Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240738

Temporal Characterization and Visualization of Revolving Therapy-Events in Lung Cancer Patients

Jonas HÜGEL[a,b,1], Donata A. SCHÄFER[c], Jan J. SCHNEIDER[a], Jiazi TIAN[d], Hossein ESTIRI[d,e], Raphael KOCH[c], Tobias R. OVERBECK[c] and Ulrich SAX[a,b]
[a] *University Medical Center Göttingen, Department of Medical Informatics, Göttingen, Germany*
[b] *University of Göttingen, Campus Institute Data Science, Göttingen, Germany*
[c] *University Medical Center Göttingen, Department of Hematology and Medical Oncology, Göttingen, Germany*
[d] *Department of Medicine, Massachusetts General Hospital, Boston, MA, USA*
[e] *Clinical Augmented Intelligence Group, Harvard Medical School, Boston, MA, USA*
ORCiD ID: Jonas Hügel https://orcid.org/0000-0002-4183-1287, Donata A. Schäfer https://orcid.org/0009-0001-2218-3974, Jan J. Schneider https://orcid.org/0000-0001-8317-875X, Jiazi Tian https://orcid.org/0000-0003-0599-3438, Hossein Estiri https://orcid.org/0000-0002-0204-8978, Raphael Koch https://orcid.org/0000-0002-2018-5685, Tobias R. Overbeck https://orcid.org/0000-0002-2579-0171, Ulrich Sax https://orcid.org/0000-0002-8188-3495

Abstract. This paper presents a comprehensive workflow for integrating revolving events into the transitive sequential pattern mining (tSPM+) algorithm and Machine Learning for Health Outcomes (MLHO) framework, emphasizing best practices and pitfalls in its application. We emphasize feature engineering and visualization techniques, demonstrating their efficacy in capturing temporal relationships. Applied to an EGFR lung cancer cohort, our approach showcases reliable temporal insights even in a small dataset. This work highlights the importance of temporal nuances in healthcare data analysis, paving the way for improved disease understanding and patient care.

Keywords. sequential pattern mining, temporal data analyses, ML workflow

1. Introduction

Electronic Health Records (EHRs) have transcended their primary billing and communication functions to emerge as invaluable repositories for unraveling patient journeys and gaining profound insights into the intricacies of complex diseases [1]. Traditional approaches, such as Pattern Mining and Machine Learning (ML), have been foundational in extracting valuable insights from healthcare data [1,2]. However, their efficacy can be hindered by the limitations of conventional techniques like one hot encoding, which overlooks the inherent temporal relationships and order of clinical

[1] Corresponding author: Jonas Hügel; E-mail: jonas.huegel@med.uni-goettingen.de.

records [3–6]. To overcome this challenge, Estiri et al. developed the transitive sequential pattern mining (tSPM) algorithm boosting downstream machine learning leveraging the Machine Learning for Health Outcomes (MLHO) framework [3,4,7,8]. Hügel et al. [9] presented tSPM+, an enhanced version of tSPM. Both algorithms define a sequence as a tuple of two events. Operating on the same principle, they transitively combine all events from a patient's history into sequences while providing seamless integration into existing machine learning workflows. Nevertheless, the oversight of inherent temporal relationships still poses a significant challenge [5,10], particularly in diseases characterized by revolving events or therapies, such as cancer, where the order of therapies is of the essence for the treatment success [11]. This paper contributes a comprehensive workflow integrating revolving events into the tSPM+ and MLHO framework. Moreover, through visualization of complex temporal relationships of significant events, we add a well understandable layer of information. We validated our approach on an EGFR lung cancer cohort comprising 200 patients.

2. Methods

2.1. Temporal Characterization with the Extended tSPM+ and MLHO Workflow

We extended the original tSPM+ and MLHO workflow [9] by a feature engineering step in the beginning and a visualization step in the end. Figure 1 visualizes the workflow.

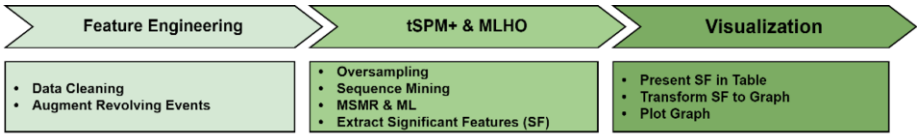

Figure 1. The extended tSPM+ and MLHO workflow: We extend the original tSPM+ and MLHO workflow [9] by a feature engineering step for revolving events and an additional step for the visualization revealing the temporal relationships of the significant sequences from the ML model as a network graph.

2.1.1. Feature Engineering

EHR data from cancer patients contains revolving events occurring multiple times in the patient trajectories, e.g. different therapies, and corresponding outcomes, such as a tumor progress or regress. Consequently, mining transitive sequences results in sequences starting with a therapy and ending with the outcome of a later therapy. To ensure encoding of direct therapy results, we add additional augmented events. We added those on different levels of coarseness, ranging from the event that the therapy took place (th_X) via the therapy type (th_X_systemic) to the specific drug used in the systemic therapy (th_X_Erlotinib). Additionally, we store a list of all unique coarse events for each coarseness level. Finally, we transform all events into the required input format of a table of triples (date, event, patient id), which we call from now on "dbMart".

2.1.2. Temporal Characterization Using tSPM+ and MLHO

We utilize the tSPM+ Docker container [9], which provides tSPM+ and MLHO in an RStudio instance, to apply tSPM+ and MLHO. We applied oversampling by quadrupling

each entry in the dbMart to stabilize the fluctuation in the ML model output introduced by our small sample size. Following the guide [9], we transformed the dbMart into numeric representations. Mining sequences that pair coarse events, such as th_X → th_X_systemic, would introduce noise. To avoid it, we erased all coarse events from the dbMart before mining all transitive sequences. Afterwards for each coarseness level, we iteratively: 1) add all events from that level to the dbMart, 2) mine sequences containing a coarse event and append them to previously mined, 3) remove coarse events from the dbMart. Finally, we applied the minimize sparsity maximize relevance (MSMR) algorithm [3] before handing significant sequences and demographics to MLHO.

2.1.3. Data Visualization

Following the aforementioned guide, we display the most significant sequences for the ML model in a table. While allowing to sort the sequences by their relevance, this approach does not reveal the overall temporal relationships of the significant events. Network visualization plays a crucial role in transforming data into insights by making complex relationships clear and easily interpretable. Therefore, plotting the sequences in a graph, where each event is a vertex and a connecting edge represents the corresponding sequence, allows us to identify patterns, trends, and potential causal effects that might be difficult to discern from a simple table.

2.2. Validation Case Study

We applied the workflow to a cohort of 200 lung cancer patients with EGFR mutations. Additionally, the data included timestamped features for the tumor classification, metastases information, co-mutations on gene level, up to twelve sequential cancer therapies as well as outcomes and demographic data. For the machine learning in MLHO, we utilized random forest (RF) as classifier with 5-fold cross validation to postdict the survival after 8 months. It is to mention, that the goal of the case study is not to derive new medical insights, but instead to show the feasibility and usability of the workflow.

3. Results

3.1. Feature Engineering

The feature engineering step has to be highly adjusted and therefore also reimplemented for each data set. We created four additional augmented events for each therapy and two augmented for the results. For therapy events, we encoded 1) if therapy number X took place (th_x), 2) the type of therapy number x (th_x_type), 3) just the type of the therapy (surgery, radiation or systemic) and 4) the name of of drug. For the outcome of the therapy, we created the augmented events 1) encoding the outcome alone (progression (PD), stable disease (SD), partial remission (PR) and complete remission (CR)) and 2) together with the number of the therapy (th_x_outcome).

3.2. Temporal Characterization Using tSPM+ and MLHO

After mining the sequences with tSPM+, we extracted the 70 most significant leveraging the MSMR algorithm. Following the MLHO workflow, we represented the most relevant

sequences from the ML model in a table including derived significance values from the underlying caret R-package [8,12]. Table 1 shows the five most significant features.

Table 1. Table displaying the five most important features for the applied ML model (RF) The feature importance is directly derived from the caret R-package and is normalized between 0 and 100, with 100 being the most significant feature. The feature importance is ML model dependent (see caret documentation).

Feature	Feature Importance
Systemic Therapy —> Best Result (BR): PD	100
th_1 —> BR: PD	86.7
th_1 —> Systemic Therapy	74.9
Systemic Therapy —> th1_BR: PD	72.8
th_1 —> Surgery	72.27

3.3. Data Visualization

We leveraged the iGraph R-package [13] to visualize the most significant sequences for the classification task in a directed graph. Figure 2 shows a detailed view of the graph.. The thickness of the edges correlates with the computed significance in the ML model allowing researchers to get a clear overview regarding the temporal relationship of the events, as well as their importance.

Figure 2. Image detail of a graph to visualize the most relevant sequences and events from the case study. Each vertex represents an event, each edge means that the sequence connecting the vertices is significant for the prediction target. The thickness of the edge is representing the feature importance value of the sequence. Reading example: Systemic Therapy->BR:PD: Progress in disease after a systemic therapy is a relevant feature to predict the death in the first 8 months, compare to Table 1. The full figure is available online [14].

3.4. Case Study

As anticipated, the case study did not report any unexpected results, instead the significant sequences in the ML model fitted our current knowledge, e.g. systemic therapy followed by progressive disease is crucial to postdict death in the first 8 months after the diagnoses. The sequences reveal that not only the type of therapy, but also its relative timing is important. Additionally, co-mutations, e.g. TP53, which is a known predictor in lung cancer patients with an EGFR mutation [15], were important features.

4. Discussion and Conclusions

Our work offers insights regarding the application, as well as the pitfalls and best practices when using the tSPM+ algorithm to reveal temporal relationships and leverage the mined knowledge in downstream ML workflows. It is to mention, that the employed case study, despite encompassing a small cohort, was still sufficient to prove the reliability of the workflow results. The targets we chose are not of groundbreaking

biomedical nature, but are rather a sanity check to prove the efficacy and reliability of the tSPM+ algorithm and the employed downstream techniques on a small data set. Unveiling new biomedical insights is subject to future works, e.g. using if the EGFR mutation is common or uncommon as postdiction target to classify the corresponding temporal relationships of the events from the patient trajectories.

Acknowledgment

This work is funded by the DFG (426671079) and BMBF (01KD2208A and 16DWWQP07C). The ethic committee (Chair: Prof. Dr. J. Brockmöller; vote: 2/5/23) of the University Medical Center Göttingen allowed the use of the data in the case study.

References

[1] Lokhandwala S, Rush B. Objectives of the Secondary Analysis of Electronic Health Record Data. In: Secondary Analysis of Electronic Health Records. Cham: Springer International Publishing; 2016. p. 3–7. doi: 10.1007/978-3-319-43742-2_1.

[2] Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. N Engl J Med. 2023 Mar 30;388(13):1201–8. doi:10.1056/NEJMra2302038.

[3] Estiri H, Strasser ZH, Klann JG, McCoy TH, Wagholikar KB, Vasey S, Castro VM, Murphy ME, Murphy SN. Transitive Sequencing Medical Records for Mining Predictive and Interpretable Temporal Representations. Patterns. 2020 Jul;1(4):100051. doi:10.1016/j.patter.2020.100051.

[4] Estiri H, Strasser ZH, Murphy SN. High-throughput phenotyping with temporal sequences. J Am Med Inform Assoc. 2021 Mar 18;28(4):772–81. doi:10.1093/jamia/ocaa288

[5] Moskovitch R. Multivariate temporal data analysis - a review. WIREs Data Min Knowl Discov. 2022 Jan;12(1):e1430. doi:10.1002/widm.1430.

[6] Moskovitch R, Choi H, Hripcsak G, Tatonetti N. Prognosis of Clinical Outcomes with Temporal Patterns and Experiences with One Class Feature Selection. IEEE/ACM Trans Comput Biol Bioinform. 2017 May 1;14(3):555–63. doi:10.1109/TCBB.2016.2591539.

[7] Estiri H, Azhir A, Blacker DL, Ritchie CS, Patel CJ, Murphy SN. Temporal characterization of Alzheimer's Disease with sequences of clinical records. eBioMedicine. 2023 Jun;92:104629.

[8] Estiri H, Strasser ZH, Murphy SN. Individualized prediction of COVID-19 adverse outcomes with MLHO. Sci Rep. 2021 Mar 5;11(1):5322. doi:10.1016/j.ebiom.2023.104629.

[9] Hügel J, Sax U, Murphy SN, Estiri H. tSPM+; a high-performance algorithm for mining transitive sequential patterns from clinical data. arXiv:2309.05671 [Preprint]. 2023 [cited 2024 June 05]:[18 p.]. Available from: https://arxiv.org/abs/2309.05671.

[10] Segura-Delgado A, Gacto MJ, Alcalá R, Alcalá-Fdez J. Temporal association rule mining: An overview considering the time variable as an integral or implied component. WIREs Data Min Knowl Discov. 2020 Jul;10(4):e1367. doi:10.1002/widm.1367.

[11] Baker K, Dunwoodie E, Jones RG, Newsham A, Johnson O, Price CP, Wolstenholme J, Leal J, McGinley P, Twelves C, Hall G. Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. Int J Med Inf. 2017 Jul;103:32–41. doi:10.1016/j.ijmedinf.2017.03.011.

[12] Kuhn M. Building Predictive Models in *R* Using the caret Package. J Stat Softw. 2008;28(5). [13] Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Syst. 2006;1695. doi:10.18637/jss.v028.i05.

[14] Hügel J, Schäfer DA, Schneider JJ, Tian J, Estiri H, Koch R, Overbeck TR, Sax U. Full Detail Figure 2 for the "Temporal Characterization and Visualization of Revolving Therapy-Events in Lung Cancer Patients" manuscript [Internet]. GRO.data; 2024 [cited 2024 Jun 5]. Available from: https://data.goettingen-research-online.de/citation?persistentId=doi:10.25625/SDJDMO.

[15] Roeper J, Falk M, Chalaris-Rißmann A, Lueers AC, Ramdani H, Wedeken K, Stropiep U, Diehl L, Tiemann M, Heukamp LC, Otto-Sobotka F, Griesinger F. TP53 co-mutations in EGFR mutated patients in NSCLC stage IV: A strong predictive factor of ORR, PFS and OS in EGFR mt+ NSCLC. Oncotarget. 2020 Jan 21;11(3):250–64. doi:10.18632/oncotarget.27430.