Extraction of Data from a Hospital Information System to Perform Process Mining

Ricardo Alfredo Quintano Neira[a,b], Gert-Jan de Vries[c], Jennifer Caffarel[c], Erin Stretton[c]

[a]Industrial Engineering Department, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ, Brazil
[b]Philips Research Brazil, Barueri, SP, Brazil
[c]Philips Research - Healthcare, Eindhoven, the Netherlands

Abstract

The aim of this work is to share our experience in the data extraction from a Hospital Information System to perform a research study using process mining techniques. We present all steps we performed: research definition, mapping the normative processes, identification of tables and fields names of the database, and extraction of data; and detail lessons learned. Any mistakes made in the extraction phase will have implications on subsequent analyses, thus it is essential to devote a great deal of attention in this phase, even if it takes a long time to perform all activities with the goal of ensuring high quality of the extracted data. We hope this work can help other researchers to plan and execute the extraction of data for process mining research studies.

Keywords:

Hospital Information Systems, Process Mining, Database Management System

Introduction

Process mining consists of a set of techniques that enable the analysis of business processes using system data. Specialized algorithms treat the data identifying patterns and trends. The use of process mining algorithms allows to discover process models (process discovery), identify deviations in a process (conformance checking), identify bottlenecks and performance indicators (performance checking), and identify how information flows between resources using social networks [1,2].

The event log is the raw material for running process mining algorithms. It contains all events used to construct a journey map and has as main attributes a case ID that represent one instance of the process (in healthcare field it could be the patient or hospitalization identification), the activity performed (e.g. "Perform triage", "Discharge of patient"), and the date and time the activity was performed [1,3].

Several research studies have applied process mining techniques for healthcare and they are present in many medical fields such as cardiology, oncology, diabetes and clinical images [4][5]. For example, Forsberg et al. [6] performed a study to identify the reading chest radiograph process in a Picture Archiving and Communication System (PACS). Rattanavayakorn and Premchaiswadi [7] applied the "working together metric" of social miner techniques to understand the behavior of health care professionals when treating patients in a hospital in Bangkok. Mans et al. [8] applied the heuristics miner algorithm to discover and compare the stroke treatment process in four Italian hospitals. Also, they applied performance checking algorithms to discover the bottlenecks and performance indicators for the pre-hospital process. Huang et al. [9] presented a technique that creates a summary of the structure of clinical pathways. They applied the approach for four different diseases (bronchial lung cancer, colon cancer, gastric cancer, and cerebral infarction) discovering essential medical behaviors with specific execution order. Another study analyzed the control flow, organizational and performance perspectives of a gynecological oncology process in a Dutch hospital to obtain insights in the care flow [10].

We are performing a research study using process mining techniques to identify deviations and bottlenecks in a sepsis treatment process in a Brazilian hospital. We expect to identify actions (changes in the process) that can improve the sepsis treatment process. Our first step was the extraction of data from a Hospital Information System (HIS).

The aim of this work is to present the steps we followed to extract data from the HIS to perform the process mining work. The main purpose of this work is to share our experience regarding the data extraction and all the preparation work associated with this task. We hope that our experience can help other researchers to plan and execute the extraction of data for process mining research studies.

Methods

Below we present all steps we followed to extract the data from the HIS database.

1. Research definition

First we defined the research questions we want to answer in our work:

1. Which is the *AS-IS* (current process) sepsis treatment process of the hospital? How does the hospital staff treat septic patients?
2. Which are the deviations in the process? Do professionals perform activities in a different order than defined in the normative sepsis treatment process?
3. Which are the bottlenecks in the process? Are there activities in the process that are taking more time than expected?

4. What is the workload of each professional in the hospital? Could a heavy workload cause delays in the process?
5. Which actions can improve the process? For example, we would like to identify changes in processes that might reduce the time it takes to administer initial antibiotic therapy.

The research questions were crucial to understand which type of information we should extract from the database. For example, if we want to know the *AS-IS* process, we need to create an event log with the case identification (in our case it is the hospitalization identification), the activity type ("Registry of patient", "Triage", "Medical evaluation") and the completion date and time. If we want to analyze the workload of health care professionals then in the event log we need to add the health professional identification and extract information about all hospitalizations (not only of sepsis) to get a complete overview of health professionals tasks.

2. Mapping the process

In a first visit to the hospital, we analyzed the process the hospital applies for treating sepsis patients. We studied the hospital documents (sepsis guidelines and sepsis screening form), performed interviews with health professionals (2 physicians, 2 nurses, 2 nurse technicians, 1 quality analyst, 1 receptionist) and performed shadowing. The number of interviewers was four.

With all the information collected, we designed two models using the Business Process Model and Notation (BPMN). One model represents the sepsis treatment in the Emergency Department and the second the sepsis treatment in the Intensive Care Unit. The models represent the way health professionals should work when treating sepsis patients (normative models). Both models were updated and validated by the staff (4 physicians, 3 nurses, 2 nurse technicians, 1 quality analyst, 1 laboratory technician, 2 pharmacists) in a second visit to the hospital. The number of interviewers was five.

Figure 1 presents a simple model of a treatment process based in the Emergency Department model. We created it to help the reader to understand the extraction process. We linked the process activities to the next steps of the extraction.

The process models and all the information collected from the hospital visits were very important to guide us in the selection of the attributes to collect from an immense amount of attributes presented in the database. For example, in the "1. Register Patient" activity we needed to know the time and date this event happened, and who performed the registration.

3. Identification of tables and fields of the database

When we started this research, we had no knowledge about the HIS database structure. We did not know from which tables and fields we should extract the data we needed to perform our research.

We asked the HIS development team to give us guidelines of which tables and fields we should collect all the information needed for our research. For this task we created a spreadsheet containing all attributes that we needed to extract from the database based on the research questions and process models. We sent this spreadsheet together with the process models to the HIS development team and asked them to complete it.

We created two versions of the spreadsheet. The first one presented just one tab with the attributes needed by each step (activity) of the treatment processes. We denominate this tab as "Process Oriented". Since it was difficult and confusing for the development team to fill it, we created a second version of the spreadsheet that contained a new tab, denominated "HIS Modules Oriented", presenting the same attributes needed grouped by functionality of the system. We believe this module oriented view would help them to more easily associate the correct tables and fields of the database, since this is closer to their way of thinking (they do not necessarily know the clinical process that is followed, but they do know the modules used by the users). The main difference between the process and module tabs is the way that the attributes needed are organized. The "Process Oriented" tab presents the attributes grouped by each activity of the process, and the "HIS Modules Oriented" tab presents the same attributes grouped by each module of the HIS. To convert an item from the "Process Oriented" tab to the "HIS Modules Oriented" tab we identified the HIS module used by the clinical user to register the information (attribute) associated to the process activity.

Below we present both tab structures of the spreadsheet.

Process oriented tab

Table 1 presents a small sample of the first tab. It contains the following columns:

- Step name (A): name of the activity from the BPMN model. E.g., "1. Register Patient", "2. Perform Triage" from Figure 1;

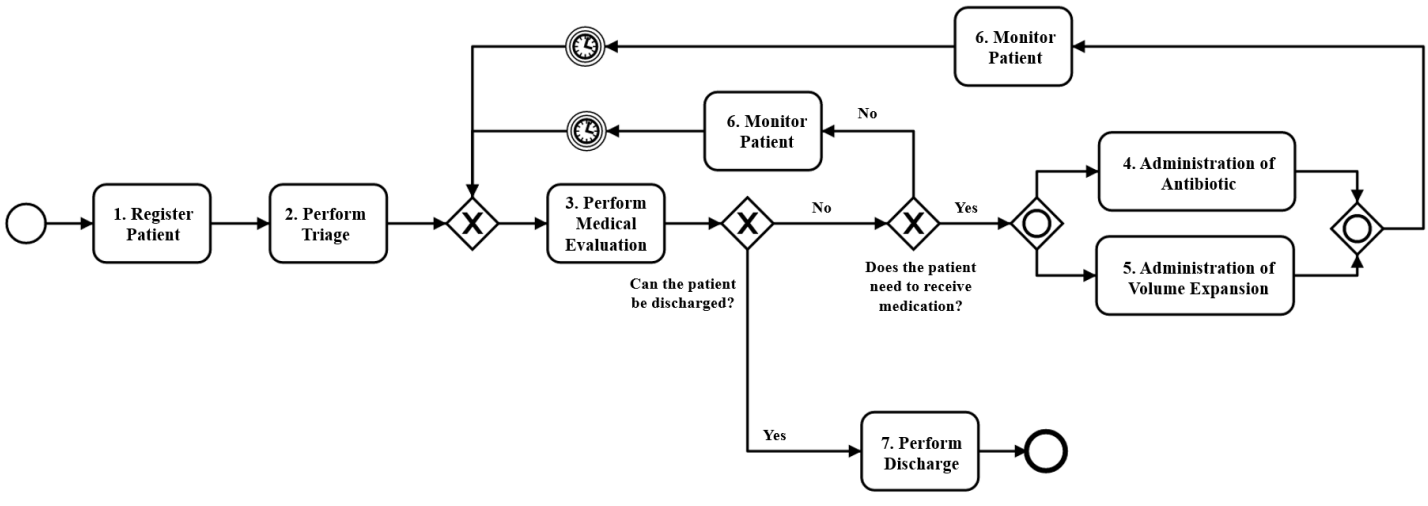

Figure 1- Simple BPMN model representing a treatment process in the emergency department (Note: this model was created as an elucidative example. It does not represent the exact and complete process as followed by the hospital)

Table 1 - Sample of process oriented tab of the spreadsheet (Note: all content data presented is fictitious. The idea is to present the structure of the table. The table presents a subset of the activities from Figure 1)

					Execution		Registry	
Step Name (A)	**Attribute (B)**	**Responsible (C)**	**Table (D)**	**Table Field (E)**	**Date and Time (F)**	**Role of User (G)**	**Date and Time (H)**	**Role of User (I)**
2. Perform Triage	Clinical Notes	ES	TRIAGE	clinical_notes	exec_date	exec_role	end_date	user_role
2. Perform Triage	Temperature	JC	VITAL_ SIGNS	temperature	start_date	perf_role	reg_date	reg_role
2. Perform Triage	Blood Pressure	JC	VITAL_ SIGNS	blood_pressure	start_date	perf_role	reg_date	reg_role
3. Perform Medical Evaluation	Clinical Notes	GJV	NOTES	clinical_notes	exec_date	exec_role	end_date	user_role
6. Monitor Patient	Temperature	JC	VITAL_ SIGNS	temperature	start_date	perf_role	reg_date	reg_role
6. Monitor Patient	Blood Pressure	JC	VITAL_ SIGNS	blood_pressure	start_date	perf_role	reg_date	reg_role

Table 2 - Sample of HIS module oriented tab of the spreadsheet (Note: all content data presented is fictitious. The idea is to present the structure of the table.)

					Execution		Registry	
Module (A)	**Attribute (B)**	**Responsible (C)**	**Table (D)**	**Table Field (E)**	**Date and Time (F)**	**Role of User (G)**	**Date and Time (H)**	**Role of User (I)**
Triage	Clinical Notes	ES	TRIAGE	clinical_notes	exec_date	exec_role	end_date	user_role
Vital Signs	Temperature	JC	VITAL_ SIGNS	temperature	start_date	perf_role	reg_date	reg_role
Vital Signs	Blood Pressure	JC	VITAL_ SIGNS	blood_pressure	start_date	perf_role	reg_date	reg_role
Electronic Health Record	Clinical Notes	GJV	NOTES	clinical_notes	exec_date	exec_role	end_date	user_role

- Attribute (B): name of the attribute that we need from the process activity. E.g., from the triage step (2. Perform Triage) we need to extract the "temperature", "blood pressure", and "clinical notes". One step name can have many attributes;
- Responsible (C): contact information of the responsible from the HIS development team who filled the spreadsheet line. This is important in case we had doubts about an attribute and thus we could contact directly the professional;
- Table (D): name of the table from the database where the field is located;
- Table field (E): name of the field from the database which contains the attribute information to be extracted;
- Execution - date and time (F): field from the database that contains the date and time that the action was performed in practice. E.g. "What was the time that the administration of the medication was performed for patient John?";
- Execution – role of user (G): field from the database that contains the role of the professional who performed the action. E.g. "Nurse", "Physician";
- Registry - date and time (H): field from the database that contains the date and time that the attribute was entered in the system. E.g. "What was the time that the administration of the medication for patient John was entered in the HIS?";
- Registry - Role of user (I): field from the database that contains the role of the professional who entered the attribute in the system. It is important to also retrieve this information as the person documenting the activity may not be the same as the one executing it (e.g. a doctor may perform an action and ask a nurse to document it in the HIS).

Columns from C to I should be filled by a professional from the HIS development team.

HIS modules oriented tab

Table 2 presents a small sample of the second tab. Columns B to I are the same from Table 1. The column "Step Name" (A) was replaced to "Module". This new column presents the module name from the HIS that the information requested can be collected. E.g. "Patient registry", "Electronic Health Record", "Computerized Physician Order Entry", "Imaging", "Transfer of Patient". Columns from C to I should be filled by a professional from the HIS development team.

Table 1 and Table 2 present the same content in different views.

Filling the spreadsheet by the HIS development team

We had one main contact person who was in charge to make the communication bridge with the development team. We sent him the spreadsheet and the BPMN process models with clear instructions on how to fill the spreadsheet. This procedure was performed with both versions of the spreadsheet. For the second

version, we made it clear that the developers could choose any of the two tabs to fill. During the "filling of the spreadsheet" step, we kept direct contact with the development team to solve their doubts. When receiving a new filled part of the spreadsheet, we immediately reviewed it to check if there was any white cell (cell not filled) and to check if the cells were with a coherent value (e.g. we asked for "prescription of medication" and we received "prescription_procedure" in the table name – this seems clearly to be not right). In case of any problem identified, we contacted them to discuss and update the information.

4. Extraction of data

Based upon the fields identified in the previous step, SQL queries were written to extract the relevant fields from the HIS. For the extraction we had to anonymize patient and hospitalization data to guarantee that no-one outside the hospital could identify a patient or link the extracted data with the hospital database. In this stage, to anonymize the data:

1. We encrypted any identification code like patient and health professional codes, chart number, hospitalization number, and prescription and administration ids;
2. Rather than storing the date of birth, we calculated the age of patients according to the admission hospitalization date. Patients older than 90 years old have a higher probability of being identified, thus all of these cases (> 90y) were classified as 90 years old;
3. Patients with a weight greater than 130kg also have a greater probability of identification, thus all of them (> 130kg) were classified as 130kg;
4. All extracted dates were shifted to a given time interval to further remove context which could lead to identification of patient data;
5. For text fields (like clinical notes and discharge summaries) we anonymized names of patients and professionals, specific numbers like chart, hospitalization, telephone, bed numbers.

Results

Regarding the "mapping the process" step (item 2 of the Methods section), several iterations were required to ensure we understood health professionals correctly and vice versa. For us it was a challenge to identify the commonalities and differences in treatment of different patient groups, as defined per severity, age, or list of comorbidities.

The "identification of tables and fields of the database" (item 3) was performed by 10 developers. All of them filled the module oriented tab. Only one developer filled both tabs. At the end we could successfully fill all cells of the spreadsheet. For this step it was fundamental to have a single contact point to orchestrate the work.

Regarding the "extraction of data" step (item 4), we extracted 4,516 sepsis hospital encounters for a period of two years. We also extracted all (not only sepsis) 61,260 hospital encounters for a period of 2 months, to collect information regarding the workload of professionals (to answer our fourth research question). All the information is present in 57 tables and more than 600 fields.

Discussion

Mapping the sepsis processes was not an easy task mainly because the communication between two different teams (health and information technology) is very challenging. The use of activity cards (cards filled by the hospital staff containing questions regarding each activity of the process; e.g. step description, notification process, tools used) helped us to understand better the processes. In addition, when validating the processes, the hospital staff could easily understand the BPMN notation (after an explanation of its elements). Thus, the BPMN models were important tools in the communication process between our team and the hospital staff.

Regarding the "identification of tables and fields of the database" step (item 3), all developers filled the module oriented tab. In our understanding, the process oriented tab was difficult for the development team to work with, since they had to search in all spreadsheet for the fields that they were responsible for. Indeed, for one step a clinical user may have to work in multiple modules, meaning that the attributes would be distributed over the system; and the development team is organized in such a way that sub-teams are responsible for individual modules. We believe that the module oriented tab helped them to easily identify the required fields.

It is important to mention that it was only possible to convert the process oriented view to the module oriented view since we had access to the hospital information system and we had shadowed clinical users during their use of the system. In addition, some of our researchers had previous experience in HIS architecture.

Columns F, G, H and I from the spreadsheet are used to collect the name of the fields regarding the time and user role that executed and registered an action. Professionals that execute actions are not necessarily the professionals who document them. These columns are very important for process mining research and they could lead to interesting insights when it is performed analyses using performance checking or social networks techniques. The executed time and user role might not always be available, and then the best alternative for process mining is to use the information of the registration of data, however taking into account the assumption made in subsequent analyses.

The extraction of non-structured fields is very important for process mining researches. These fields may have some information that can be converted to process activities or they can help to find answers for some questions. For example, when a deviation is discovered in a treatment process it is important to understand why certain cases took this unexpected path; some answers can be found studying clinical notes discovering for example, that patients that followed this deviation were in a critical stage of the disease. Friedrich et al. [11] discussed that it is estimated that 85% of the information of companies is stored in non-structure format and this source of information can be important for creating models. Most of process mining techniques need a very structured event log as input, and the use of natural language processing (NLP) techniques takes an important role for process mining research since it allows to convert unstructured data to concrete events.

Performing this work we could understand that to extract data from a HIS system is not an easy task: a lot of pre-work, knowledge about the treatment process but also about the HIS and its use, and collaboration with multi-disciplinary teams are needed. In general, the complete extraction process took us more time than we expected. Below we present the main reasons:

1. The HIS was not originally designed with the sepsis management pathway, and therefore it was a challenge for us to collect all the information that is relevant for the pathway and to identify where it was located in the system;
2. Limited time availability of the development team to support this initiative;
3. We had to deal with challenges to understand the data structure of the system;
4. We had to guarantee that all patient and hospitalization information was properly anonymized by our scripts (especially in the non-structured fields);
5. We had limited time to run the scripts to not compromise the use of the system by the hospital.

Our initial attempts of process mining analyses encompassed the evaluation of conformance of a simple sepsis process [12]. It took us great effort to check the quality of the data and to create a simple event log, however this opened up subsequent analyses. Later we started a more complete analysis and we could identify the AS-IS sepsis treatment process in the emergency department, identify deviations, and identify bottlenecks. During this investigation we had to deal with many challenges (e.g. creating specific events, filtering the right cases, dealing with missing timestamps) and it was very important to have the close participation of health professionals. The results of the research study are very promising. The next step is to validate the results with the hospital staff.

The investigation of health care processes is far from a trivial task. Health care processes tend to be very flexible and complex. Health care professionals play an important role providing health information, analyzing and interpreting results, getting insights and guiding to new analyses. Data scientists know methods and tools to organize and consolidate data in a proper way, and they can propose solutions to optimize resources and improve processes. To reconstruct, analyze and improve health care processes it is important to use smart approaches combining informatics/engineering techniques with health care knowledge. That means that data scientists and health professionals should work always together in all stages of a research to provide meaningful results. This approach should also be applied in the extraction phase.

We believe that the extraction steps we presented in this work (applying specific adjustments) could be applicable to other healthcare fields of research that uses, for example, data mining, simulation and neural networks techniques.

The extraction phase requires a lot of attention, active and clear communication with external teams, to guarantee that the extracted data will have quality and will allow to perform the research correctly. Getting the wrong data will result in wrong results for the entire research ("garbage in, garbage out" as they say).

Conclusions

With this work we shared our experience in extracting data from a hospital information system to perform a process mining research. Any mistakes made in the extraction phase will have implications on subsequent analyses, thus it is essential to devote a great deal of attention in this phase, even if it takes a long time to perform all activities with the goal of ensuring high quality of the extracted data.

Acknowledgements

The authors thank all professionals from the hospital and the HIS development team for supporting us in the extraction phase.

The authors thank CAPES and Philips Research for funding the development of this work.

References

[1] W. van der Aalst, *Process Mining: Data Science in Action*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2016. doi:10.1007/978-3-662-49851-4.

[2] Process Mining Group - Math&CS department - Eindhoven University of Technology, Process Mining, (2016). http://www.processmining.org/ (accessed December 11, 2016).

[3] A.O. García, D.P. Alfonso, and O.U.L. Armenteros, Analysis of Hospital Processes with Process Mining Techniques, in: *MedInfo*, 2015, pp. 310–314. doi:10.3233/978-1-61499-564-7-310.

[4] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, Process mining in healthcare: A literature review, *J. Biomed. Inform.* **61** (2016), 224–236. doi:10.1016/j.jbi.2016.04.007.

[5] W. Yang, and Q. Su, Process mining for clinical pathway: Literature review and future directions, in: *2014 11th Int. Conf. Serv. Syst. Serv. Manag.*, IEEE, 2014, pp. 1–5. doi:10.1109/ICSSSM.2014.6943412.

[6] D. Forsberg, B. Rosipko, and J.L. Sunshine, Analyzing PACS Usage Patterns by Means of Process Mining: Steps Toward a More Detailed Workflow Analysis in Radiology, *J. Digit. Imaging*. **29** (2016), 47–58. doi:10.1007/s10278-015-9824-2.

[7] P. Rattanavayakorn, and W. Premchaiswadi, Analysis of the social network miner (working together) of physicians, in: *2015 13th Int. Conf. ICT Knowl. Eng. (ICT Knowl. Eng. 2015)*, IEEE, 2015, pp. 121–124. doi:10.1109/ICTKE.2015.7368482.

[8] R. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini, and W. van der Aalst, Process mining techniques: an application to stroke care, *Stud. Health Technol. Inform.* **136** (2008), 573–578. http://www.ncbi.nlm.nih.gov/pubmed/18487792.

[9] Z. Huang, X. Lu, H. Duan, and W. Fan, Summarizing clinical pathways from event logs, *J. Biomed. Inform.* **46** (2013), 111–127.

[10] R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker, Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital, in: *BIOSTEC*, Springer, 2008, pp. 425–438. doi:10.1007/978-3-540-92219-3_32.

[11] F. Friedrich, J. Mendling, and F. Puhlmann, Process model generation from natural language text, in: *Int. Conf. Adv. Inf. Syst. Eng.*, 2011, pp. 482–496.

[12] G.-J. de Vries, R.A.Q. Neira, G. Geleijnse, P. Dixit, and B.F. Mazza, Towards Process Mining of EMR Data - Case Study for Sepsis Management, in: *BIOSTEC*, 2017.

Address for correspondence

Ricardo Alfredo Quintano Neira
e-mail: ricardo.quintano.pucrio@philips.com