DOI: 10.1561/9781638282372.ch15

Chapter 15

Monitoring and Decision Support in Treatment Modalities for Lung Cancer

By Alejandro Rodriguez Gonzalez, Ernestina Menasalvas, Fotis Aisopos, Dimitrios Vogiatzis, Anastasia Krithara, Georgios Paliouras, Samaneh Jozashoori, Ariam Rivas, Ahmad Sakor, Maria-Esther Vidal, Maria Torrente, Mariano Provencio Pulla, Anna Trinatafyllou and Athanasios Dalianis

15.1 Introduction

In the European Union, the total cost of cancer was €126 billion in 2009, with health care amounting to €51.0 billion (40%). The health care cost reached €102 per citizen. Because of early death, productivity loss costs €42.6 billion, while there were €9.43 billion lost working days. Lung Cancer (LC), in particular, had the highest economic impact (€18.8 billion or 15% of overall cancer costs), followed by breast cancer (€15.0 billion, 12%), colorectal cancer (€13.1 billion, 10%), and prostate cancer (€8.43 billion, 7%) [1]. The current clinical approach to LC is standardized with reference ratios, regardless of the patient. However, the sub-optimal care and management of cancer patients affect the well-being of patients, as well as the healthcare cost [2–4].

In particular, the limitations of the current treatments can be detected at: (1) diagnosis — inter-patient variability preventing personalized treatments [5–7]; (2) therapy and response — determining the optimal duration of chemotherapy,

[8, 9] and more effective therapies are sought with fewer toxic effects; (3) adverse effects — over 5% of hospital admissions are due to the adverse effects of drugs [10]; and (4) comorbidities and side effects. The presence of comorbidities complicates the decisions about treatments. They are often underrepresented in clinical trials, and information regarding treatment effectiveness is often extrapolated from studies of younger patients without comorbidities. Thus, patient management is often sub-optimal and not personalized, affecting survival [11].

We designed and implemented an LC study application based on big data technologies to address the above concerns. First, the study harvests heterogeneous data from open sources and the Electronic Health Records (EHR) of LC patients from the Hospital Universitario Puerta de Hierro Majadahonda (HUPHM). The harvested data are analyzed, and a structure is extracted. Then the data are annotated with concepts from the Unified Medical Language System (UMLS). Next, all the data are integrated into a knowledge graph with semantic web technologies. The knowledge graph contains clinical data, and open data about LC, and can be accessed by oncologists via a web-based dashboard. The oncologists can focus on specific patient cohorts and obtain information about survival curves, toxicities, and drug–drug interactions. The knowledge graph also integrates structured representations of scientific publications; they can be traversed and ranked according to relevance to an input request.

This chapter presents the main outcomes achieved in developing the LC study in the context of BigMedilytics. It is organized as follows: in Section 15.2, we refer to the data analysis requirements for LC and how they are addressed by the study, i.e., the kind of information that can be obtained by an oncologist via a dashboard. In Section 15.3, we refer to data harvesting and analysis from various resources. Next, in Section 15.5, we refer to the data integration process that creates the knowledge graph. This is followed by a description of the software framework in Section 15.6. The results are presented in Section 15.7, and conclusions are drawn in Section 15.8.

15.2 Requirements for the Lung-Cancer Study

High-level requirements for the LC study as posed by the oncologists of HUPHM are concerned with the investigation of the following pieces of information:

- Over-treatment and non-scheduled visits
- Number of visits to the Emergency Room (ER)
- Time to spend searching for related cases in the bibliography
- Observed adverse events due to comorbidities

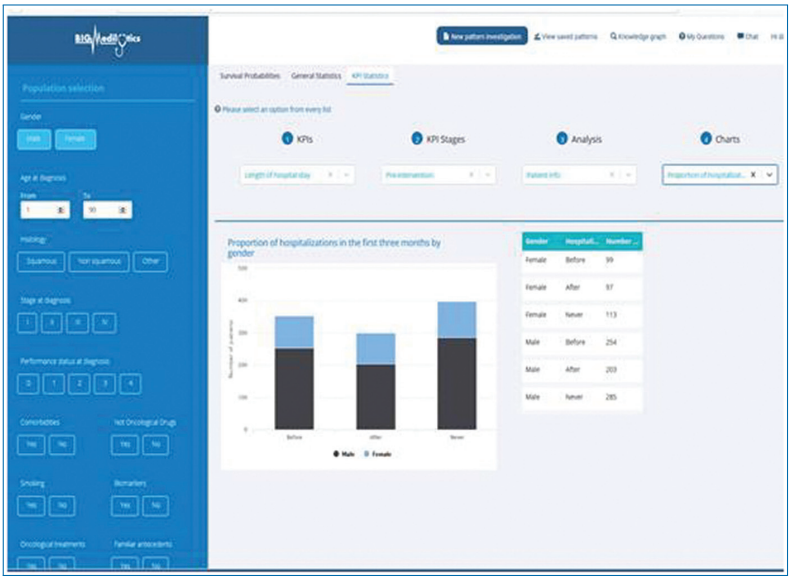

Figure 15.1. LC dashboard.

How does the study address the requirements: A web-based dashboard enables access to analytics of the clinical data and the LC knowledge graph. Figure 15.1 depicts an exemplary visualization of the clinical data outcomes. This dashboard allows clinicians to perform the following analysis: Patient data segmentation: First, the study can focus on segments of the population based on demographic and behavioral filters, e.g., gender, age, smoking habit, familial antecedents, as well as on the presence of biomarkers. Then there are filters related to the diagnosis of cancer, such as the histology (squamous, non-squamous, other), the stage of cancer at diagnosis (I–IV), performance status, comorbidities, the reception of non-oncological drugs, systemic and local progression, and finally, the existence of brain metastasis.

Hospitalization statistics: Based on the above filters, the dashboard can produce not only survival curves but also information about the length of hospitalization, number of toxicities, time from diagnosis till the first hospitalization, and the 10 most frequent diagnoses before being diagnosed with LC.

The Knowledge Graph Exploration: The open data integrated into the knowledge graph can be explored. These data included scientific publications, drug–drug interactions, and side effects of treatment.

Dashboard question answering: This provides answers and contextual information to questions posed in free text. The question types are as follows:

- Yes/No, e.g., "Is TREM2 associated with Lung Cancer?"—Factoid, e.g., "What type of LC is Afatinib used for?"—List of questions, e.g., "List drugs interacting with Afatinib."

15.3 Data Sources

EHRs: The HUPHM provided information about 8,901 patients diagnosed and treated for LC from 2008 to 2018. After data cleaning to remove corrupted EHRs, or EHRs with many missing values, we ended up with 988 EHRs of patients. Out of the 988 patients, 416 had been hospitalized. The EHRs contained 315,891 notes and 16,550 reports representing clinical variables of LC patients and services consulted by patients before and after diagnosis; each EHR had 320 attributes. The EHR contained both structured and unstructured information in the form of free text in Spanish.

Data anonymization: The study uses confidential information about patients of the HUPHM. Only the oncologists and the IT department have access to it and must conform to the applicable laws and HUPHM policies. The patients included in the project were informed of the project's aims, signed the informed consent form, and could request further information at any time. The patients' data usage was limited to the current investigation.

The EHRs were anonymized by removing entities such as name and address and replacing them with an ID. Only clinical data (e.g., regarding follow-up, treatments, and toxicities) related to the disease were shared with the rest of the study partners. The database with the relation between the anonymized data and the actual patient is stored in a server, which is not connected to the Internet. The transfer of anonymized data between the study partners was done physically or exceptionally via a corporate email of password-protected databases.

The project follows the Organic Law 15/1999 on the Protection of Personal Data and Anonymization of the Data and the EU General Data Protection Regulation 2016/679 (GDPR), regarding lawful data processing. The investigator and the promoter must keep the collected data for at least 25 years after completion. Thus, patients' data will be kept by the HUPHM as the promoter for the patients' benefit and further scientific research.

Open data sources: The following open data sources were used:

- PubMed: Provides access to the MEDLINE database of references and abstracts of scholarly articles for the life sciences. This repository was harvested for LC-related publication abstracts and the MeSH topics and relevant metadata. The PubMed Central (PMC) was also used to provide full-text access to some of the articles found in PubMed.
- DrugBank: An open database of drugs and targets.
- OBO Foundry: A repository of a wide range of interoperable biomedical and chemical ontologies. It is used for hierarchical harvesting relations for genes and diseases from the Gene Ontology and Disease Ontology, respectively.

Open data harvesting: Open data included 163,000 articles, 1.5M drug–drug interactions, 10K drug–target interactions, and OBO LC ontologies.

15.4 Analysis of Harvested Data, Natural Language Processing

NLP on Clinical Data: Natural Language Processing (NLP) is applied to the EHR text in Spanish [12]. The NLP pipeline is also depicted in Figure 15.2:

- Annotation: Rule-based annotators are deployed to extract: LC diagnosis using UMLS Metathesaurus, [13] the cancer stage using the TNM notation3, dates and times expressions, and family members mentions.
- Disambiguation: In the previous step, there can be generated ambiguous annotations due to the presence of negation, speculation, and annotations that do not belong to the patient as a subject. This step filters annotations affected by negation, speculation, and annotations that mention family history but do not refer to the patient. The disambiguation process automatically generates a new data set containing annotations without negation, speculations, or family history.
- Diagnosis Extraction: The cancer diagnosis and the diagnosis date are extracted from disambiguated annotations obtained in the previous step.

NLP on Open Data 163,000 harvested scientific publications (from PubMed and PMC) has been processed with NLP using named entity recognition and relation extraction to create an open graph of 402,020 nodes and 12,256,983 edges (see also Figure 15.3). SemRep was used to extract UMLS-based biomedical information [14]. The output of SemRep is semantic triples of the form subject–predicate object, where the subject and object entities are concepts from the UMLS, and the predicate is the relation between them. SemRep uses MetaMap, an entity extraction tool [15]. MetaMap uses symbolic NLP and computational linguistic techniques to

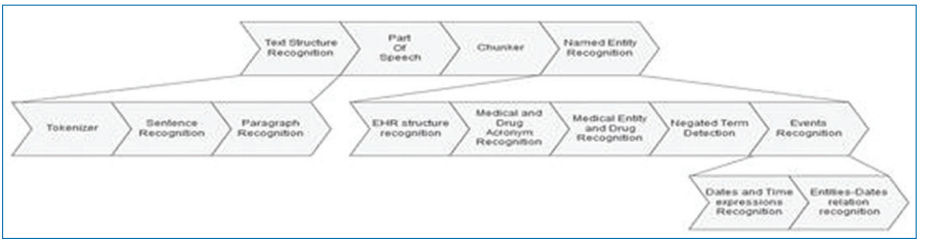

Figure 15.2. Information extraction from EHR.

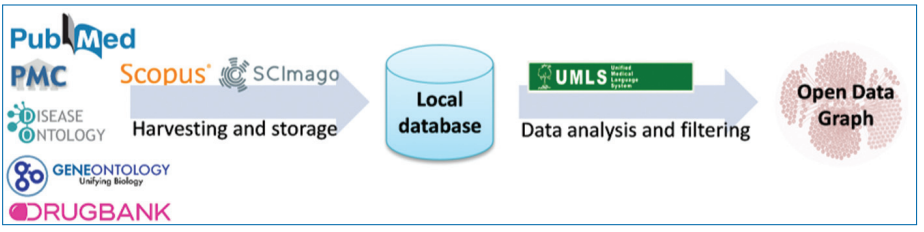

Figure 15.3. Open data harvesting and analysis.

map biomedical text to Metathesaurus concepts. An example of an extracted triplet is: (subject: radiation dose, relation: treats, object: body_tissue_injury).

15.5 The Lung Cancer Knowledge Graph

Knowledge integration is used to create and populate the knowledge graph with data provided by the Open Data graph, the analysis of the EHR, and other data sources. A unified schema allows for describing the integrated data into a knowledge graph. Additionally, NLP techniques are used for extracting relevant knowledge from the short text in available data sets, e.g., indications of contracts are extracted from the drug description in the Drugbank data set. Furthermore, linking techniques enable linking the knowledge graph with existing biomedical, e.g., Bio2RDF,[i] and general domain knowledge graphs, e.g., DBpedia[ii] and Wikidata.[iii] The knowledge graph is accessible via SPARQL endpoints or a federated query engine. Figure 15.4 depicts a portion of the knowledge graph. Pattern discovery is performed on top of the knowledge graph to identify communities or constellations of patterns that are similar. Ontologies express contextual information, and novel similarity measures are defined to decide when two patients are identical in a given context. Community detection algorithms (e.g., semEP [16]) are used for partitioning the knowledge graph into communities that represent meaningful patterns. They are described in terms of contextual information encoded in the knowledge graph.

i. https://bio2rdf.org/.

ii. https://www.dbpedia.org/.

iii. https://www.wikidata.org/wiki/Wikidata:Main_Page.

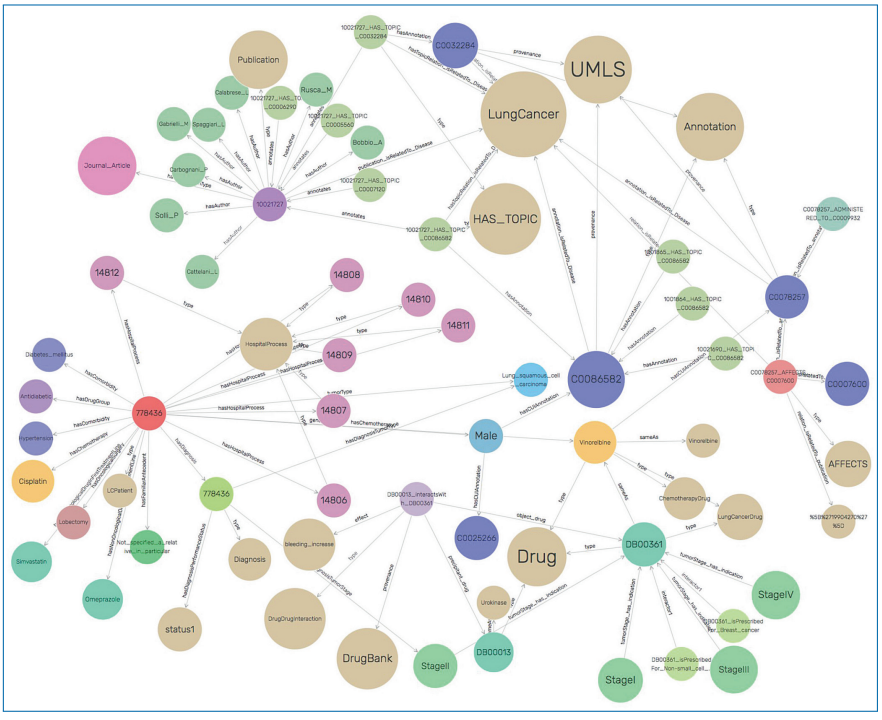

Figure 15.4. The LC knowledge graph.

15.6 Study Software Framework

The high-level software architecture of the LC study follows a layered model, where each layer provides certain functionality. The layers interact in a top-down way, with each layer having access to all layers below; they are as follows:

- Hardware Layer: Docker is used as a packaging and deployment mechanism for the study components.
- Resource Management Layer: Manages docker containers.
- Data Layer: Archive of all heterogeneous data sources.
- Extraction Layer: Information extraction from heterogeneous data sources.
- Platform Layer and Semantic Layer: Contains the big data tools used by the study's components. The Semantic Layer part of the layer semantically integrates all the heterogeneous data sources and the knowledge extracted at the Extraction layer. The result of the semantically integrated heterogeneous data sources is a knowledge graph.
- Analytics Layer: Pattern discovery on the knowledge graph.
- Presentation Layer: Dashboard and visualization modules.
- Support Layer: Offers utilities in multiple layers and components.

- Access Control Layer: Provides a centralized authentication and authorization service so that all components can communicate securely and reliably. Furthermore, some study components are remotely hosted, and they communicate with the rest of the platform over REST APIs.

15.7 The Lung Cancer Study Results

The oncologists, via the use of the dashboard, were able to identify some initial patterns that could have an immediate impact on the well-being of the patients. First, it was researched whether there is evidence before the diagnosis of LC that may lead physicians to clinical suspicion of LC. The risk of developing LC was associated with medical services used before diagnosis. The top-5 medical services used between 4 and 15 months before diagnosis were as follows: *cardiology, pneumology*, and *emergencies* (see Figure 15.5a).

Second, the patients who visit the ER and are discharged from the hospital were analyzed. The aim was to investigate whether this visit corresponds to a predictable, expected, and avoidable event. Whether it was necessary to conduct a proper initial evaluation. In particular, the number of ER admittances was associated with the EHR's features (e.g., age, gender, or comorbidities). For instance, in Figure 15.5b, we depict the number of ER services related to comorbidities.

Third, drug toxicities between non-oncological and oncological drugs and their association with long- and short-term survival were studied. For instance, in Figure 15.6, the survival curve combinations of oncological drugs (e.g., Vinorelbine, Pemetrexed, and Cisplatin) with a non-oncological drug (e.g., Omeprazole) are depicted.

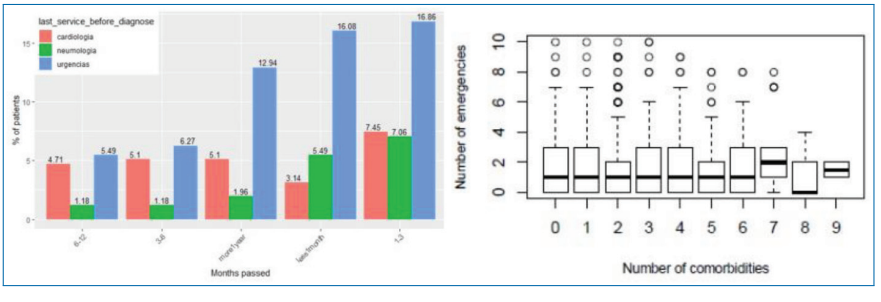

Figure 15.5. Analytical results. (a) Medical services used before diagnosis. (b) Emergency service usage per number with LC of comorbidities.

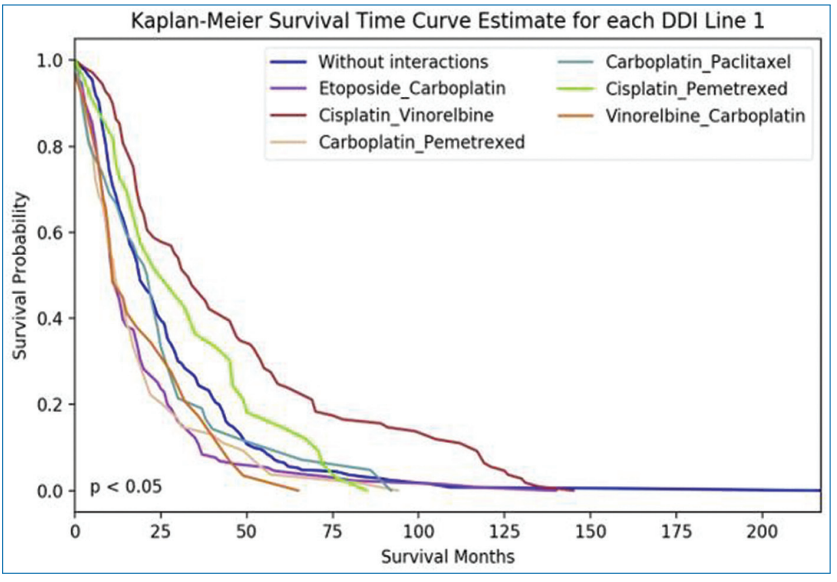

Figure 15.6. Drug interactions and toxicities.

15.8 Conclusion

The LC study has achieved some important results. First is integrating biomedical literature, structured databases (such as Drugbank), and EHR in a knowledge graph comprising 150M triples for the LC. Second, the study offers a web-based dashboard that allows access to the knowledge graph, including the usage of free text questions. Third, we were able to obtain certain associations that will ultimately benefit the patients. In particular, some evidence for the prediction of LC was extracted; then the pattern of the patient characteristics that visit the ER was analyzed; finally, drug interactions and toxicities for combinations of oncological and non-oncological drugs were associated with survival curves. This marks not the end of the investigation, as the information is being periodically updated with new open data, and further investigations are underway to allow oncologists to select the most appropriate treatment according to the patient's profile.

References

[1] Luengo-Fernandez R, Leal J, Gray A, Sullivan R. Economic burden of cancer-across the european union: a population-based cost analysis. The lancet oncology. 2013. 14(12), 1165–1174. DOI: 10.1016/S1470-2045(13)70442-X.

[2] Al-Badriyeh D, Alameri M, Al-Okka R. Cost-effectiveness research in cancer therapy: a systematic review of literature trends, methods and the influence of funding. BMJ open. 2017. 7(1), e012648. DOI: 10.1136/bmjopen-2016-012648.

[3] Gill J, Sullivan R, Taylor D. Overcoming cancer in the 21st century. UCL School of Pharmacy. 2015.

[4] Poonacha TK, Go RS. Level of scientific evidence underlying recommendationsarising from the national comprehensive cancer network clinical practice guidelines. Journal of clinical oncology. 2011. 29(2), 186–191. DOI: 10.1200/JCO.2010.31.6414.

[5] Detterbeck FC, Marom EM, Arenberg DA, Franklin WA, Nicholson AG, Travis WD, Girard N, Mazzone PJ, Donington JS, Tanoue LT, *et al.* The iaslc lung cancer staging project: background data and proposals for the application of tnm staging rules to lung cancer presenting as multiple nodules with ground glass or lepidic features or a pneumonic type of involvement in the forthcoming eighth edition of the tnm classification. Journal of Thoracic Oncology. 2016. 11(5), 666–680. DOI: 10.1016/j.jtho.2015.12.113.

[6] Group NMaC. Preoperative chemotherapy for non-small-cell lung cancer: asystematic review and meta-analysis of individual participant data. The Lancet. 2014. 383(9928), 1561–1571. DOI: 10.1016/S0140-6736(13)62159-5.

[7] Sculier JP, Berghmans T, Meert AP. Tnm classification and clinicopathological factors: what is helpful for adjuvant chemotherapy decision after lung cancer resection? Journal of thoracic oncology. 2014. 9(3), 266–270. DOI: 10.1097/JTO.0000000000000110.

[8] Carcereny E, Ramirez JL, Sanchez-Ronco M, Isla D, Cobo M, Moran T, de Aguirre I, Okamoto T, Wei J, Provencio M, *et al.* Blood-based chrna3 single nucleotide polymorphism and outcome in advanced non-small-cell lung cancer patients. Lung cancer. 2010. 68(3), 491–497. DOI: 10.1016/j.lungcan.2009.08.004.

[9] Provencio M, Isla D, Sánchez A, Cantos B. Inoperable stage iii non-small celllung cancer: Current treatment and role of vinorelbine. Journal of thoracic disease. 2011. 3(3), 197. DOI: 10.3978/j.issn.2072-1439.2011.01.02.

[10] Garrido P, González-Larriba JL, Insa A, Provencio M, Torres A, Isla D, Sanchez JM, Cardenal F, Domine M, Barcelo JR, *et al.* Long-term survival associated with complete resection after induction chemotherapy in stage iiia (n2) and iiib (t4n0-1) non-small-cell lung cancer patients: the spanish lung cancer group trial 9901. Journal of clinical oncology. 2007. 25(30), 4736–4742. DOI: 10.1200/JCO.2007.12.0014.

[11] Søgaard M, Thomsen RW, Bossen KS, Sørensen HT, Nørgaard M. The impact of comorbidity on cancer survival: a review. Clinical epidemiology. 2013. 5(Suppl 1), 3. DOI: 10.1002/ijc.10882.

[12] Solarte Pabón O, Torrente M, Provencio M, Rodríguez-Gonzalez A, Menasalvas E. Integrating speculation detection and deep learning to extract lung cancer diagnosis from clinical notes. Applied Sciences. 2021. 11(2), 865. DOI: 10.3390/app11020865.

[13] Bodenreider O. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research. 2004. 32(suppl_1), D267–D270. DOI: 10.1093/nar/gkh061.

[14] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguisticstructure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of biomedical informatics. 2003. 36(6), 462–477. DOI: 10.1016/j.jbi.2003.11.003.

[15] Aronson AR. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. Proceedings of the AMIA Symposium. 2001. p. 17. American Medical Informatics Association.

[16] Palma G, Vidal M, Raschid L. Drug-target interaction prediction using semantic similarity and edge partitioning. Proceedings of The Semantic Web – ISWC. 2014. https://doi.org/10.1007/978-3-319-11964-9_9.