search engines to retrieve the information. However, our aim here is to extract specific patterns that can be used for reconstruction of treatment lines.

Afterwards, the information stored in XMI files and generated by the temporal relation identification process is stored into a document-based relational database. As this database only provides insights to the information at document level and does not facilitate the integration of information for patients, so we cannot query the treatment lines for patients. Therefore, the integration of structured information is still required for reconstructing the patient's treatment lines.

At this stage, a specific module is developed to integrate the information of document-oriented database and to deal with information redundancy. As each patient can have many EHRs generated for him during his treatment course, several redundant information is included in these clinical documents. Therefore, there is a need for development of a specific post-process module to deal with information redundancy to be able to then reconstruct the lines of treatments. For this purpose, an algorithm with a set of heuristics rules is being developed that is based on the clinician's knowledge and experience for determining what kind of treatments with specific dosage can be prescribed for the patients at the same or different time intervals. This algorithm accepts the structured information of document-oriented database and follows the steps discussed below:

- For each treatment type, temporally order the treatment concepts. Then, select the earliest mention of its drug and dosage from EHRs, and start the treatment line X.

- While the end time point of X is not found, include all the mentioned unique drug concepts with their dosage from EHRs in X. The end time point of X is found when:
  - More than N months have been passed without the mention of the last drug concept of X within EHRs or without the mention of a new compatible drug. Note that, the value of N for months is different for each treatment type.
  - A new drug with a specific dosage is mentioned in an EHR that is not compatible with other drugs in X.

Once, the above algorithm is implemented and the lines of treatments are identified for each patient, they will be stored into a patient-oriented database from which query and answering process can be followed for having the detailed information for each line. Figure 2 presents an example of the output stored by our framework in the patient-oriented database for a patient, who has gone through the chemotherapy treatments.

Finally, Fig. 3 depicts the summary of the concepts extracted from EHRs towards generating the patient's treatment lines.

Id	Treatment	Line	Init	Finish	Drugs	Dosage
6695	chemotherapy	1	2017-06-01	2017-08-07	carboplatino	396 mg
6695	chemotherapy	2	2017-11-08	2018-02-22	docetaxel|pemetrexed	75 mg/m2|500 mg/m2

Fig. 2. Example of the output of our framework in the patient-oriented database

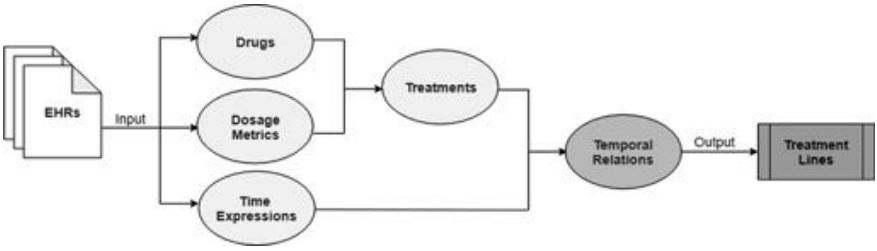

Fig. 3. Concepts extracted From EHRs towards reconstructing the patient's treatment lines

4 Conclusion and Future Work

In this paper, we have analyzed challenges associated with the process of reconstructing the patient's treatment lines from their EHRs. We have focused on the NER, temporal relation identification and the integration of structured information. This work is an ongoing research in which future works will be aimed at the validation and the improvement of the framework. However, the validation of each of the modules in the framework is a difficult and time-consuming task as it requires the manual inspection of the EHRs to check their performance accuracy. In addition, it is significant to note that some of the steps of temporal relation identification and the integration of structured results are not yet completely automatic as they are dependent on the way the clinical texts are written. Thus, future improvements go to automatizing these processes completely.

Acknowledgment. This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 727658, project IASIS (Integration and analysis of heterogeneous big data for precision medicine and suggested treatments for different types of patients). MN is also supported by UPM (Universidad Politécnica de Madrid) Programa Propio of PhD grants.

References

1. Ursano, R.J.: Disease and illness: prevention, treatment, caring, and health. Prev. Chronic Dis. **8**(6), A128 (2011)
2. Roth, C.P., Lim, Y.-W., Pevnick, J.M., Asch, S.M., McGlynn, E.A.: The challenge of measuring quality of care from the electronic health record. Am. J. Med. Qual. **24**(5), 385–394 (2009)
3. Ghitza, U.E., Sparenborg, S., Tai, B.: Improving drug abuse treatment delivery through adoption of harmonized electronic health record systems. Subst. Abuse Rehabil. **2**, 125–131 (2011)
4. Liu, M., et al.: Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J. Am. Med. Inform. Assoc. **19**(e1), e28–e35 (2012)
5. Najafabadipour, M., Tuñas, J.M., Rodríguez-González, A., Menasalvas, E.: Lung cancer concept annotation from Spanish clinical narratives. In: Auer, S., Vidal, M.-E. (eds.) DILS 2018. LNCS, vol. 11371, pp. 153–163. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-06016-9_15
6. Savova, G.K., et al.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Assoc. **17**(5), 507–513 (2010)
7. Friedman, C., Hripcsak, G., DuMouchel, W., Johnson, S.B., Clayton, P.D.: Natural language processing in an operational clinical information system. Nat. Lang. Eng. **1**(1), 83–108 (1995)
8. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium, pp. 17–21 (2001)
9. Menasalvas, E., Rodriguez-Gonzalez, A., Costumero, R., Ambit, H., Gonzalo, C.: Clinical narrative analytics challenges. In: Flores, V., et al. (eds.) IJCRS 2016. LNCS (LNAI), vol. 9920, pp. 23–32. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47160-0_2
10. Menasalvas Ruiz, E., et al.: Profiling lung cancer patients using electronic health records. J. Med. Syst. **42**(7), 126 (2018)
11. SNOMED International. https://www.snomed.org/. Accessed 13 Jul 2018
12. Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/. Accessed 4 May 2018
13. Najafabadipour, M., et al.: Recognition of time expressions in Spanish electronic health records. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), pp. 69–74 (2019). https://doi.org/10.1109/CBMS.2019.00025