A data-driven approach for analyzing healthcare services extracted from clinical records

Manuel Scurti[1], Ernestina Menasalvas-Ruiz[1,2], Maria Esther Vidal-Serodio[3], Maria Torrente[4], Dimitrios Vogiatzis[5], George Paliouras[5], Mariano Provencio[4], Alejandro Rodríguez-González[1,2]

[1]Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Pozuelo de Alarcón, Madrid, Spain

[2]ETS Ingenieros Informáticos, Universidad Politécnica de Madrid, Madrid, Spain
{alejandro.rg, ernestina.menasalvas}@upm.es

[3]TIB Leibniz Information Centre for Science and Technology, Germany

[4]Hospital Universitario Puerta de Hierro de Majadahonda, Spain

[5]Inst. of Inf. and Tel. National Center for Scientific Research "Demokritos", Greece

***Abstract*— Cancer remains one of the major public health challenges worldwide. After cardiovascular diseases, cancer is one of the first causes of death and morbidity in Europe, with more than 4 million new cases and 1.9 million deaths per year. The suboptimal management of cancer patients during treatment and subsequent follows up are major obstacles in achieving better outcomes of the patients and especially regarding cost and quality of life In this paper, we present an initial data-driven approach to analyze the resources and services that are used more frequently by lung-cancer patients with the aim of identifying where the care process can be improved by paying a special attention on services before diagnosis to being able to identify possible lung-cancer patients before they are diagnosed and by reducing the length of stay in the hospital. Our approach has been built by analyzing the clinical notes of those oncological patients to extract this information and their relationships with other variables of the patient. Although the approach shown in this manuscript is very preliminary, it shows that quite interesting outcomes can be derived from further analysis.**

Keywords – data-driven, lung cancer, patient management, EHR, hospitalization processes

I. INTRODUCTION

Cancer is a major public health issue both for its mortality and incidence per year as well as its high costs. According to data provided by the European Commission, in 2018, around 4 million people have been diagnosed with cancer and about 1.9 million have died. Quantification of the real economic burden needs not only an estimation of the costs in the health-care system but also an estimation of the lost earnings associated with the inability to work. According to [1], treating cancer costed the EU €126 billion in 2009. The productivity losses because of early death sum around €42.6 billion and lost working days around €9.43 billion. Lung cancer has the highest economic cost (€18.8 billion, 15% of overall cancer costs) [1]. As can be seen, cancer is a disease that not only causes deaths, but also economic loss in terms of the disease treatment but also on its management. Different variables that are present during the diagnosis, treatment, and follow-up of cancer patients represents also an economical cost that needs to be improved. Detecting the main problems associated with the management of the disease can impact directly in the economy as well as in improving patient's quality of life. For this reason, it is necessary to analyse the disease also from a management point of view.

Several approaches have focused in the past in the use of different sources of information to improve the knowledge available about a certain disease. Some of the most relevant works, closely related with the main aspects to be discussed in this paper, are referred to Length of Stay (LOS); it is one of the main variables that affect both the economic impact of a disease in a hospital, as well as the quality of life of the patients suffering from a specific disease. In [2], an analysis of how different variables correlate with LOS by using Electronic Health Record (EHR) admission data is presented. In this case, data is collected from all the hospital services and the main goal of the analysis is to improve the management of LOS among patients. Another similar approach is followed in [3] to create a model for predicting survival and length of stay in critically ill patients using sequential organ failure scores, or in [4] with older adults after cardiac surgery or in [5] about the determining factors associated with prolonged LOS in patients following cardiac surgery.

In this paper, we present a data-driven pipeline to improve lung cancer patient management. In particular, we show how clinical notes have been used after a process of Natural Language Processing (NLP) that extracts information that allows for retrieving several clinical parameters [6] and for calculating others such as the length of stay in hospital (objective 1) and the reasons of emergency assistance, number of days of hospital stay prior and after cancer treatments. These features can be used, along with patient's clinical characteristics, to detect those patients that, depending on their clinical characteristics, can consume more often hospital care resources (e.g., emergency visit, admission, and readmission). From this information, the hospital can create proactive interventions to avoid it in safe conditions and improving the patient satisfaction, their outcomes, and reducing costs for the healthcare system. On the other hand, but with a related objective, we aim to identify the most frequently used services

and the most common clinical causes that lead to their use, before cancer diagnosis. This will enable the hospital services to recognise patients that might be at risk of developing lung cancer (objective 2), and provide an early diagnosis, which will result in an improvement of the treatment effectiveness and the patient´s outcome in such cases.

II. MATERIALS AND METHODS

Data recorded between January 2008 and December 2018 of patients diagnosed with lung cancer tumor were extracted from the EHR of "Hospital Universitario Puerta de Hierro de Majadahonda", in Madrid, Spain. Data contains around 300K clinical notes of 967 patients. The notes contain written clinical information in Spanish. For each patient, all the clinical notes generated in their visits and tests performed in the hospital were collected. In this context, a process is a specific event that had happened within the clinical history of a patient, related to a specific hospital service. However, there is no information regarding the processes in the clinical notes. Hence, only by analyzing the dates and the content of the notes, the associated processes could be inferred.

To analyze information regarding the two mentioned objectives (length of stay and identification of risk cancer), it is necessary to group notes containing semantically related content. For each patient we can describe the main reason for visiting the hospital. The following set of processes has been validated by the health professionals:

1. *Consultations*: Correspond to scheduled visits to a hospital service to see a specialist doctor.
2. *Day Hospital-Home*: A patient goes to the Day-care hospital to receive treatment (chemotherapy) after visiting the oncologist and returns home.
3. *Day Hospital-Hospitalization*: As the previous process, but the patient has to be hospitalized after or during treatment.
4. *Emergency-Hospitalization:* The patient visits the ER due to poor health status or severe condition, and has to be hospitalized.
5. *Emergency-Home*: As in the previous process, but due to the patient´s improvement he is discharged and goes home.
6. *Home-Hospitalization:* Normally, it corresponds to processes of scheduled surgeries and the patient is admitted straightaway.
7. *Hospitalization-Death*: The patient dies during their hospitalization.
8. *Emergency-Death:* The patient dies after admission in the ER.
9. *Non-Classified:* This category groups clinical notes that could not be classified in any of the previous categories.

The NLP process [6] is performed to annotate clinical notes with concepts such as: hospitalization/ER date, date of discharge (from hospital or from ER), diagnosis, reason for ER visit, and service. Figure 1 depicts the pipeline followed to annotate and categorize the notes in the set of processes.

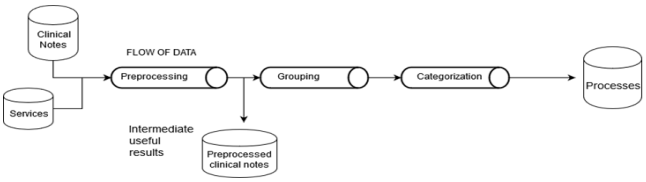

Figure 1. Pipeline to assigning processes to clinical notes.

As observed in Figure 1, inferring processes from raw clinical notes is performed in a 3-step process:

1. Pre-processing of data: The clinical notes contents are cleaned according to the category (type) of the clinical note. Hospitalization, discharge dates and service names are extracted from text (by looking for patterns in reports that are already annotated).
2. Grouping of semantically related clinical notes into processes: This task looks for group of notes with consecutive creation dates and use the hospitalization and discharge dates extracted from the text to find groups of notes created in the same day and whose group size includes at least two notes. After this, it is possible to assign distinct process id for each note.
3. Categorization of processes: Each process is assigned to one of the nine categories by using a set of different features describing the contents of the process (i.e., the clinical notes clustered together). The main features used in the process are number of clinical notes in the cluster, types of reports, number of clinical notes per category of the note and presence of hospitalization or emergency process prior to the currently analysed process.

III. RESULTS

The extraction and analysis of our clinical notes allows us to categorize them on groups based on the aforementioned processes, allowing us to obtain how the patients are distributed among the different hospital services based on their clinical features. Figure 2 shows the distribution of processes being the most frequent the consultations and following one the day hospital to come processes. Furthermore, Figure 3 reports on the distribution depending on the stage of the tumour of the patient. One can see that stages III and IV corresponding to advanced cancer stages, are the ones that have more hospital processes associated. Figure 4 depicts a process length; it is interesting to note that

the longest processes are those hospitalization processes that end with the cure of the patient.

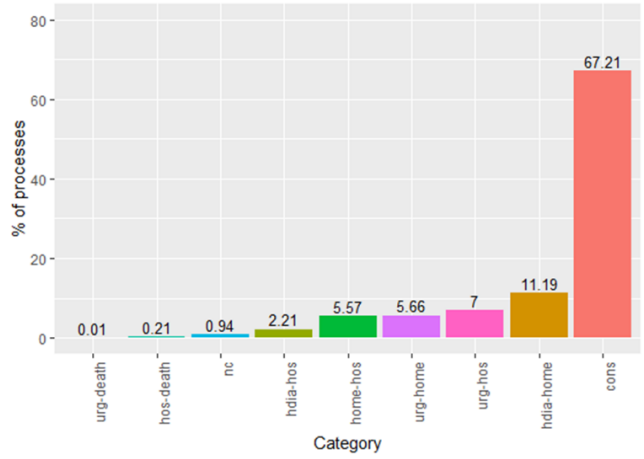

Figure 2. Distribution of processes

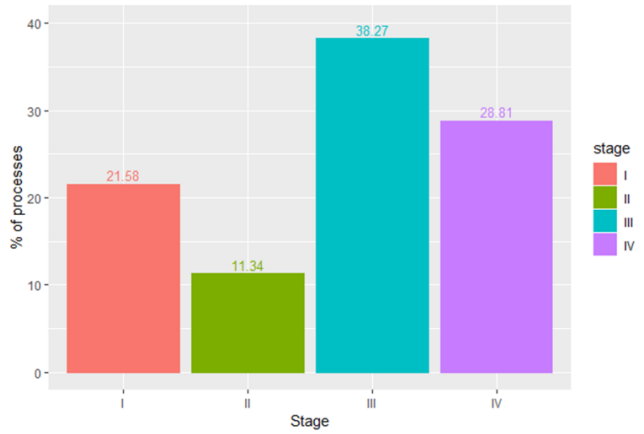

Figure 3. Processes per tumor stage

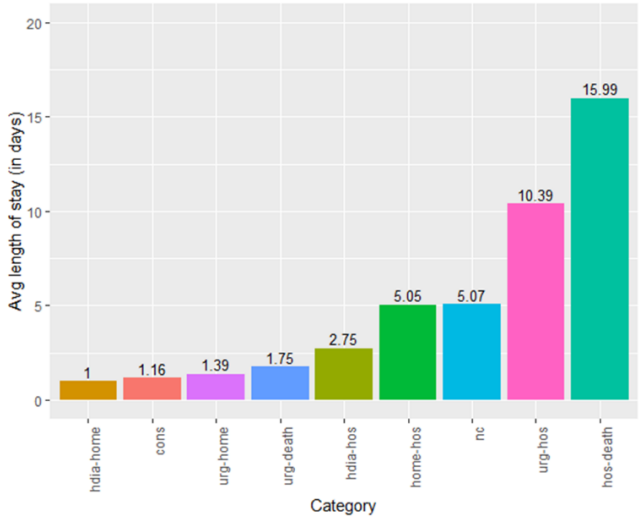

Figure 4. Length of stay per process

Regarding the identification of people at risk of developing lung cancer; we aim to identify the patients that might have lung cancer but have not been diagnosed yet. Having the diagnosis date available, the goal is to identify if there is evidence before the date of diagnosis that can lead the physicians to think that a specific patient might already have lung cancer, being thus necessary an oncology consultation as soon as possible. The current data obtained represents the distribution of the services that have been consulted by the patients before diagnosis date. A total of 4,938 reports were used for this analysis, as they were the only ones that had both service and cause of admission successfully extracted. The services with the largest number of reports are: Cardiology, Emergency Room, external consultations, Nursing, and Pneumology. We also have calculated percentage of patients and time duration (in months) between the last service visited by the patient and the date diagnosis. These results are presented in Figures 5 and 6, where we can see that certain services such cardiology and emergencies are visited more often. We can also observe that most of the patients had their last visit to the hospital before diagnosis in the last month.

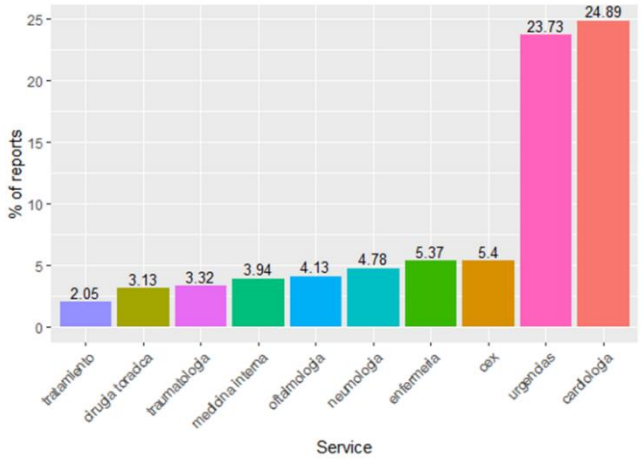

Figure 5. (a) Services visited prior to lung cancer diagnosis

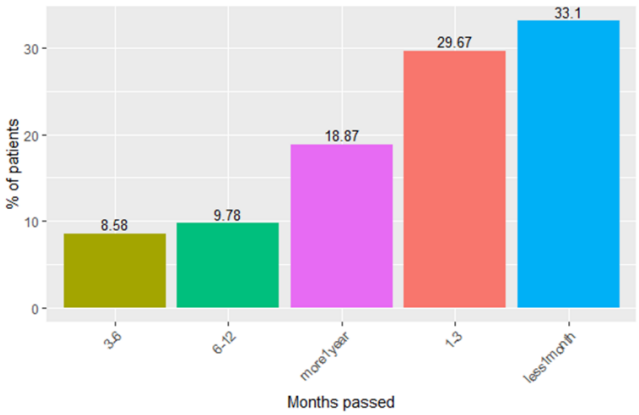

Figure 6. Months since the last visit to hospital prior to diagnosis.

IV. DISCUSSION AND FUTURE WORK

The paper has presented preliminary results regarding the analysis of clinical notes to extract knowledge that can help establishing policies for a better care of patients and to early diagnose lung cancer. The data-driven approach presented enables the extraction from text (clinical notes and reports (summarize report information)) information. This information can be used first to cluster the clinical notes into oncological processes, and then, to calculate relevant information regarding processes such as length of stay, admission cause, most frequent processes and analyze processes prior to the death of the patient. All these parameters are of paramount importance to establishing polices for improving patient´s care. In particular, we have focused on the analysis of the services the patients have visited prior to the tumor diagnosis in order to be able to find if there is any indicator that can help doctors to anticipate the tumor diagnosis at an early stage. Despite the promising results, this work only presents preliminary results, and further work is required to find among others, the relation of the objectives analyzed with demographics, habits, and comorbidities. Further results are also expected from the association of this information with scientific literature information.

ACKNOWLEDGMENT

This work is supported by the EU Horizon 2020 innovation programme under grant agreement No. 780495, project BigMedilytics (Big Data for Medical Analytics). The work is also a result of a 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).

REFERENCES

[1] R. Luengo-Fernandez, J. Leal, A. Gray, and R. Sullivan, “Economic burden of cancer across the European Union: a population-based cost analysis,” *Lancet Oncol.*, vol. 14, no. 12, pp. 1165–1174, Nov. 2013, doi: 10.1016/S1470-2045(13)70442-X.

[2] H. Baek, M. Cho, S. Kim, H. Hwang, M. Song, and S. Yoo, “Analysis of length of hospital stay using electronic health records: A statistical and data mining approach,” *PLOS ONE*, vol. 13, no. 4, p. e0195901, Apr. 2018, doi: 10.1371/journal.pone.0195901.

[3] R. Houthooft *et al.*, “Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores,” *Artif. Intell. Med.*, vol. 63, no. 3, pp. 191–207, Mar. 2015, doi: 10.1016/j.artmed.2014.12.009.

[4] J. Zuckerman *et al.*, “Psoas Muscle Area and Length of Stay in Older Adults Undergoing Cardiac Operations,” *Ann. Thorac. Surg.*, vol. 103, no. 5, pp. 1498–1504, May 2017, doi: 10.1016/j.athoracsur.2016.09.005.

[5] A. Almashrafi, H. Alsabti, M. Mukaddirov, B. Balan, and P. Aylin, “Factors associated with prolonged length of stay following cardiac surgery in a major referral hospital in Oman: a retrospective observational study,” *BMJ Open*, vol. 6, no. 6, p. e010764, Jun. 2016, doi: 10.1136/bmjopen-2015-010764.

[6] E. Menasalvas Ruiz *et al.*, “Profiling Lung Cancer Patients Using Electronic Health Records,” *J. Med. Syst.*, vol. 42, no. 7, p. 126, May 2018, doi: 10.1007/s10916-018-0975-9.