The 7th IEEE International Conference on E-Health and Bioengineering - EHB 2019
Grigore T. Popa University of Medicine and Pharmacy, Iasi, Romania, November 21-23, 2019

Bridging the Gap between Process Mining and DES Modeling in the Healthcare Domain

Oscar Tamburis
Dept. of Veterinary Medicine and Animal Productions, Federico II University, Naples, Italy, oscar.tamburis@unina.it

***Abstract*—The present paper aims at describing an original approach to link Process Mining techniques and Discrete Event Simulation modeling, via the implementation of ProM6 framework, in a hospital–based scenario. A specific methodology was elaborated, each step of which was validated in order to investigate the conformance between the original Event Log, and the Simulation tool data.**

Keywords—Process Mining; Descrete Event Simulation; ProM6; Simul8; ophthalmology.

I. INTRODUCTION

An increasing interest is currently ongoing for business processes simulation, because of the possibility to analyze the behavior of any kind of processes and to know how to improve them. More specifically, a Discrete Event Simulation model (DES model) is a stochastic model through which a system, along its workflows, is figured out as a network of queues and activities. The activities are presented as a discrete sequence of events in time, interspersed with queues. DES modeling allows understanding how systems work, thus allowing users to improve, among the other things, skills of problem setting and solving, bottleneck analysis, as well as decision-making. Nonetheless, issues exist as to the treatment of the data, the building time of a Simulation model, the knowledge of the attributes of the process, or the validation of the model [1]. The present paper investigates a first attempt of how DES model building can be coped with via the implementation of Process Mining (PM) techniques, which allow the analysis of business processes based on event logs. The Event log is the amount of information data about a process, recorded sequentially from the case ID (the physical customer), plus the events, the related timestamps and other attributes as resources, costs, etc. PM allows gaining insight into various aspects, such as the process (or control flow) perspective, as well as the performance, data, and organizational perspectives [2]. In particular, the research purpose was to build a DES Model, starting from the Event log data stored in Process Mining tools, related to the '*cataract process*' of the ophthalmology ward of a Dutch Hospital.

II. METHODS

As depicted in Fig.1, in order to obtain the control flow of the process, data were analyzed implementing ProM6 (an extensible, platform–independent framework, as it is implemented in Java, which supports a wide variety of process mining techniques in the form of plug-ins). Once obtained the structure, it was possible to extract from ProM6 the routing out probabilities between the activities in form of absolute occurrences during the process [3]; following some calculations it became then possible to obtain a distribution percentage profile that could be imported or put manually for each activity in Simul8. The model was enriched with: (i) *timing data* (ProM6 allows the user to obtain, in form of tables, the timing of the transactions between the process activities); (ii) *resources performances* (these can be obtained in form of number of executions per each activity; the data can be treated in Excel as well, in order to obtain a resources distribution profile). Worth noticing that the research only focused on these attributes, as directly case–related. In a second step, an adaptation algorithm between the software was created, and the methodology to validate the data imported to enrich the DES model in Simul8 was investigated: in particular, a viable way was found out to extract process data, in form of Event log, during the Simulation run. To that end, two more tools were deployed: Excel (as it is directly connected to Simul8) and Disco tool (as part of the PM techniques, and used to convert the Excel format into the '.xes' format, supported by ProM6). Moreover, in the ProM6 environment the user can treat the data with 'Conformance Checking' plug-ins: this technique allows finding valid parameters to demonstrate the conformance between Simulation data and process data [4], in order to guarantee a simulation as close as possible to reality thanks to a carefully tuning made by analyzing information extracted by logging data.

III. MODEL DESIGN

In order to build up the Event log to analyze, only patients of the ophthalmology ward that had their process starting point after the March 29th at 11:30:00 CEST 2014 until the November 22nd at 10:30:00 CEST 2018, were taken into account [5]. The Event log contains more than four thousand events and seventeen activities. Table 1 reports the main activities considered (both in Dutch and English languages) as well as their characterization within the whole process organization. The Event log of the process was represented via Inductive miner Infrequent Petri-Net (IMi) [6], so as on the one hand to respect particular PN-related formal criteria, such as Fitness and Generalization [7], and on the other hand to deploy a robust enough algorithm to treat noisy data. The Petri-Net was then converted in a BPMN model [8] via the 'Convert a petri-net into BPMN diagram' plug-in, and eventually exported directly in Simul8 – a commercial DES tool that does not allow any direct Petri-Net import – undergoing specific qualitative

978-1-7281-2603-6/19/$31.00 ©2019 IEEE

checks, i.e.: numbers and names of events (some activities could be lost in the importing, or the names could be changed), or layout of the diagram (the position of the events must be the same).

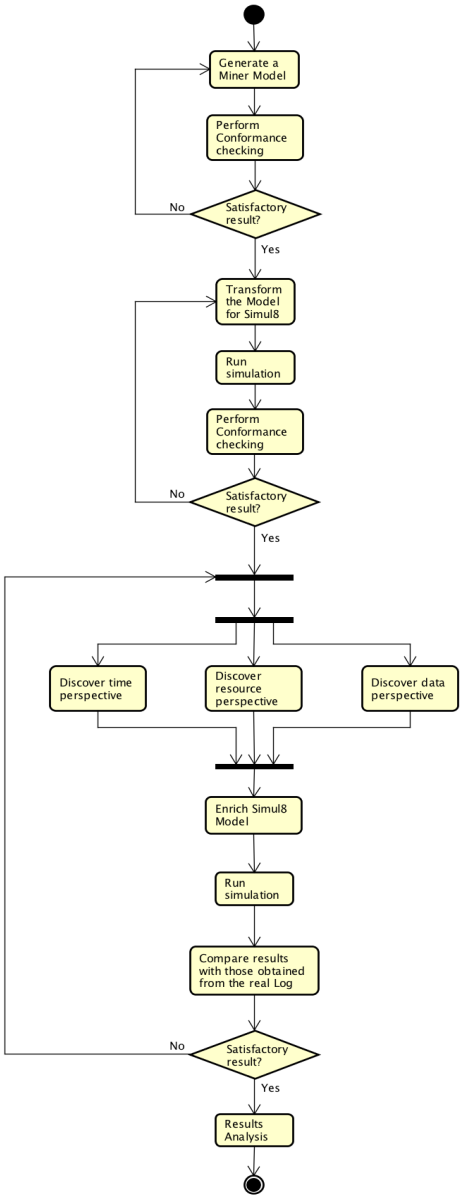

Fig. 1. Activity Diagram showing the general algorithm of the project that links ProM6 and Simul8 environments

A. Routing probabilities profile

Using ProM6 makes possible to obtain good information as to the probability distributions profile related to the process activities; this can be used as input to refine the behavior of the DES model. The ProM6 Log summary reports all the information related to the system, such as: (i) absolute occurrences of the activities in the process; (ii) occurrences of the start of the activities in the process; (iii) percentage of the utilization of the resources in the process. Therefore, Simul8 allows the user to create the routing probability profile of each activity in a specific routing out window. In order to create the

TABLE I. MAIN PROCESS ACTIVITIES

Activity (Dutch)	Activity (English)	Refers to:
Cornea-corneatopografie	Eye scans	Preoperative examination
Oog-IOL-master	Optical biometry	Preoperative examination
Dagverpleging-zwaar	Heavy day time nursing	Services performed on the day of surgery
Oogl.-extractaps. Impl. Lens.a.o.k.	Eye lens extraction and implanting a new lens	Services performed on the day of surgery
Snijtijd enkelvoudige zitting	Making an incision in a surgical session in which only cataract is treated	Services performed on the day of surgery
Zittingduur enkelvoudige OK	Being in surgical session in which only cataract is treated	Services performed on the day of surgery
Telefonisch consult	Phone consultation	Non–hospital–based activity
Vervolgconsult algemeen	General consultation	Non–hospital–based activity
1e consult algemeen	First general consultation	Non–hospital–based activity

right distribution profile for each activity to be exported in the DES model, the Petri Net–related BPMN diagram is defined by parallel and exclusive gateways to represent different constructs of the process; for each possible construct, a formula was figured out to obtain the right routing probabilities profile of each activity. The routing probability is expressed in percentage by means of the P_{ai} value, where: 'a' represents the activity the user is taking the actual value from, and; 'i' represents one of the three cases that the user can find in the BPMN structure, i.e. transition gateway-parallel; transition gateway-activity (w/o back); transition gateway-activity (with back). Once obtained the P_{ai} for all the activities, it is possible to use these probabilities in the Simul8 model, so as to define the routing out percentages for each single activity. For space reasons, only the formula for the third construct, used to evaluate P_{end3}, $P_{aspraak3}$ and $P_{AN\ Anker\ verrichtigen3}$ (Fig.2), is shown,

$$P_{a3} = \frac{n}{\#of\ activities} \qquad (1)$$
$$\sum_{n=1} (n - ns)$$

where: 'n' = number of the absolute occurrences of the i–th process activity; 'ns' = number of the occurrences of start of the i–th process activity.

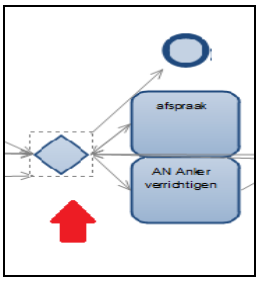

Fig. 2. Example of the the third possible case the user can find in a BPMN model. The gateway is followed from three activities with possibility of back

B. Timing perspective

Timing is a process attribute that affects all the components of the process. It is not allowed to record an Event log of a

process without its timing view. Timestamps and frequencies can be used to learn probability distributions that adequately describe waiting and service times, e.g. working times of the healthcare operators. PM techniques demonstrate that the replay techniques used for conformance checking can be modified to add the time perspective to process models [9]. In our case, the 'Replay a Log on Petri-net for Conformance analysis' plug-in was implemented to replay the Event Log on the Petri-net model, in order to optimize the mapping of the Log traces and, as a consequence, improve the alignment between the original Log and the final DES model. The plug-in gives back timing view tables as output, where time is expressed in average time, minimum and maximum value; the tables can be exported in Excel and used for a triangular distribution in the DES model. Simul8 uses such distribution to represent the transaction time between two activities.

C. Evaluation of resources performances

In a DES model it is possible to define the actors (called "resources") who actually perform the activities, along with their characteristics, such as the overall workload, or the percentage of working time for a specific time period. It is important to get exact information concerning the resources, since the better the resource characterization is, the more realistic the DES model becomes. For such reason, it was important to investigate in particular the workload-dependent behavior: traditionally the "Yerkes-Dodson Law of Arousal" suggests that a worker under time pressure may become more efficient and thus accomplish tasks in a faster way. However, if the pressure is too high, the worker's performance may degrade [10]. In our case, the ProM6 'Perform predictions of Business process features' plug-in was exploited, which made possible to discover some important rules underlying a process, i.e. 'in which conditions a resource performs an activity?', 'who does what?', or 'how often?'. In this way, ProM6 also provided information like the Correctly Classified Instances, Incorrectly Classified Instances, mean absolute square error, the total number of Instances and the Confusion matrix. The latter in particular contains the information regarding the number of resources that perform a specific activity. Rows represent the single resources; columns report the number of executions, as a function of the selected activity. For each row it is possible to calculate: 'X_n executions' = sum of the executions for each resource; 'X_n executions Av' = average number of executions of one resource for a specific activity, and number of resources involved for that activity. In formulae:

$$X_n\ executions\ Av = \frac{\sum_{resources} X\ executions(resource)}{\#\ resources} \qquad (2)$$

The resource probability profile, in form of conditioned probability, is then calculated as follows:

$$P(resource \mid Activity) = \frac{\sum_{resources} X\ executions(resource)}{\sum_{resources} X\ executions(resources\ chosen)} \qquad (3)$$

A timely criterion was then adopted to determine the resources with the greater amount of work, in order to choose them before the evaluation of the P in (3): in particular, the resources were chosen for the present experimentation which featured a number of executions greater than the average of the total executions of the resources related to a specific activity. Of course, the choice of the criteria generally depends on the purposes of the analysis performed [see e.g. 11]. The information about the working percentages in function of the executions of the single resource as well as the total number of executions for that activity, were then treated in Excel, in order to obtain for each activity a 'resource perspective summary'. Following, in Simul8 environment a distribution profile of the work of the resources for a single activity was defined, and a series of labels were used to link the profile to the resources as well as to the activities.

IV. Main Findings

The adaptation in a Simulation tool environment of a model elaborated from the Event log of a given process via PM techniques, is not a trivial matter: the focus is in fact on verifying in our case the overall compatibility of the files exported from ProM6 and to be imported in Simul8. Moreover, most of the plug-ins do not always support export operations or, even if it is possible, the data need to be timely treated, in order to fill the final DES model – sometimes information need to be inserted in the DES tool manually. For these reasons, for the '*cataract process*' under exam a validation system, articulated in three phases corresponding to the three steps of the model design, was figured out. The Conformance Checking was conducted in each phase starting from the extraction of all the data (work-item, events, timestamps, etc.) during the Simulation run of the DES model in form of Event log; to this purpose a Visual logic (programming system in Simul8) code interface of the Event log, which runs for each activity, was built. This made possible for Simul8 to report the Event log in a spreadsheet that was exported directly in Excel and, after that, treated through Disco tool, in order to get to the '.xes' format, necessary in turn to import the file in ProM6. The most timely plug-in was then implemented to check the compatibility between the IMi Petri-Net model (obtained in ProM6 from the original data) and the Event log (obtained from the Simulation run) through the abovementioned Fitness parameter. According to the literature, a reliable alignment occurs for a Fitness parameter value comprised between 0.95 and 1 [12]. For what concerns the validation of the routing probabilities profile of the BPMN model, a Trial simulation was performed, so as to obtain statistically significant result. In particular, it was chosen to make a ten-runs Trial to validate the results for the study case. As previously introduced, it was necessary to confront the ProM6–related log summary with the Event log obtained from the ten runs of the DES model. No plug-ins were in this case available to make such comparison, so it was chosen to follow a quantitative approach [13], articulated as follows: (i) creation of the '.xes' format for each of the ten Simul8 runs; (ii) import each run in ProM6, so to obtain the ProM Log summary, which contains the absolute and relative occurrences of each process activity; (iii) export the summary in HTML, so to have the HTML file for each run-related Event Log; (iv) export the HTML files in Excel, so as to have also the possibility to calculate, for instance, the average between the absolute occurrences of the same activities in the Trial; (v) making up a whole Simulation Event log,

which represent the average of the runs in the Trial, to be further compared with the original Event Log. The comparison occurred between the working time percentage of each activity in the original Log and the average working time percentage of each activity in the Simul8 Event log. The two-sample Kolmogorov-Smirnov (KS) test [14], which compares two data sets to decide whether they were sampled from similarly shaped population distributions, was deployed via Matlab, considering a I type error $\alpha = 0{,}05$. Since it resulted $p \approx 0{,}02 < \alpha$, it was possible to refuse the H_0 hypothesis (the result origins from random causes) and to affirm that the two distributions were from the same population. As said, it is possible to obtain the correct timing view (average, minimum and maximum values of time in the transaction between activities) in minutes using ProM6 plug-ins. These values can represent a triangular distribution to add to the DES model in Simul8: the upper, lower and modal values correspond to maximum, minimum and average ones in the ProM6 table. The Simul8 model had to be enriched with the real start date (Timestamps) of the process activities; this means that at the end of the already described steps the 'Replay a Log on Petri-net for Conformance analysis' plug-in, after being deployed to confront the IMi Petri-Net model and the Simulation-related Event log, was also used to obtain the timing perspective tables for each Trial run, exportable as CSV file. It was so possible to calculate in Excel an average of the timing of the single run-related Logs, which were compared with the original tables. In this case, the comparison was only qualitative, since the Event log obtained from the original data of the cataract process only featured the Start Events of the process activities – thus missing End Events as well as activity durations. The evaluation of the Resources distribution profile in the DES model coped eventually with the work-load of the resources as well as with the possibility to obtain, starting from the number of the executions of the resources in each activity, the percentage of working time, both free and blocked. This perspective is called Resources Event log, and turns out as extremely meaningful to understand the organizational side of the process. Also in this case, the '.xes' format was obtained as already seen and exported in ProM6. The deployment of the 'Perform predictions of Business process feature' plug-in was then needed to confront for each activity the Prom Log summary with the Simulation-related Event log, in order to detect the presence of errors in the manual transport of data in Simul8. No matching points were found between the ProM6 data and the DES model data in this case, so the formers could only be used to somehow oversee the organizational perspective of the process in the DES model.

V. Conclusions

In this study a new approach was described to figure out a DES model using Process Mining techniques, for a specific hospital-related process, trying to overcome issues related to: treatment of the data provided from the organization; building time of a DES tool; gaining knowledge as of the meaningful process attributes; validation of the resulting model. The main idea of the project – i.e. to adapt the process mining Tool, ProM6, with the commercial Simulation tool, Simul8 – was in the overall successfully completed. The original concept idea comes from a critical evaluation of both tools: Simul8 is a commercial tool, which does not allow any system changes, but the user can import, export, connect with Excel, save the work, and many other high level features; ProM6 is a platform that makes possible to build different projects related to the PM techniques in Java source; this allows in many case to export, and to import '.xes' format files. At the same time, it was necessary to influence both the Event log characteristics in the Process Mining tool, as well as the DES model, therefore the adaptation basically included other tools: Disco tool and Microsoft Excel. This project shows how the adaptation was possible, describing the steps carried out to get to the final result. More studies are clearly required to find more solid evidence as to the usefulness of PM techniques for determining the DES models construction. This study provides however a robust starting point, since links between the two techniques effectiveness were found and demonstrated.

References

[1] M.J. Glover, E. Jones, K.L. Masconi, M.J. Sweeting, S.G. Thompson, SWAN Collaborators, "Discrete Event Simulation for Decision Modeling in Health Care: Lessons from Abdominal Aortic Aneurysm Screening", Medical Decision Making, **38**(4), 2018, pp. 439-451.

[2] R.S. Mans, M.H. Schonenberg, M. Song, W.M. van der Aalst, P.J. Bakker, "Application of process mining in healthcare–a case study in a dutch hospital", in International joint conference on biomedical engineering systems and technologies, Springer, Berlin, Heidelberg, 2008, pp. 425-438.

[3] W.M. Van Der Aalst, Process mining: discovery, conformance and enhancement of business processes (Vol.2), Heidelberg: Springer, 2011.

[4] G. Sedrakyan, J. De Weerdt, M. Snoeck, "Process-mining enabled feedback: tell me what I did wrong vs. tell me how to do it right", Computers in human behavior, **57**, 2016, pp. 352-376.

[5] M.M. Overduin, Exploration of the link between the execution of a clinical process and its effectiveness using process mining techniques, Technische Universiteit Eindhoven, 2013.

[6] www.simul8.com

[7] T. Allweyer, BPMN 2.0: introduction to the standard for business process modeling, BoD–Books on Demand, 2016.

[8] N. Lohmann, M. Song, P. Wohed (Eds.), Business Process Management Workshops: BPM 2013 International Workshops, Beijing, China, August 26, 2013, Revised Papers (Vol. 171). Springer.

[9] F. Mannhardt, M. De Leoni, H.A. Reijers, W.M. van Der Aalst, "Balanced multi-perspective checking of process conformance", Computing, **98**(4), 2016, pp. 407-437.

[10] L.E. Chaby, M.J. Sheriff, A.M. Hirrlinger, V.A. Braithwaite, "Can we understand how developmental stress enhances performance under future threat with the Yerkes-Dodson law?", Communicative & integrative biology, **8**(3), 2015, e1029689.

[11] W.M. van der Aalst, M. Netjes, M., H.A. Reijers, "Supporting the full BPM life-cycle using process mining and intelligent redesign", in Contemporary issues in database design and information systems development, Igi Global, 2007, pp. 100-132.

[12] S. Bernardi, J.I. Requeno, C. Joubert, A. Romeu, "A systematic approach for performance evaluation using process mining: the POSIDONIA operations case study", in Proceedings of the 2nd International Workshop on Quality-Aware DevOps, ACM, 2016, pp.24-29.

[13] H. Nguyen, M. Dumas, A.H. ter Hofstede, M. La Rosa, F.M. Maggi, "Business process performance mining with staged process flows", in International Conference on Advanced Information Systems Engineering, Springer, Cham, 2016, pp. 167-185.

[14] H. Hassani, E. Silva, "A Kolmogorov-Smirnov based test for comparing the predictive accuracy of two sets of forecasts", Econometrics, **3**(3), 2015, pp. 590-609.