Towards the Application of Process Mining for Supporting the Home Hospitalization Service*

Roberto Aringhieri[1], Guido Boella[1], Enrico Brunetti[1 2], Luigi Di Caro[1], Chiara Di Francescomarino[3], Mauro Dragoni[3], Roger Ferrod[1], Chiara Ghidini[3], Renata Marinello[2], Massimiliano Ronzani[3], and Emilio Sulis[1]

[1] University of Turin, Torino, Italy
{roberto.aringhieri,guido.boella,enrico.brunetti, luigi.dicaro,roger.ferrod,emilio.sulis}@unito.it
[2] City of Health and Science, Torino, Italy
{ebrunetti,rmarinello}@cittadellasalute.to.it
[3] Fondazione Bruno Kessler, Trento, Italy
{dfmchiara,dragoni,ghidini,mronzani}@fbk.eu

Abstract. Providing quality hospital services, especially when decisions need to be made, highly depends on the suitability and efficiency of the underlying processes, as well as on the capability of monitoring, analysing and using the data of process executions so as to provide operational support to decision makers. Process Mining can be a useful instrument in this setting. In this extended abstract we report about a real-life healthcare scenario, that is supporting the Home Hospitalization Service Team of an Italian hospital in making decisions about the home hospitalization of patients. We sketch the high-level idea of a solution leveraging Natural Language Processing and Process Mining for achieving the goal and report about some preliminary results, as well as about criticalities and challenges arisen so far.

Keywords: Healthcare processes · Predictive Process Monitoring · Natural Language Processing

1 Introduction

Improving healthcare processes and supporting clinical personnel in making decisions might have a high impact on the quality of life of patients. Process Mining (PM) [1], which deals with the analysis of business processes based on their behaviour - observed and recorded in event logs - can be a useful instrument in this setting. PM deals with the analysis of business process event logs in different ways [2], including process discovery (i.e., extracting process models from an event log) [1], predictions of the future of ongoing cases [7] and process optimization [1]. PM techniques can be leveraged for the discovery and analysis of both

This research has been partially carried out within the "Circular Health for Industry" project funded by "Compagnia di San Paolo" under the call "IA, uomo e società".

clinical and administrative processes in healthcare. The literature related to PM applied to the healthcare domain is not negligible and keeps on growing [9].

The application of PM techniques is further encouraged by the wide availability of administrative and clinical data in hospitals. This data could be leveraged for discovering (and improving) processes, as well as for supporting hospital teams in making decisions on clinical and administrative issues [3, 11]. Furthermore, it often happens that this data are collected in national standard forms and documents, shared among several hospitals on the national area. For instance, in Italy, one of these documents is the Hospital Discharge Form (HDF), which collects information related to the clinical history of a patient during his/her hospitalization. The data collected in the discharge form range from data (with temporal information) related to the hospital admission, discharge and examinations carried out during the hospitalization to data such as the number of days of hospitalization. While for some of these fields the content is also standardized, as in the case of the ICD-9-CM[4] codes for the examinations, this is not the case for textual fields, which, although very informative, are also highly unstructured.

In this extended abstract we report about a real-life healthcare scenario. We first introduce the scenario and the goal we would like to achieve in such a context (Section 2); we then report about the high-level idea of a possible solution to achieve such a goal (Section 3); and we finally conclude discussing some of the technological and methodological challenges arisen so far (Section 4).

2 The Home Hospitalization Service Scenario

The Home Hospitalization Service (HHS) of the City of Health and Science (CHS), which has been in operation for over 30 years, has proven to be a valid alternative to hospitalization for a variety of acute and chronic exacerbated diseases [10], such as uncomplicated ischemic stroke, congestive heart failure, exacerbations of chronic obstructive pulmonary disease, onco-hematological diseases with high transfusion requirements, dementia with behavioral disorders [6]. The HHS consists of a multidisciplinary team. The essential criteria for taking care of an acute patient at home are threefold: (i) clinical aspects, e.g., no need for continuous or invasive monitoring of vital parameters, as well as to perform invasive diagnostic-interventions; (ii) geographical aspects (residence in the area of competence of the HHS); (iii) social welfare (constant presence of one or more caregivers, formal or informal). Every year, the service manages about 500 admissions of patients coming in most cases from the same hospital and in small part upon direct request of the General Practitioner (GP). At the end of the treatment period, more than 80% of patients are discharged to the GP, 10.5% die during hospitalization and about 8% is moved to hospital. Over the past 8 years, the percentage of patients unable to continue care management at home has remained constant, despite the increase in clinical complexity and care burden of patients taken into care. In 2018, HHS patients were 492 with a high

[4] https://www.cdc.gov/nchs/icd/icd9cm.htm

average age (about 84 years).The overall goal is supporting the HHS team in the timely identification and notification of the patients that can be managed through the HHS, as well as in the efficient management of the HHS processes.

Data Description. The administrative and clinical data available so far for the specific case study are related to Emergency Department Discharge Forms (EDDF) and to the Hospitalization Discharge Forms (HDF) of about 400 CHS patients benefitting from the HHS. The EDDF contains information collected at the ED such as: (i) date and time information related to the ED admission, triage, discharge, last and latest update of the anamnesis; (ii) structured information e.g., on the patient triage colour code and on the ICD-9-CM diagnosis code; and (iii) textual notes e.g., on the diagnosis. The HDF contains instead information about the clinical history of the patient during the hospitalization, such as: (i) date and time information related to e.g., the hospital admission, discharge, main intervention; (ii) structured information related to e.g., patients' personal data, number of visits; and (iii) textual information related to e.g., the hospitalization cause and the anamnesis.

3 The Overall Idea

In order to support the HHS team in making decisions on the home hospitalization of a patient, the high-level idea is applying existing approaches of Predictive Process Monitoring (PPM) to the data automatically produced by the administrative and clinical management of ED patients. PPM approaches, indeed, learn from historical data (e.g., the history of past patients) and predict the future of incomplete executions (e.g., whether the patient will be successfully hospitalized or not). To this aim, the available data need first to be preprocessed and analysed. Moreover, since the available data contain an important quantity of informative textual data, Natural Language Processing (NLP) techniques can be applied and combined with PPM approaches.

Fig. 1 summarises the three steps of the pipeline. After a first phase, in which data are preprocessed, integrated and cleaned, textual data (e.g., clinical diary, diagnostic hypothesis) can be analysed through NLP techniques, so as to extract structured information from unstructured data. Finally, PPM approaches are applied to structured data (converted in an event log) and unstructured data transformed into structured ones.

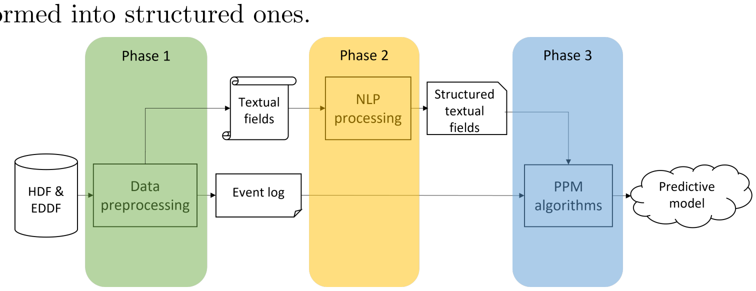

Fig. 1. The overall pipeline

Data Preprocessing and Analysis. The dataset related to the HDFs extracted from the hospital information systems has first been cleaned by removing hospitalizations of few days and then joined with the one of the ED. The dataset has then been transformed into an event log. The hospital discharge id number has been used as *trace id*. For the HDF data, hospital admission (`H_admission`), discharge (`H_discharge`) and the interventions carried out (labelled with the corresponding ICD-9-CM code) have been used as activities, while the corresponding date and time fields as timestamps. Patient personal data and other structured data, such as the ICD-9-CM code of the diagnosis or the setting of referral, have been added as case attributes. Similarly, for EDDF data, date and time fields related to the ED admission, discharge, triage, anamnesis and diagnostic hypothesis have been used as timestamps for the `ED_admission`, `ED_discharge`, `ED_triage`, `ED_anamnesis`, `ED_diagnostic_hp`, respectively. Anamnesis, diagnosis and other few attributes have been instead used as case attributes. The resulting event log is composed of 413 cases with 412 different paths and 219 different activities. Fig. 2 shows the the directly-follows graph (related to the 20% most frequent activities) extracted from the event log, as visualized by Disco[5]. While the figure is not meant to be readable, it gives an idea of the different paths characterizing the log - with only few activities shared among several paths.

Moreover, data have been labelled according to whether (i) the patient has been hospitalized at home and the hospitalization had a positive outcome (HH, i.e., Home Hospitalization); or (ii) she/he has been hospitalized in a different ward or the home hospitalization had a negative outcome (NO-HH). Out of the 413 cases, 368 (89%) were labeled with HH and 45 (11%) with NO-HH.

Predicting the Home Hospitalization Outcome. Once data have been preprocessed, NLP techniques can be used to extract structured data from textual fields carrying useful information for deciding about the home hospitalization of a patient. For instance, the presence of a caregiver who lives with the patient is a critical factor for the home hospitalization. However, this information is not explicitly tracked in a structured field; it is instead hidden in textual fields describing with whom the patient has reached the ED. In the analysed dataset, we have overall about 2,500 notes related to patients' anamnesis or to other clinical aspects with an average length of about 94 words.

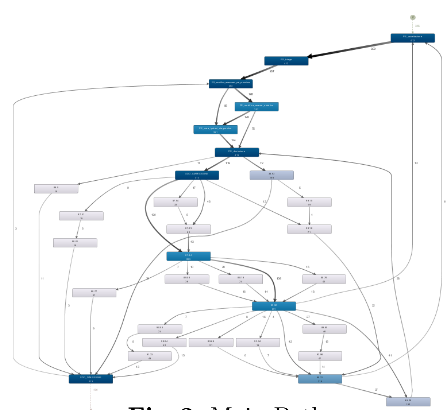

Fig. 2. Main Paths

One of the main criticalities of this phase is related to the fact that the textual fields of the analysed data contain several typos, acronyms and medical (but not

[5] https://fluxicon.com/

necessarily technical) terms, thus hampering the extraction of structured features and requiring a further preprocessing step for the typos correction. In order to face the issue of typos introduced by doctors quickly writing down notes, we conducted a first analysis on a sample of 200 clinical notes, for a total of 9374 words. Among these, 241 (2.57%) contain easily recognizable typos (i.e. errors that give rise to non-existent words in the vocabulary), while 28 (0.30%) are real-word errors (i.e. meaningful words but not the intended words in the context), thus requiring an analysis of the context of the sentence. We also observed that around 60% of the errors are related to common-sense words, while the remaining errors are related to medical terms, thus demanding for a rich medical dictionary.

The structured data, either extracted from the non-structured fields or already stored in structured fields, can then be provided as input to PPM algorithms that use these features to learn a predictive model (as shown in Fig. 1). At runtime, when the HHS team has to decide whether a new patient should undergo the home hospitalization, given the features of the new patient, the predictive model will predict whether it is likely that she/he will successfully undergo home hospitalization (HH) or whether it is better to proceed with the hospitalization in another ward (NO-HH). PPM algorithms, e.g., the ones available in Nirdizati [8], a PPM tool that collects a rich set of state-of-the-art approaches based on machine learning algorithms, can be used to train a predictive model able to learn the correlations between variables that describe the patient data and examinations he/she has carried out (features) and the hospitalization at home or in another hospital ward. We plan to evaluate the approach on a test set, obtained by ordering the event log and taking the last 20% as test set, while using the first 80% as training set. Predictions are then obtained for each of the cases of the test set before the ED discharge. A preliminary analysis carried out by using as features only structured data, i.e., without leveraging any information from the textual fields, with different PPM algorithms shows promising results. We are indeed able to obtain a F-measure score higher than 0.75[6]. Furthermore, explanation techniques can be used in order to understand which features impact most on the home hospitalization of a patient.

One of the main criticalities of this phase is related to the level of granularity to be used with examinations (i.e., whether examinations should be considered at a very low level of detail or abstracted and aggregated), as well as with other features, such as the patient's setting of referral (e.g., ED, GP, in-hospital ward).

Process mining techniques can be used to strengthen the development of online optimization algorithms with lookahead to manage processes in real time, which is extremely important and challenging especially in hospitals. Hospital processes are indeed characterized by many sources of uncertainty making ex-ante planning not robust enough and, by consequence, determining inefficiency in terms of outputs and outcomes [5]. Online optimization algorithms, which demonstrated their validity in the management of several health processes (e.g., operating room planning, emergency medical services) can hence be used to ensure the delivery of efficient and of quality health services [4].

[6] We considered as positive the HH outcome and as negative the NO-HH outcome.

4 Discussion and Conclusions

We have described a real-life healthcare scenario related to the HHS and, with the aim of supporting the HHS team in making decisions, we have introduced a possible pipeline. From a technological viewpoint, the main challenges we have faced up to now are related to the quality of the unstructured data containing several typos and acronyms, as well as to the activity granularity. We plan (i) to leverage more advanced techniques that take into account contextual information and other medical texts in Italian to solve the issues related to unstructured data; and (ii) to use NLP and semantic knowledge to identify the right level of abstraction for activities. However, the technological challenges are not the only ones to face. From a methodological viewpoint, the main challenge is indeed enabling the communication and the collaboration of a team of persons with different background and expertise. We plan to further strengthen the collaboration among the team members and continue building a shared vocabulary.

References

1. van der Aalst, W.M.P.: Process Mining - Data Science in Action, Second Edition. Springer (2016)
2. van der Aalst W. M. P. et al.: Process mining manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part I. LNBI, vol. 99, pp. 169–194. Springer (2011)
3. Amantea, I.A., Sulis, E., Boella, G., Marinello, R., Bianca, D., Brunetti, E., Bo, M., Fernandez-Llatas, C.: A process mining application for the analysis of hospital-at-home admissions. Stud Health Technol Inform **270**, 522–526 (2020)
4. Aringhieri, R.: Online optimization in health care delivery: Overview and possible applications. Operations Research Proceedings 2019 pp. 357–363 (2020)
5. Duma, D., Aringhieri, R.: An ad hoc process mining approach to discover patient paths of an emergency department. Flex. Serv. Manuf. J. **32**(1), 6–34 (2020)
6. Isaia, G., Bertone, P., Isaia, G.C., Ricauda, N.: Home care for patients with chronic obstructive pulmonary disease. Arch Phys Med Rehabil **100**, 664–5 (2010)
7. Maggi, F.M., Di Francescomarino, C., Dumas, M., Ghidini, C.: Predictive monitoring of business processes. In: Advanced Information Systems Engineering - 26th International Conference, CAiSE 2014 June 16-20, 2014. LNCS, vol. 8484, pp. 457–472. Springer (2014)
8. Rizzi, W., Simonetto, L., Di Francescomarino, C., Ghidini, C., Kasekamp, T., Maggi, F.M.: Nirdizati 2.0: New Features and Redesigned Backend. In: Demonstration Track at BPM 2019, September 1-6, 2019. CEUR Workshop Proceedings, vol. 2420, pp. 154–158. CEUR-WS.org (2019)
9. Rojas, E., Munoz-Gama, J., Sepúlveda, M., Capurro, D.: Process mining in healthcare: A literature review. J. Biomed. Informatics **61**, 224–236 (2016)
10. Sulis, E., Amantea, I.A., Boella, G., Marinello, R., Bianca, D., Brunetti, E., Bo, M., Bianco, A., Cattel, F., Cena, C., et al.: Monitoring patients with fragilities in the context of de-hospitalization services: An ambient assisted living healthcare framework for e-health applications. In: 23rd ISCT. pp. 216–219. IEEE (2019)
11. Sulis, E., Terna, P., Di Leva, A., Boella, G., Boccuzzi, A.: Agent-oriented decision support system for business processes management with genetic algorithm optimization: an application in healthcare. J. Med. Syst. **44**(9), 1–7 (2020)