A Process Mining approach to statistical analysis: application to a real-world advanced melanoma dataset

Erica Tavazzi[1,2], Camille L. Gerard[2], Olivier Michielin[2,3], Alexandre Wicky[2], Roberto Gatta[2], and Michel A. Cuendet[2,3,4]

[1] Department of Information Engineering, University of Padova, Italy
erica.tavazzi@phd.unipd.it
[2] Precision Oncology Center, Lausanne University Hospital (CHUV), Switzerland
[3] Molecular Modeling Group, Swiss Institute of Bioinformatics, Lausanne, Switzerland
[4] Department of Physiology and Biophysics, Weill Cornell Medicine, New York, USA

Abstract. Thanks to its ability to offer a time-oriented perspective on the clinical events that define the patient's path of care, Process Mining (PM) is assuming an emerging role in clinical data analytics. PM's ability to exploit time-series data and to build processes without any *a priori* knowledge suggests interesting synergies with the most common statistical analyses in healthcare, in particular survival analysis. In this work we demonstrate contributions of our process-oriented approach in analyzing a real-world retrospective dataset of patients treated for advanced melanoma at the Lausanne University Hospital. Addressing the clinical questions raised by our oncologists, we integrated PM in almost all the steps of a common statistical analysis. We show: (1) how PM can be leveraged to improve the quality of the data (data cleaning/preprocessing), (2) how PM can provide efficient data visualizations that support and/or suggest clinical hypotheses, also allowing to check the consistency between real and expected processes (descriptive statistics), and (3) how PM can assist in querying or re-expressing the data in terms of pre-defined reference workflows for testing survival differences among sub-cohorts (statistical inference). We exploit a rich set of PM tools for querying the event logs, inspecting the processes using statistical hypothesis testing, and performing conformance checking analyses to identify patterns in patient clinical paths and study the effects of different treatment sequences in our cohort.

Keywords: Process mining · Oncology · Melanoma · Statistical analysis

1 Introduction

Process Mining (PM) is a family of process analysis methods that aim at discovering, monitoring and improving the efficiency of real processes by extracting knowledge from the Event Logs (EL) recorded by an information system. Analytic algorithms are applied to ELs with the main goals of: (i) mining the data

in order to represent the process able to produce them (*Process Discovery*, PD), (ii) measuring to which extent a given process can represent an input EL or how much an EL complies with a given process (*Conformance Checking*, CC), and (iii) improving process efficiency, by allowing problem diagnosis and delay prediction, recommending process redesigns or supporting decision making (*Process Enhancement*) [2].

In PM for Healthcare (PM4HC), processes are meant as a graph of activities which can be performed with the aim of diagnosing, treating and/or preventing diseases to improve the patients' health status. The activities can be clinical and non-clinical and may represent different behaviours according to the specific organization [12]. Often, such processes are highly dynamic, complex, increasingly multidisciplinary [8]. Notably, the complexity increased recently due to the advent of personalized approaches to care, in which treatments are tailored to the specific profile of the patient and disease, such that the diversity of therapeutic pathways exploded compared to traditional standardized care guidelines.

Pragmatically, PM4HC has shown interesting applications in many domains, and in Oncology in particular, PM4HC was successfully applied to identify the most common patterns of care for many kinds of tumors, even though the purpose remained exploratory. Rectal cancer [7], gynecological cancer [11], and melanoma [13] were investigated both in terms of PD and CC, even if in most cases the focus was more on CC, while the application of PD remained descriptive of the general trend [9]. From this perspective, there were only few cases where the PM4HC analysis was used for statistical inference, *i.e.* to concretely develop predictive models assessing the role of covariates in determining disease evolution or patient clinical pathway. While the idea of applying a combination of PM and statistics for a complete statistical analysis is not entirely new [4][10], it is not a very common approach and still requires to be consolidated, in particular to integrate survival analysis, which plays a forefront role in Oncology.

In this work, we focus on exploring the contributions of PM when performing statistical analyses in Oncology. As an application, we examined a real-world cohort of advanced melanoma patients treated at the Lausanne University Hospital (CHUV); here we show how PM can guide and/or assist researchers in all the classical steps of statistical analysis, that is, data preprocessing, descriptive statistics, and inferential statistics. Figure 1 summarizes these steps.

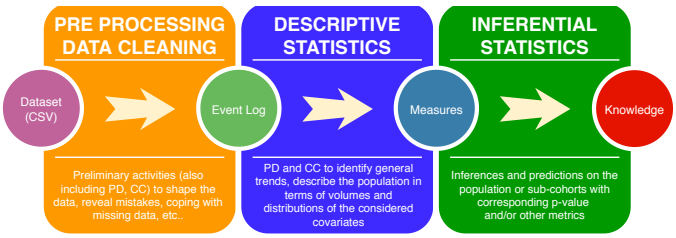

Fig. 1. Workflow of the classical steps of a statistical analysis, here implemented exploiting a process-oriented approach.

In the preprocessing step, we approached the data inspecting their structure, their information content, and their quality: after identifying the clinical milestones of interest (like diagnosis, treatments, survival outcome), data were first shaped as EL. We then employed the visualization tools provided by PM to detect data inconsistencies due to input errors or missing values. This allowed us to go back to the data sources, recheck and correct the recorded information, thus recursively improving the data quality.

In the descriptive analysis step, we first employed the EL time-oriented structure to inspect cardinality and order of the administered pharmacological treatments. Then, we implemented both unsupervised and supervised methods to capture the flow of the patients' pathways over data-driven graphs (PD approach) or user-defined graphs (CC approach), respectively. In this part of the analysis, the graphical output provided by PM allows a fast access to the design and/or interpretation of the models, and an immediate assessment of the treatments in terms of type, order and timing of consecutive administrations.

Finally, in the inferential statistics step, we build upon the processes constructed in the previous step to quickly select sub-cohorts of patients characterized by similar patterns of care and/or clinical attributes. The cohorts were then compared in terms of time-to-event outcome and overall survival (OS), using Kaplan-Meier analysis and log-rank test.

2 Material and Methods

2.1 Material

In this work, we analyzed the data of a cohort of patients treated at the CHUV and diagnosed with advanced melanoma.

Melanoma is an aggressive cancer that arises from melanocytes (pigment cells). Cutaneous melanoma is the most common type. However, it exists also uveal and mucosal melanomas, which occur in the eye and in the mucosa (such as the mouth or the vulva), respectively. The primary risk factor of cutaneous melanoma is ultraviolet light exposure. As outdoor activities are a way of life in Switzerland, the melanoma incidence is high in the country [3]. The extent of the disease progression is described by a staging system, ranging from I to IV: Stage IV indicates metastatization of melanoma cells to distant organs. Surgery is the most common and resolutive approach for the lowest stages, but when the disease is more extensive, systemic treatments such as Immunotherapy are required, with Radiotherapy also used as palliative or local treatment.

The study cohort includes 184 patients diagnosed with advanced melanoma between March 18th, 2008 and November 17th, 2019, with follow-up up to 2019, December 30th.[1] Data were sourced from the electronic healthcare records available at CHUV and curated by trained oncologists.

[1] This study was approved by the Research Ethical Committee of Canton de Vaud (CER-VD) and includes only patients who did not oppose usage of their data, and was conducted according to the Swiss Federal Act on Research involving Human Beings.

Data includes: sex, date of birth, primary tumor type, stage and diagnosis date, advanced tumor diagnosis date and mutation type (among BRAF-V600, BRAF-nonV600, NRAS, wild type (wt)), pharmacological treatments, and survival information (date of death or last follow-up). In this study, only the medications administered after the stage IV diagnosis were considered.

2.2 Methods

We implemented the classical statistical analysis pipeline shown in Figure 1 by employing PM4HC techniques to achieve the goals of each step. To perform the analyses, we used pMineR, an open source R library implementing PM4HC functionalities [5]. By handling data in the form of EL, it allows, among its features, to implement PD and CC analyses.

We started with the raw data set, which we first assumed to be *clean* from mistakes. First, we cast the data in the form of EL, by selecting the main clinical milestones of interest for the analysis and defining the rules to cope with missing values. Then, we implemented a PD algorithm based on First Order Markov Models (FOMMs)[5], to provide a fast and easy-to-understand representation of the subsequent events. This representation allowed us to identify visually some unexpected links between clinical events (*e.g.* due to mistakes in some dates). With the help of a physician, we iteratively reviewed the data and rerun the PD algorithm in order to increasingly approach the expected graph and thus refine the data quality.

To describe the general statistics of the population and quantify the flux of patients though different patterns of cares (the second step in Figure 1), we exploited both PD and CC techniques. The unsupervised PD analysis is based on the same FOMM model as described above. The supervised CC approach is based on a pre-defined representation of the different treatment lines implemented with the Pseudo-Workflow formalism (PWF) available in the software tool. Once performed PD and CC, the patients were grouped according to their paths through the graphs using the selection language provided by the tool. Then Kaplan-Meier survival curves and log-rank tests were used to quantify statistical differences between the groups, considering as end-points time-to-event in PD and OS in CC.

Process Discovery In PD, one of the most diffused process representation exploits the directly-follows graphs (DFGs): in this graphical representation, directed edges link all the couples of nodes representing subsequent activities in the EL. Even if DFGs have some well-known limitations [1], they are very intuitive and can be helpful to share with clinicians a first representation of the data. In the pMineR implementation, DFGs correspond to FOMMs.

Conformance Checking CC was performed by using the PWF, designing a diagram that describes the expected flow of events in terms of diagnoses, treatment lines, and survival events. Graphically, this results in a set of nodes,

representing the *status* that the subjects can assume, and a set of conditions (*triggers*) which fire transitions between status [6]. This representation allows to count which triggers/status are activated while automatically running down the events of each subjects, thus capturing the population behaviours through the diagram.

3 Results

3.1 Data preprocessing

Event Log For each patient, we built the EL with the following events, each associated with a time stamp:

- *Primary Stage*: the primary diagnosis, with melanoma type, tumor stage at the diagnosis, and somatic mutation harboured by the tumor as attributes;
- *Stage IV*: the diagnosis of stage IV;
- *T-Begin*: the begin of a line of treatment, with the type of the given drug(s) as attribute;
- *T-End*: the end of a line of treatment, with the type of the given drug(s) as attribute;
- *Dead, Censored*: the survival information, consisting in the dead of the patient or in the last follow-up date, respectively.

The collected treatments belong to the following categories:

- *Immunotherapy (IO)*: anti-CTLA4, anti-PD1, anti-CTLA4 + anti-PD1 (in combination), or other IO;
- *Chemotherapy (Chemo)*;
- *Targeted therapy*: tyrosine kinase inhibitors (TKI), other targeted therapy (TT).

In this study, only the treatments after stage IV diagnosis were considered.

Missing data In time-oriented analyses, missing information can consist either in unrecorded events or in missing dates associated to the events themselves. In order to preserve the clinical information we kept only complete treatments lines: the EL of patients with an incomplete line were thus truncated to the last available certain information (stage IV diagnosis or end of a previous line), artificially introducing a *Censored* event before the line with missing information.

Data Cleaning To detect mistakes in the data, we adopted an iterative approach: a FOMM process was discovered and visually analyzed to detect inconsistencies on unexpected edges. Then, the data were updated and the the procedure repeated until no more mistakes were found.

To give a practical example of detection, we report in Figure 2 a) the FOMM resulting from an intermediate version of the dataset, where unexpected edges emerge because the beginning of the first line of treatment was erroneously dated

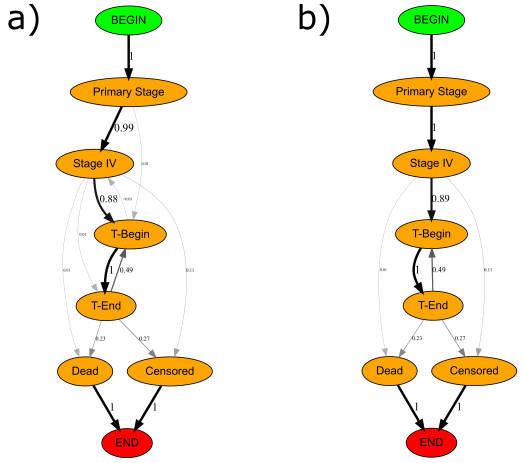

Fig. 2. First Order Markov Models obtained on all the events constituting the EL: a) before cleaning the information of a subject with an error in the dates, b) after data cleaning.

before the stage IV diagnosis for one patient in the source data. In Figure 2b) we can observe the FOMM after correction of the inaccurately collected information. This updated graph presents, conversely, only relations fully compliant with the nature (and the collection design) of the data.

With this approach we revealed some previously uncaught mistakes in the original data, such as inconsistency in data representation (*e.g.* dd/mm/yy vs dd/mm/yyyy), or temporal event inversion (*e.g.* cancer treatment begin before a tumor diagnosis).

3.2 Descriptive statistics

A first descriptive statistics was performed by querying the input EL, consisting of 1196 records: this allowed us to explore in the first instance cardinality and order of the administered treatments. Then, we delved into the data by using the FOMM, to obtain an agnostic data representation, and a PWF diagram, to verify the consistency of the process with respect to the expected behaviour.

Event Log querying By analysing the EL it was possible to perform some first descriptive investigations. We focused, specifically, on the treatments administered to the patients. Considering the events of all the patients, regardless of the position in the path of care, we extracted a total of 322 administered treatments. Table 1 reports, for each treatment category, its absolute and relative frequency of occurrence, and its duration in terms of median and inter-quartile range (25%-75%).

Out of 163 patients that received at least one recorded line of treatment, we identified 49 distinct patterns of treatment sequence. The most frequent ones are reported in Table 2.

Table 1. Occurrences and duration (in days) of the administered treatments collected in the data. The inter-quartile ranges (IQR) are computed at 25% and 75%.

Drug category	**Occurrences (%) (n=322)**		**Median (IQR) duration [days]**
TKI	76	(23.6)	122 (76.5–228.0)
anti-CTLA4 + anti-PD1	70	(21.7)	46.5 (0.0–167.8)
anti-PD1	66	(20.5)	84.0 (33.0–253.2)
anti-CTLA4	66	(20.5)	61.5 (31.0–63.0)
Chemo	29	(9.0)	44.0 (22.0–67.0)
Other IO	13	(4.0)	92.0 (22.0–203.0)
TT	2	(0.6)	461.5 (300.7–622.2)

Table 2. Most frequent patterns of treatment recorded in the data. The relative frequency of occurrence is computed over the total number of patients with at least one recorded treatment.

First line	**Second line**	**Occurrence (%) (n=163)**	
anti-CTLA4 + anti-PD1	-	36	(22.1)
anti-PD1	-	22	(13.5)
anti-CTLA4	-	11	(6.7)
anti-CTLA4 + anti-PD1	TKI	11	(6.7)
Chemo	anti-CTLA4	9	(5.5)
anti-CTLA4	anti-PD1	8	(4.9)
TKI	anti-CTLA4	6	(3.7)
anti-CTLA4	TKI	5	(3.1)
TKI	-	3	(1.8)

Process discovery on treatment sequences Figure 3 shows the FOMM obtained from the clean EL considering only the administered treatments (ignoring diagnosis and survival events). Such a process allows to inspect the temporal causality of the treatments, highlighting the most frequent connections over all the population. It also provides a first overview of the position of the treatments in the paths.

Conformance checking for treatment sequences We designed a PWF able to capture the chronological order of the events: at the top, we represented the events related to the staging, and then the different treatment lines. In order to be able to define treatments paths at different levels of granularity we added a further status for each treatment line, that is, *IO* (immunotherapy). This is doable thanks to the possibility in the PWF formalism to define simultaneous activation of multiple status. Finally, we introduced two additional status to catch the survival outcomes, namely *Dead* and *Censored*, that can be activated without constraints on the previous status, as soon as a survival event is read in the EL. The activation of the survival status terminates the inspection of the flow of events for that patient.

Figure 4 reports the result of the run on our cohort. Nodes and boxes report the number of times that a status/trigger was reached/fired. Due to space constraints, we limited the plot to the first two lines of treatment, even if the PWF included all the 7 lines of treatments available in the data.

By inspecting the graph, it is possible to follow the population's paths and read the corresponding number of subjects that run specific patterns. For in-

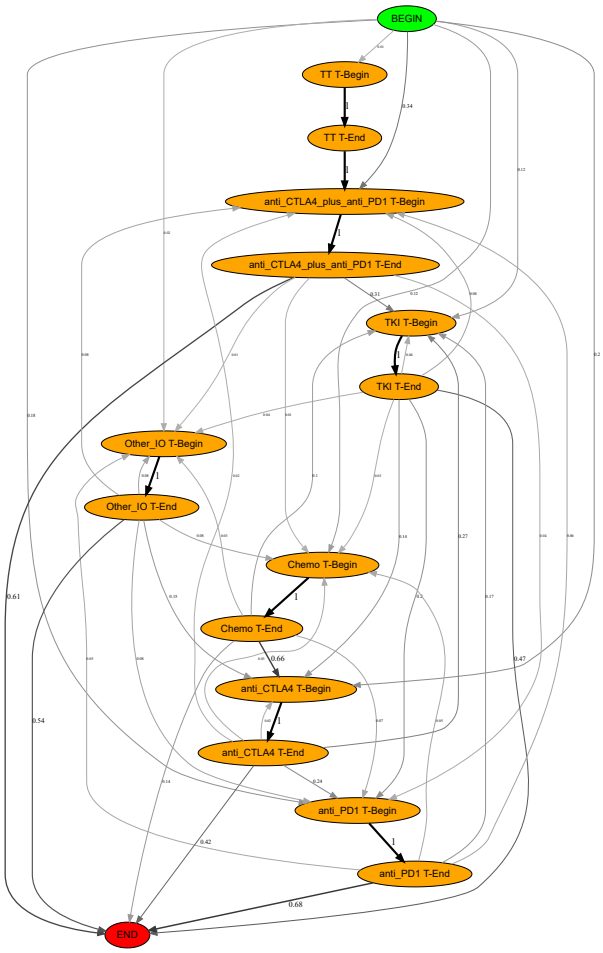

Fig. 3. First Order Markov Models obtained on the treatments.

stance, we can observe that all the patients included in the dataset (and thus with a BEGIN event) had a Stage IV diagnosis (expected by design), that the most frequent first line of treatment was the combination of anti-CTLA4 and anti-PD1 with a total of 56 occurrences, or that only 163 over 184 patients had a first line recorded, followed in 89 cases by a second line.

The survival nodes (*Dead* and *Censored*) are graphically separated from the others in order to limit the number of edges in the graph. However they can be reached from any point in the graph, and the available query tool can inspect at what precise point they were activated.

3.3 Inferential statistics

By exploiting the EL, the FOMM and the PWF diagrams of the previous analyses, we could easily select cohorts characterized by specific patterns of interest

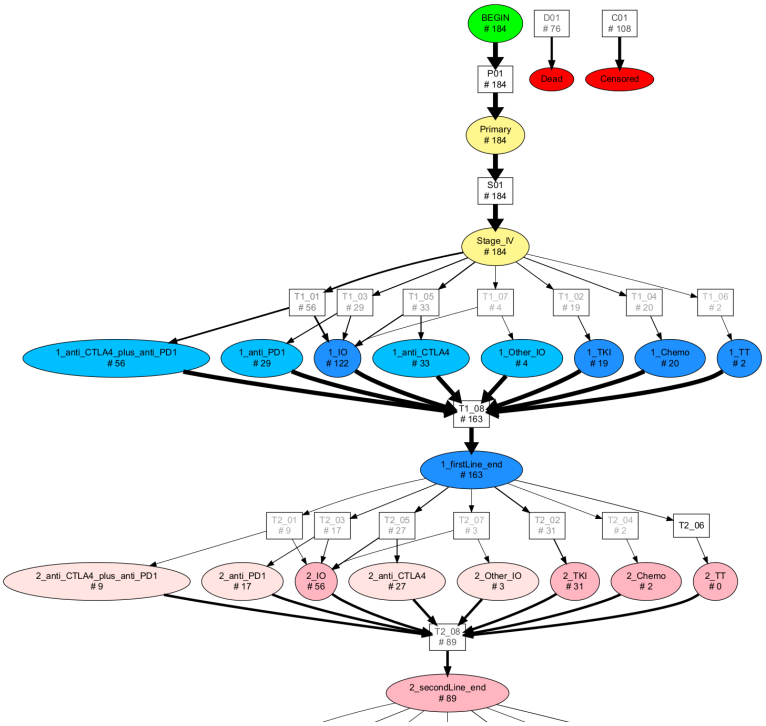

Fig. 4. Conformance Checking model (limited to the first two lines of treatments) reporting the status activated by the patients' processes over the used-defined PWF.

and perform survival analyses. While the FOMM strongly reflects (and is limited to) the events and the information present in the EL, the PWF represents an abstraction where the user has the opportunity to provide additional knowledge in the definition of the PWF structure itself. This enhanced semantic expressiveness is one of the main reasons why PWF was previously used in structuring Clinical Guidelines [5]. Descriptive statistics can help in suggesting hypotheses: in our case, the previous PWF and FOMM diagrams allowed to easily identify and query cohorts for statistical inference analyses. We report below two examples of the investigations we performed.

First, we inspected the relationship between type of somatic tumor mutation and time between primary and Stage IV diagnosis. Here, we consider the following mutation status: BRAF V600 mutated, BRAF non-V600 mutated, NRAS mutated, and wt. For this study, we limited the cohort to cutaneous melanoma patients, exploiting filtering tool to easily query the EL attributes.

We implemented a survival analysis by first using the FOMM structure of Figure 2 to query the path of interest (between the nodes Primary Stage and Stage IV) and obtain the time between the two events. Then, the Kaplan-Meier estimator is computed, with patients stratified by mutation status, as shown in Figure 5a). Even if a difference between the BRAF v600 mutated and the NRAS

mutated sub-cohorts seems to emerge, the log-rank test computed between all the survival distributions pairs report no significant differences (all p-values were >0.05) for any combinations.

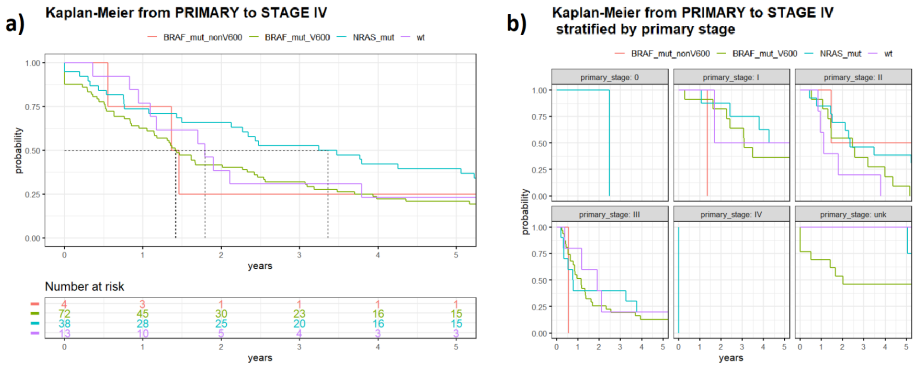

Fig. 5. Time-to-event analysis based on a mined FOMM: time from primary to stage IV diagnosis, stratified by: a) mutation, b) mutation and type of primary.

To demonstrate the potential of the analysis – even if in this case limited by the sample cardinality – we performed a further stratification of the data, distinguishing patients by their primary stage. Also here, pMineR facilitates this step, by allowing direct selection on the patient attributes. Figure 5b) reports the plot of the corresponding Kaplan-Meier estimator. Even if, as expected, no statistically significant clinical evidence emerges from this analysis, mainly due to the low number of subjects per category, it is interesting to observe how rapidly this approach allows to enrich the analysis' level of detail.

The second survival analysis exploits the PWF defined in Figure 4. We queried the data in order to identify any differences in terms of OS based on the following patterns of interest: (1) only IO (any BRAF status), (2) IO $\rightarrow$ TKI, (3) TKI $\rightarrow$ IO, (4) only TKI. In defining the rules, we grouped together consecutive lines belonging to the same category. Patterns interspersed with TT or Chemo treatments were excluded. Upon the suggestions of clinicians, in case of sequences with multiple treatment lines, only the first occurring pattern was considered. The resulting OS survival curves are shown in Figure 6. Table 3 reports the frequency of occurrence of each pattern, the median OS time (in years), and the percentage of patients alive at 1.5 and 3 years (CI at 95%), respectively. Statistical significance of OS differences was assessed with the log-rank test, which turned out to be significant for IO vs IO $\rightarrow$ TKI (p-value<0.0001) and IO vs TKI $\rightarrow$ IO (p-value: 0.012). The difference between IO and IO $\rightarrow$ TKI is expected because patients who receive TKI after IO are those who did not respond to IO. Knowing that the benefits of TKI are usually only temporary, it is not surprising that these patients have shorter OS. The difference between IO and TKI $\rightarrow$ IO is interesting, as it may be related to recent biological findings showing that acquired resistance to TKI may hinder IO efficacy.

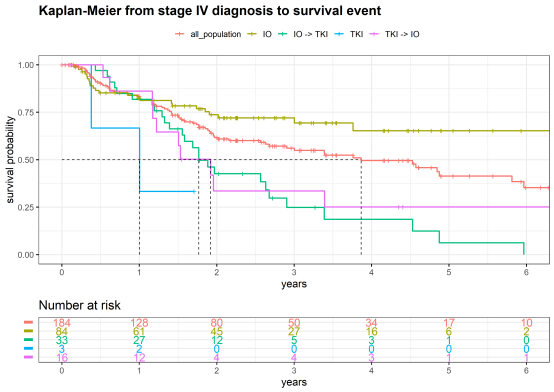

Fig. 6. Overall survival analysis based on a CC graph: time from stage IV diagnosis to death, stratified by treatment pattern.

Table 3. OS for the main treatment patterns of interest.

treatment path	**frequency**	**median OS [years]**	**1.5-year OS % (95% CI)**	**3-year OS % (95% CI)**
all	100 %	3.87	72.7 (66.1 - 80.1)	54.9 (47.1 - 64.1)
IO	45.7 %	NA	76.9 (68.0 - 86.9)	69.4 (59.1 - 81.5)
IO → TKI	17.9 %	1.77	63 (48.3 - 82.1)	18.6 (7.7 - 45.2)
TKI → IO	8.7 %	1.92	57.4 (36.6 - 90.1)	25.1 (9.7 - 65.3)
TKI	1.6 %	1.00	0	0

4 Discussion and Conclusion

PM4HC is expected to have an increasingly relevant role in the analysis of healthcare data, in particular in Oncology. Process-oriented representations, together with tools able to interrogate the data in terms of temporal patterns identified through paths in a workflow, are efficient ways to easily generate clinically-relevant hypotheses and measure statistical significance, in particular in survival analysis.

In this preliminary work, we demonstrated the added value of a process-oriented approach when performing three classical steps of data analysis: preprocessing, descriptive statistics, and inferential statistics. The main remarkable points emerging from this experience are: (a) query languages for EL, PD and CC are efficient tools for data cleaning and preprocessing, by quickly identifying previously unrecognized mistakes; (b) graphical representations can promote dialogue between clinicians and data scientists, suggesting alternative perspectives and possible research questions; (c) PD gives a relevant contribute in representing the data in an agnostic way; on the other hand CC (with formalisms such as PWF) allows implementing multi-scale data abstractions and identifying patterns or inconsistencies of the data in pre-defined workflows; (d) the process representations, both in PD and CC, effectively support survival analysis techniques, allowing rapid definition of sub-cohorts of interest and providing immediate statistical measures of differences between various paths of the graph.

Noticeably, each step of this study was performed in close cooperation between clinicians and PM scientists, in the effort of creating a multidisciplinary team with shared PM skills. The final goal will be to give full autonomy to physicians to perform PM analyses themselves.

In the future, PM4HC has great potential to be developed further in synergy with classical statistical tools to analyze healthcare-related data. In particular, the fast-growing amount of real-world clinical data produced in modern hospitals, each patient's therapeutic journey being by nature a temporal process, represents a formidable opportunity for PM4HC to contribute to the advent of precision medicine.

References

1. van der Aalst, W.: A practitioner's guide to process mining: Limitations of the directly-follows graph. Procedia Computer Science **164**, 321–328 (2019)
2. van der Aalst, W., Adriansyah, A., et al.: Process mining manifesto. In: Int. Conf. on Business Process Management. pp. 169–194. Springer (2011)
3. Bulliard, J., Panizzon, R., Levi, F.: Melanoma prevention in Switzerland: where do we stand? Revue medicale suisse **2**(63), 1122–1125 (2006)
4. Cowey, C.L., Liu, F.X., Boyd, M., Aguilar, K.M., Krepler, C.: Real-world treatment patterns and clinical outcomes among patients with advanced melanoma: A retrospective, community oncology-based cohort study (A STROBE-compliant article). Medicine (Baltimore) **98**(28), e16328 (Jul 2019)
5. Gatta, R., Lenkowicz, J., et al.: pMineR: an innovative R library for performing process mining in medicine. In: Conf. on Artificial Intelligence in Medicine in Europe. pp. 351–355. Springer (2017)
6. Gatta, R., Vallati, M., Lenkowicz, J., et al.: Generating and comparing knowledge graphs of medical processes using pminer. In: Proceedings of the Knowledge Capture Conference. K-CAP 2017, Association for Computing Machinery, New York, NY, USA (2017)
7. Geleijnse, G., Aklecha, H., et al.: Using process mining to evaluate colon cancer guideline adherence with cancer registry data: a case study. In: AMIA (2018)
8. Homayounfar, P.: Process mining challenges in hospital information systems. In: 2012 Federated Conf. on Computer Science and Information Systems. pp. 1135–1140. IEEE (2012)
9. Kurniati, A.P., Johnson, O., Hogg, D., Hall, G.: Process mining in oncology: A literature review. In: 2016 6th Int. Conf. on Information Communication and Management. pp. 291–297. IEEE (2016)
10. Lenkowicz, J., Gatta, R., et al.: Assessing the conformity to clinical guidelines in oncology: An example for the multidisciplinary management of locally advanced colorectal cancer treatment. Management Decision **56**(10), 2172–2186 (Oct 2018)
11. Mans, R., Schonenberg, H., Song, M., Aalst, W.V., Bakker, P.: Application of process mining in healthcare - a case study in a dutch hospital. In: BIOSTEC (2008)
12. Mans, R.S., van der Aalst, W.M., Vanwersch, R.J.: Process mining in healthcare: evaluating and exploiting operational healthcare processes. Springer (2015)
13. Rinner, C., Helm, E., Dunkl, R., et al.: Process Mining and Conformance Checking of Long Running Processes in the Context of Melanoma Surveillance. Int J Environ Res Public Health **15**(12) (12 2018)