Analyzing Medical Data with Process Mining: a COVID-19 Case Study

Marco Pegoraro ✉[1], Madhavi Bangalore Shankara Narayana[1],
Elisabetta Benevento[1,3], Wil M.P. van der Aalst[1], Lukas Martin[2], and
Gernot Marx[2]

[1] *Chair of Process and Data Science (PADS), Department of Computer Science,*
RWTH Aachen University, Aachen, Germany
{`pegoraro, madhavi.shankar, benevento, vwdaalst`}`@pads.rwth-aachen.de`
[2] *Department of Intensive Care and Intermediate Care,*
RWTH Aachen University Hospital, Aachen, Germany
{`lmartin, gmarx`}`@ukaachen.de`
[3] *Department of Energy, Systems, Territory and Construction Engineering, University of Pisa, Pisa, Italy*

Abstract

The recent increase in the availability of medical data, possible through automation and digitization of medical equipment, has enabled more accurate and complete analysis on patients' medical data through many branches of data science. In particular, medical records that include timestamps showing the history of a patient have enabled the representation of medical information as sequences of events, effectively allowing to perform process mining analyses. In this paper, we will present some preliminary findings obtained with established process mining techniques in regard of the medical data of patients of the Uniklinik Aachen hospital affected by the recent epidemic of COVID-19. We show that process mining techniques are able to reconstruct a model of the ICU treatments for COVID patients.

Keywords: Process Mining · Healthcare · COVID-19.

Colophon

This document is an Author Accepted Manuscript (AAM) corresponding to the following scholarly paper:

Pegoraro, Marco, Madhavi Bangalore Shankara Narayana, Elisabetta Benevento, Wil M. P. van der Aalst, Lukas Martin, and Gernot Marx. "Analyzing Medical Data with Process Mining: a COVID-19 Case Study". In: *Business Information Systems Workshops*. Ed. by Abramowicz, Witold, Sören Auer, and Milena Stróżyna. Springer, 2022, pp. 39–44

Please, cite this document as shown above.

Publication chronology:

- 2021-05-01: full text submitted to the Workshop on Applications of Knowledge-Based Technologies in Business, work-in-progress track
- 2021-05-14: notification of acceptance
- 2021-05-19: camera-ready version submitted
- 2021-06-15: presented
- 2022-04-06: proceedings published

Correspondence to:
Marco Pegoraro, Chair of Process and Data Science (PADS), Department of Computer Science,
RWTH Aachen University, Ahornstr. 55, 52074 Aachen, Germany
Website: `http://mpegoraro.net/` · Email: `pegoraro@pads.rwth-aachen.de` · ORCID: 0000-0002-8997-7517

Content: 9 pages, 5 figures, 11 references. Typeset with pdfLaTeX, Biber and BibLaTeX.
Please do not print this document unless strictly necessary.

1 Introduction

The widespread adoption of Hospital Information Systems (HISs) and Electronic Health Records (EHRs), together with the recent Information Technology (IT) advancements, including e.g. cloud platforms, smart technologies, and wearable sensors, are allowing hospitals to measure and record an ever-growing volume and variety of patient- and process-related data [7]. This trend is making the most innovative and advanced data-driven techniques more applicable to process analysis and improvement of healthcare organizations [5]. Particularly, *process mining* has emerged as a suitable approach to analyze, discover, improve and manage real-life and complex processes, by extracting knowledge from event logs [1]. Indeed, healthcare processes are recognized to be complex, flexible, multidisciplinary and ad-hoc, and, thus, they are difficult to manage and analyze with traditional model-driven techniques [9]. Process mining is widely used to devise insightful models describing the flow from different perspectives—e.g., control-flow, data, performance, and organizational.

On the grounds of being both highly contagious and deadly, COVID-19 has been the subject of intense research efforts of a large part of the international research community. Data scientists have partaken in this scientific work, and a great number of articles have now been published on the analysis of medical and logistic information related to COVID-19. In terms of raw data, numerous openly accessible datasets exist. Efforts are ongoing to catalog and unify such datasets [6]. A wealth of approaches based on data analytics are now available for descriptive, predictive, and prescriptive analytics, in regard to objectives such as measuring effectiveness of early response [8], inferring the speed and extent of infections [2, 10], and predicting diagnosis and prognosis [11]. However, the process perspective of datasets related to the COVID-19 pandemic has, thus far, received little attention from the scientific community.

The aim of this work-in-progress paper is to exploit process mining techniques to model and analyze the care process for COVID-19 patients, treated at the Intensive Care Unit (ICU) ward of the Uniklinik Aachen hospital in Germany. In doing so, we use a real-life dataset, extracted from the ICU information system. More in detail, we discover the patient-flows for COVID-19 patients, we extract useful insights into resource consumption, we compare the process models based on data from the two COVID waves, and we analyze their performance. The analysis was carried out with the collaboration of the ICU medical staff.

The remainder of the paper is structured as follows. Section 2 describes the COVID-19 event log subject of our analysis. Section 3 reports insights from preliminary process mining analysis results. Lastly, Section 4 concludes the paper and describes our roadmap for future work.

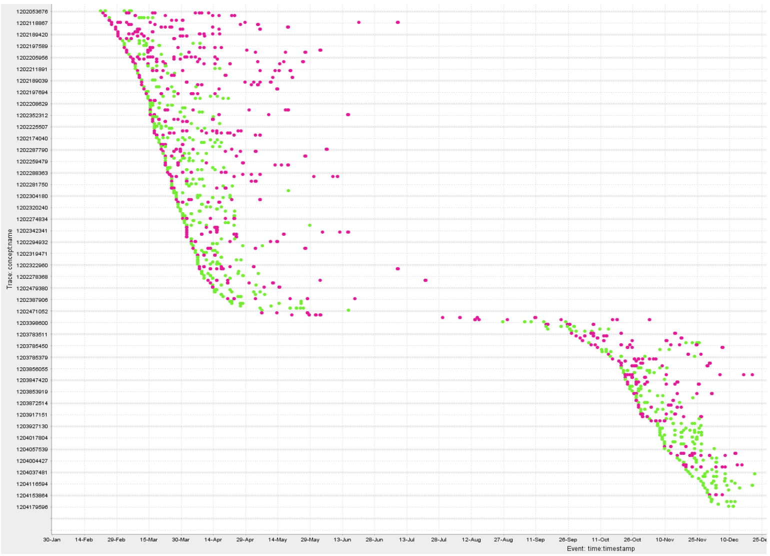

Figure 1: Dotted chart of the COVAS event log. Every dot corresponds to an event recorded in the log; the cases with Acute Respiratory Distress Syndrom (ARDS) are colored in pink, while cases with no ARDS are colored in green. The two "waves" of the virus are clearly distinguishable.

2 Dataset Description

The dataset subject of our study records information about COVID-19 patients monitored in the context of the COVID-19 Aachen Study (COVAS). The log contains event information regarding COVID-19 patients admitted to the Uniklinik Aachen hospital between February 2020 and December 2020. The dataset includes 216 cases, of which 196 are complete cases (for which the patient has been discharged either dead or alive) and 20 ongoing cases (partial process traces) under treatment in the COVID unit at the time of exporting the data. The dataset records 1645 events in total, resulting in an average of 7.6 events recorded per each admission. The cases recorded in the log belong to 65 different variants, with distinct event flows. The events are labeled with the executed activity; the log includes 14 distinct activities. Figure 1 shows a dotted chart of the event log.

3 Analysis

In this section, we illustrate the preliminary results obtained through a detailed process mining-based analysis of the COVAS dataset. More specifically, we elaborate on results based on control-flow and performance perspectives.

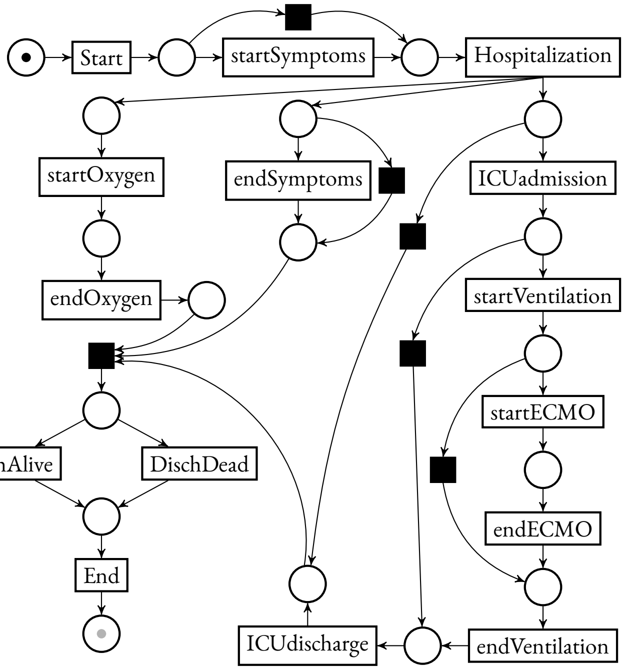

Figure 2: A normative Petri net that models the process related to the COVAS data.

Firstly, we present a process model extracted from the event data of the COVAS event log. Among several process discovery algorithms in literature [1], we applied the Interactive Process Discovery (IPD) technique [3] to extract the patient-flows for COVAS patients, obtaining a model in the form of a Petri net (Figure 2). IPD allows to incorporate domain knowledge into the discovery of process models, leading to improved and more trustworthy process models. This approach is particularly useful in healthcare contexts, where physicians have a tacit domain knowledge, which is difficult to elicit but highly valuable for the comprehensibility of the process models.

The discovered process map allows to obtain operational knowledge about the structure of the process and the main patient-flows. Specifically, the analysis reveals that COVID-19 patients are characterized by a quite homogeneous high-level behavior, but several variants exist due to the possibility of a ICU admission or to the different outcomes of the process. More in detail, after the hospitalization and the onset of first symptoms, if present, each patient may be subject to both oxygen therapy and eventually ICU pathway, with subsequent ventilation and ECMO activities, until the end of the symptoms. Once conditions improve, patients may be discharged or transferred to another ward.

We evaluated the quality of the obtained process model through conformance check-

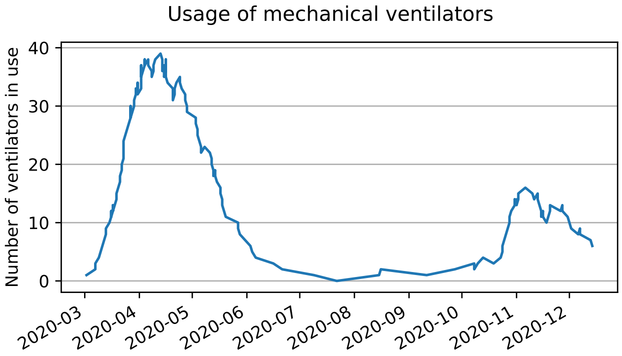

Figure 3: Plot showing the usage of assisted ventilation machines for COVID-19 patients in the ICU ward of the Uniklinik Aachen. Maximum occupancy was reached on the 13th of April 2020, with 39 patients simultaneously ventilated.

ing [1]. Specifically, we measured the token-based replay fitness between the Petri net and the event log, obtaining a value of 98%. This is a strong indication of both a high level of compliance in the process (the flow of events does not deviate from the intended behavior) and a high reliability of the methodologies employed in data recording and extraction (very few deviations in the event log also imply very few missing events and a low amount of noise in the dataset).

From the information stored in the event log, it is also possible to gain insights regarding the time performance of each activity and the resource consumption. For example, Figure 3 shows the rate of utilization of ventilation machines. This information may help hospital managers to manage and allocate resources, especially the critical or shared ones, more efficiently.

Finally, with the aid of the process mining tool Everflow [4], we investigated different patient-flows, with respect to the first wave (until the end of June 2020) and second wave (from July 2020 onward) of the COVID-19 pandemic, and evaluated their performance perspective, which is shown in Figures 4 and 5 respectively. The first wave involves 133 cases with an average case duration of 33 days and 6 hours; the second wave includes 63 patients, with an average case duration of 23 days and 1 hour. The difference in average case duration is significant, and could have been due to the medics being more skilled and prepared in treating COVID cases, as well as a lower amount of simultaneous admission on average in the second wave.

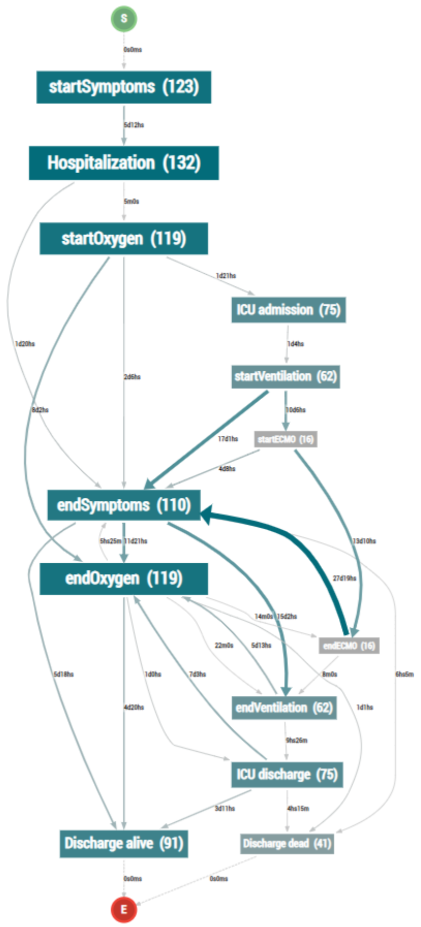

Figure 4: Filtered directly-follows graph related to the first wave of the COVID pandemic.

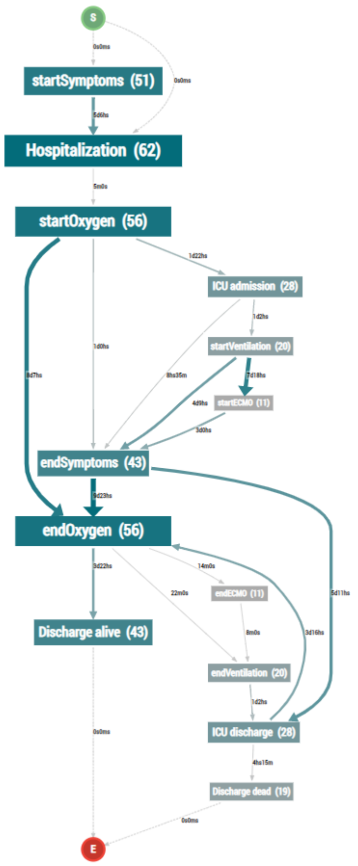

Figure 5: Filtered directly-follows graph related to the second wave of the COVID pandemic.

4 Conclusion and Future Work

In this preliminary paper, we show some techniques to inspect hospitalization event data related to the COVID-19 pandemic. The application of process mining to COVID event data appears to lead to insights related to the development of the disease, to the efficiency in managing the effects of the pandemic, and in the optimal usage of medical equipment in the treatment of COVID patients in critical conditions. We show a normative model obtained with the aid of IPD for the operations at the COVID unit of the Uniklinik Aachen hospital, showing a high reliability of the data recording methods in the ICU facilities.

Among the ongoing research on COVID event data, a prominent future development certainly consists in performing comparative analyses between datasets and event logs geographically and temporally diverse. By inspecting differences only detectable with process science techniques (e.g., deviations on the control-flow perspective), novel insights can be obtained on aspects of the pandemic such as spread, effectiveness of different crisis responses, and long-term impact on the population.

Acknowledgements

We acknowledge the ICU4COVID project (funded by European Union's Horizon 2020 under grant agreement n. 101016000) and the COVAS project for our research interactions.

References

[1] van der Aalst, Wil M. P. *Process Mining - Data Science in Action, Second Edition*. Springer, 2016. ISBN: 978-3-662-49850-7. DOI: 10.1007/978-3-662-49851-4.

[2] Anastassopoulou, Cleo, Lucia Russo, Athanasios Tsakris, et al. "Data-based analysis, modelling and forecasting of the COVID-19 outbreak". In: *PloS one* 15.3 (2020), e0230405.

[3] Dixit, Prabhakar M., H. M. W. Verbeek, Joos C. A. M. Buijs, et al. "Interactive Data-Driven Process Model Construction". In: *Conceptual Modeling - 37th International Conference, ER 2018, Xi'an, China, October 22-25, 2018, Proceedings*. Ed. by Trujillo, Juan, Karen C. Davis, Xiaoyong Du, et al. Vol. 11157. Lecture Notes in Computer Science. Springer, 2018, pp. 251–265. DOI: 10.1007/978-3-030-00847-5_19.

[4] *Everflow Process Mining*. https://everflow.ai/process-mining/. [Online; accessed 2021-05-17].

[5] Galetsi, Panagiota and Korina Katsaliaki. "A review of the literature on big data analytics in healthcare". In: *Journal of the Operational Research Society* 71.10 (2020), pp. 1511–1529. DOI: 10.1080/01605682.2019.1630328.

[6] Guidotti, Emanuele and David Ardia. "COVID-19 Data Hub". In: *Journal of Open Source Software* 5.51 (2020). Ed. by Rowe, Will, p. 2376. DOI: 10.21105/joss.02376.

[7] Koufi, Vassiliki, Flora Malamateniou, and George Vassilacopoulos. "A Big Data-driven Model for the Optimization of Healthcare Processes". In: *Digital Healthcare Empowering Europeans - Proceedings of MIE2015, Madrid Spain, 27-29 May, 2015*. Ed. by Cornet, Ronald, Lacramioara Stoicu-Tivadar, Alexander Hörbst, et al. Vol. 210. Studies in Health Technology and Informatics. IOS Press, 2015, pp. 697–701. DOI: 10.3233/978-1-61499-512-8-697.

[8] Lavezzo, Enrico, Elisa Franchin, Constanze Ciavarella, et al. "Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'". In: *Nature* 584.7821 (2020), pp. 425–429.

[9] Mans, Ronny S., Wil M. P. van der Aalst, and Rob J. B. Vanwersch. *Process Mining in Healthcare - Evaluating and Exploiting Operational Healthcare Processes*. Springer Briefs in Business Process Management. Springer, 2015. ISBN: 978-3-319-16070-2. DOI: 10.1007/978-3-319-16071-9.

[10] Sarkar, Kankan, Subhas Khajanchi, and Juan J Nieto. "Modeling and forecasting the COVID-19 pandemic in India". In: *Chaos, Solitons & Fractals* 139 (2020), p. 110049.

[11] Wynants, Laure, Ben Van Calster, Gary S Collins, et al. "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal". In: *British Medical Journal* 369 (2020).