*Caring is Sharing – Exploiting the Value in Data for Health and Innovation*
*M. Hägglund et al. (Eds.)*

*doi:10.3233/SHTI230056*

# Application of Process Mining for Modelling Small Cell Lung Cancer Prognosis

Luca MARZANO[a,1], Sebastiaan MEIJER[a], Asaf DAN[b], Salomon TENDLER[b], Luigi DE PETRIS[b], Rolf LEWENSOHN[b], Jayanth RAGHOTHAMA[a] and Adam S. DARWICH[a]

[a] *Division of Health Informatics and Logistics, School of Engineering Sciences in Chemistry, Biotechnology and Health (CBH), KTH Royal Institute of Technology, Huddinge, Sweden.*

[b] *Department of Oncology-Pathology, Karolinska Institutet and the Thoracic Oncology Center, Karolinska University Hospital, Stockholm, Sweden*

**Abstract.** Process mining is a relatively new method that connects data science and process modelling. In the past years a series of applications with health care production data have been presented in process discovery, conformance check and system enhancement. In this paper we apply process mining on clinical oncological data with the purpose of studying survival outcomes and chemotherapy treatment decision in a real-world cohort of small cell lung cancer patients treated at Karolinska University Hospital (Stockholm, Sweden). The results highlighted the potential role of process mining in oncology to study prognosis and survival outcomes with longitudinal models directly extracted from clinical data derived from healthcare.

**Keywords.** Process mining, Real-world Data, oncology, small cell lung cancer, treatment decision

## 1. Introduction

Real-world data (RWD) from healthcare has the potential to inform real-world evidence of treatment effects. This includes time-to-event survival analysis to predict outcomes of administered therapies, and prognosis [1]. However, several aspects need to be considered when using RWD to inform decision-making, to avoid biases and extract robust results [2]. An important aspect is to consider the different treatment processes that the dataset represents [3]. In recent years, there has been an upswing in process mining applications within the healthcare domain [4]. Process mining is an approach that bridges process modelling and computational sciences, with the goal of extracting and analysing processes from the data. Process mining is usually carried out in several steps, encompassing process discovery, conformance check, and system enhancement [5]. This approach aims to improve the interpretation of outcomes generated from the data records

[1] Corresponding Author: Luca Marzano, Division of Health Informatics and Logistics, School of Engineering Sciences in Chemistry, Biotechnology and Health (CBH), KTH Royal Institute of Technology, Huddinge, Sweden, address: Hälsovägen 11, Huddinge, E-mail: lmarzano@kth.se .

produced during real-world processes [4]. Process mining in healthcare has mainly focused on production data, such as hospital clinical pathways, resource allocation, and scheduling [6]. Very little is currently known about process mining applied with the purpose of studying disease progression and survival outcomes. Oncology as a whole field, and small cell lung cancer (SCLC) in particular, is one of the diseases that would benefit most from this. Prognosis of oncological patients is mainly assessed with time-to-event analysis, such as Cox regression [7], with little consideration of treatment decision points over the course of the disease.

This paper showcases a process mining analysis of oncological survival data. The approach describes the design of a pipeline aimed to directly extract the processes from the clinical real-world database with the purpose of modelling longitudinally the prognosis and survival of the SCLC patients.

## 2. Methods

The data consisted of consecutive SCLC cases diagnosed and treated at the Karolinska University Hospital between 2008 and 2016 (n=705). The study was approved by the institutional review boards at Karolinska Institutet and at Stockholm County Council (2016/8-31). The present cohort was previously used to validate the eighth version of TNM (Tumour, Nodes, Metastases) staging system [8], study the prognostic impact of baseline patient characteristics [9], and the identification of subgroups of patients with similar survival times using unsupervised machine learning [3].

The type of treatment received by the patients in the cohort was chemotherapy (CT), chemotherapy and concomitant radiotherapy (CT+RT), radiotherapy (RT), and surgery (SR). No one received immunotherapy. Treatment choice is usually based on clinical variables, including patient's general condition (ECOG performance status), tumor extension (cancer stage) and response to the previous therapy as well as residual toxicity. According to the documented clinical or radiological progression of the tumor, the clinician may choose to change or reuse the same therapy. A rechallenge is considered as a new treatment line [9]. The progress free survival (PFS) for each line of therapy is defined as the interval between the start of the therapy and the earliest date of documented clinical or radiological progression according to standard clinical practice, or death.

Figure 1 shows the process mining workflow applied in this study. The main outcome of the pipeline is the longitudinal survival model extracted from the patients' records. Prior to extracting the processes, patient survival data were converted into an event-log format retrieving the PFS timestamps and the treatment decision follow up. After formatting, process mining was applied to extract the process map of the chemotherapy cycles with directly-follows graph technique [10]. This map could be defined as a directed graph, where nodes represent the selected therapy, and edges the follow up decision. Self-loops represented the re-challenge with the same therapy. Once the graph was extracted, the process behind PFS and treatment decision was studied. Following the querying of graphs related to patient groups of interest (e.g., patients with a specific tumor stage), process visualisation with metrics of interest were produced (edge frequency and node median PFS), sequence of the therapies, and transition matrices were provided, thus obtaining the longitudinal survival model developed directly from the patients' records.

All the analysis were carried out in R using the dplyr package for data processing, and bupaR package for process mining.

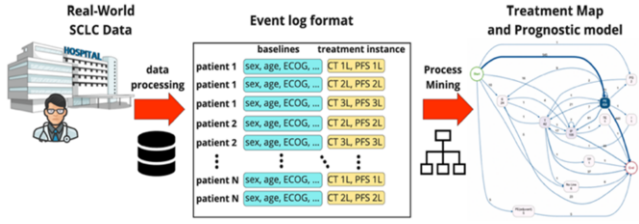


**Figure 1.** SCLC process mining workflow.

## 3. Results

A treatment decision event log with 1,622 instances was created from the patient data. The treatment map was extracted, and a dynamic process visualisation of the patients was produced detailing patient TNM staging. The developed pipeline in Figure 1 allowed to explore several cohorts by querying the subgraphs from the treatment map filtering the patients' characteristics of interest (such as cancer stage, and ECOG performance status). As example, Figure 2 shows the results for one of the cohorts: patient with IVB TNM stage (n=311). The 8th TNM classification introduced subclasses of stage IV SCLC patients, with most patients having multiple distant metastasis. The majority of SCLC cases are diagnosed with stage IVB and therefore the study of this case was of high clinical interest. From the longitudinal model we extracted the treatment-PFS graph, chemotherapy sequences, and treatment transition frequencies. The pipeline was tested also to extract longitudinal subgroup models for other TNM stages, for different treatment decisions (e.g., CT and CT+RT), and detected clusters in a previous study [3], thus showing the flexibility of the approach to different purposes.

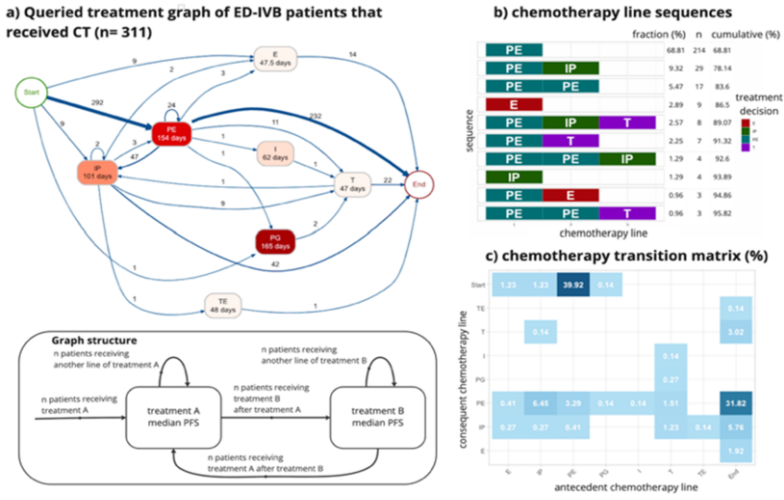


**Figure 2.** treatment process map, and chemotherapy (CT) line traces of ED-IVB patients. PE: platinum with etoposide, IP: platinum with irinotecan, T: topotecan, E: etoposide, I: irinotecan, TE: topetecan and etoposide, PG: platinum and gemcitabine

## 4. Discussion

The study presents an analysis of treatment pathways and associated outcomes in subgroups of SCLC patients treated with chemotherapeutic agents.

The chosen methodology was a pipeline that directly extracted treatment processes from the patient data, thus defining graphs from treatment and prognostic outcomes.

Results showed that process mining has potential for enriching the analysis of oncological cohorts by developing an increased understanding of the underlying treatment pathways and decision points. In addition to the rapid processing and the impactful visualisation, the technique can be used to inform longitudinal modelling of disease progression and subsequent impact on treatment decisions, as well as patient outcomes. The ready-to-use longitudinal models directly extracted from RWD constitute suitable objects for the design of multi-state survival models, or other process model solutions, such as causal networks [11].

To the best knowledge of the authors, this work constitutes the first application of process mining in healthcare that focuses on studying longitudinal survival outcomes in SCLC. In addition to process discovery, conformance check, and system enhancement, process mining could find reliable applications also for prognostic assessment and treatment evaluation.

This study has some limitations. For instance, the granularity of the data did not allow sufficient confidence to make inferences about the treatment effects. This is mainly due to the lack of longitudinal information in the retrospective cohort. On one hand, highly detailed information regarding the baseline patient and disease characteristics allowed the identification of patient groups of interest. On the other hand, follow-up information was limited to overall survival, PFS, and chemotherapy regimen. Further, most of the patients had extensive disease, TNM stage IVB, and short follow-up due to the early relapse during the first line therapy. Additional information on dose levels, adverse effects, and cancer progression would further improve the analysis. Increasing the sample size of the cohort, and the extension of the study involving multiple centres would benefit to the achievement of reliable evidence of SCLC studies.

Future work includes the collection of longitudinal information. Alternatives to the graph definition will be explored (e.g., adverse effects nodes, definition of edge weights, or new surrogate endpoints to study with the PFS). For what concerns the process mining, a larger variety of algorithms [10] will be tested to assess the variation and validity of the models. Involvement of the clinical experts in the model development will be a key factor to assess the clinical reliability of the approach, especially to leverage the gap between real processes and quality of the real-world data. Further considerations regarding the implementation challenges will be explored and discussed. Adherence with the current community standards for clinical pathways analytics (e.g., Observational Health and Data Science and Informatics [12]) and data interoperability in a large-scale infrastructure, such as the European Health Data Space [12], would be a key characteristic to implement.

## 5. Conclusions

Process mining applied to real-world healthcare data has the potential to allow visualisation of follow-up decisions in function of disease progression. The method can also be used to inform the design of multi-state models extracted directly from the data,

enabling more sophisticated statistical modelling of longitudinal data. The study provides insights into the role of process mining in oncology to study prognosis and survival outcomes, and additional indications on how to develop this further in the future.

## References

[1] Schurman B. The Framework for FDA's Real-World Evidence Program. Appl Clin Trials 2019;28:15–7.

[2] Miksad RA, Abernethy AP. Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality. Clin Pharmacol Ther 2018;103:202–5. https://doi.org/10.1002/CPT.946.

[3] Marzano L, Darwich AS, Tendler S, Dan A, Lewensohn R, de Petris L, et al. A novel analytical framework for risk stratification of real-world data using machine learning: A small cell lung cancer study. Clin Transl Sci 2022;15:2437–47. https://doi.org/10.1111/cts.13371.

[4] Munoz-Gama J, Martin N, Fernandez-Llatas C, Johnson OA, Sepúlveda M, Helm E, et al. Process mining for healthcare: Characteristics and challenges. J Biomed Inform 2022;127:103994. https://doi.org/10.1016/j.jbi.2022.103994.

[5] Martin N, de Weerdt J, Fernández-Llatas C, Gal A, Gatta R, Ibáñez G, et al. Recommendations for enhancing the usability and understandability of process mining in healthcare. Artif Intell Med 2020;109:101962. https://doi.org/10.1016/j.artmed.2020.101962.

[6] Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D. Process mining in healthcare: A literature review. J Biomed Inform 2016;61:224–36. https://doi.org/10.1016/J.JBI.2016.04.007.

[7] de Neve J, Gerds TA. On the interpretation of the hazard ratio in Cox regression. Biometrical Journal 2020;62:742–50. https://doi.org/10.1002/bimj.201800255.

[8] Tendler S, Grozman V, Lewensohn R, Tsakonas G, Viktorsson K, De Petris L. Validation of the 8th TNM classification for small-cell lung cancer in a retrospective material from Sweden. Lung Cancer 2018;120:75–81. https://doi.org/10.1016/j.lungcan.2018.03.026.

[9] Tendler S, Zhan Y, Pettersson A, Lewensohn R, Viktorsson K, Fang F, et al. Treatment patterns and survival outcomes for small-cell lung cancer patients–a Swedish single center cohort study. Acta Oncol (Madr) 2020;59:388–94. https://doi.org/10.1080/0284186X.2019.1711165.

[10] van der Aalst W. Advanced Process Discovery Techniques. Process Mining 2016:195–240. https://doi.org/10.1007/978-3-662-49851-4_7.

[11] Krishnan SM, Friberg LE, Bruno R, Beyer U, Jin JY, Karlsson MO. Multistate model for pharmacometric analyses of overall survival in HER2-negative breast cancer patients treated with docetaxel. CPT Pharmacometrics Syst Pharmacol 2021;10:1255–66. https://doi.org/10.1002/PSP4.12693.

[12] European Commission D-G for H and FS. European Health Data Space 2022. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en#more-information (accessed February 1, 2023).