Process Mining Model to Guarantee the Privacy of Personal Data in the Healthcare Sector

Sebastian Saavedra [1], José Llatas [1] and Jimmy Armas-Aguirre [1,2]

[1] *Universidad Peruana de Ciencias Aplicadas, Lima, Perú*
[2] *Pontificia Universidad Católica del Perú, Lima, Perú*

Abstract

In the paper, we propose a model to guarantee the privacy of patient data in critical processes in the healthcare sector through the application of process mining. Process mining is a discipline that discovers process models by analyzing event logs in order to identify bottlenecks and establish alternatives to improve their performance. In healthcare institutions, process mining is used to improve critical processes. However, event data logs containing confidential healthcare patient data are not protected when process mining and data visualization are applied. This definitely increases the risk of theft of this sensitive data and, therefore, the risk of patients being affected. The proposed model aims to mask event logs containing sensitive data so that they are inaccessible when process mining is applied. The model comprises four main stages: 1. target definition and data transformation; 2. data masking; 3. inspection and pattern analysis; 4. application of process mining techniques and data visualization. The model was validated using data from an appointment request process of a state health organization in Lima, Peru. Preliminary results showed that complete event logs containing sensitive data were protected, flow compliance increased by 68% and average processing time increased by 89.4%.

Keywords

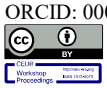
Process mining, Healthcare, Data privacy

1. Introduction

The healthcare sector is among the three sectors with the highest number of data breach and security incidents, in 2016 the healthcare sector was the most affected, with 116 incidents, representing 37.2% of all incidents, while the second most affected sector reported only 34 incidents [1]. The World Economic Forum shows in its 2020 Global Risks Report that digital data theft and the risk of cyberattacks on critical infrastructure (including those in the healthcare sector) were among the top 10 risks most likely to occur in that year [2].

Process mining is a very useful technique for the discovery of real process models by analyzing event logs. Because of its benefits, many institutions from different business areas use it to optimize their processes. However, being an emerging technique, process mining also faces challenges that have not yet been solved. One of them is to consider security and privacy issues when applying it [3]. The challenge is greater when using this discipline in the healthcare sector, since Electronic Medical Records (EMR) are the most important asset in the healthcare sector, because of the detail data they contain about patients.

This paper evaluates the creation of a reference model to ensure the privacy of sensitive data in the dating process, supported by Process Mining and Data Visualization. We expect this model will reduce the existing security gaps in Process Mining in the healthcare sector.

CISETC 2021: International Congress on Educational and Technology in Sciences, November 16-18, 2021, Chiclayo, Peru
EMAIL: u201613718@upc.edu.pe (A. 1); u201617237@upc.edu.pe (A. 2 jimmy.armas@upc.pe, a20184064@pucp.edu.pe (A. 3,4)
ORCID: 0000-0002-7521-4981 (A. 1); 0000-0001-8121-3160 (A. 2); 0000-0002-1176-8969 (A. 3,4)


This paper is structured as follows: we will review Process Mining models in the healthcare sector and then we will focus on describing the proposed model as a solution to the problem. Finally, conclusions and recommendations based on the results got in a case study are presented.

2. State of the Art: Process Mining Models

A Three-step framework for privacy preservation during the application of process mining is presented in [4]. In the first step, sensitive information is protected; in the second step, privatized metadata is created. Finally, the third step comprises of applying process mining on this metadata. However, a case study was not carried out to validate the variables of the framework, so the authors mention that the effects of the application of data transformation methods to preserve of privacy in the event logs of healthcare sector organizations should be investigated.

In [5], a five-phase reference model is developed for the evaluation of operational variables in healthcare using process mining and data visualization. In the first phase, data mining is performed, while in the second phase the event logs are processed, which will be analyzed through process mining in the third phase and represented in dashboards using the data visualization techniques applied in the fourth phase. Finally, the results are evaluated in the fifth phase. This model allows the identification of the effects of the application of process mining on healthcare sector records, but does not include techniques or practices that preserve the privacy of these event logs.

In [6], a protection model for event data privacy is designed using differential privacy, which allows the sharing of public information about a dataset without allowing the sensitive data of the individuals to be compromised. This model protects sensitive data using queries so that process analysts do not have access to it. However, the authors show that data protection only applies to one of the 3 activities of process mining: process discovery, and that it does not extend to compliance verification and process improvement activities.

In [7], details the analysis of a series of tools used to carry out cyberattacks on healthcare institutions in order to identify the most appropriate defensive techniques.

These techniques do not include the protection of optimized processes through process mining.

In short, the literature reviewed includes models and frameworks focused on process mining applied to the healthcare sector, but they do not satisfactorily cover the privacy aspect, sometimes because it is not addressed at all [5], or because it does not protect data throughout the entire process of applying process mining [6].

3. Data privacy process model: proposed solution

3.1. Description of the proposal

The model designed to be presented in Figure 1 comprises of four phases, taking as a reference the method of [8]. This model ensures the privacy of sensitive data that allows the identification of patients whose event logs are within the base used for the application of Process Mining and Data Visualization. Based on the regulations defined by Ministerial Resolution No. 688 - 2020 MINSA [9].

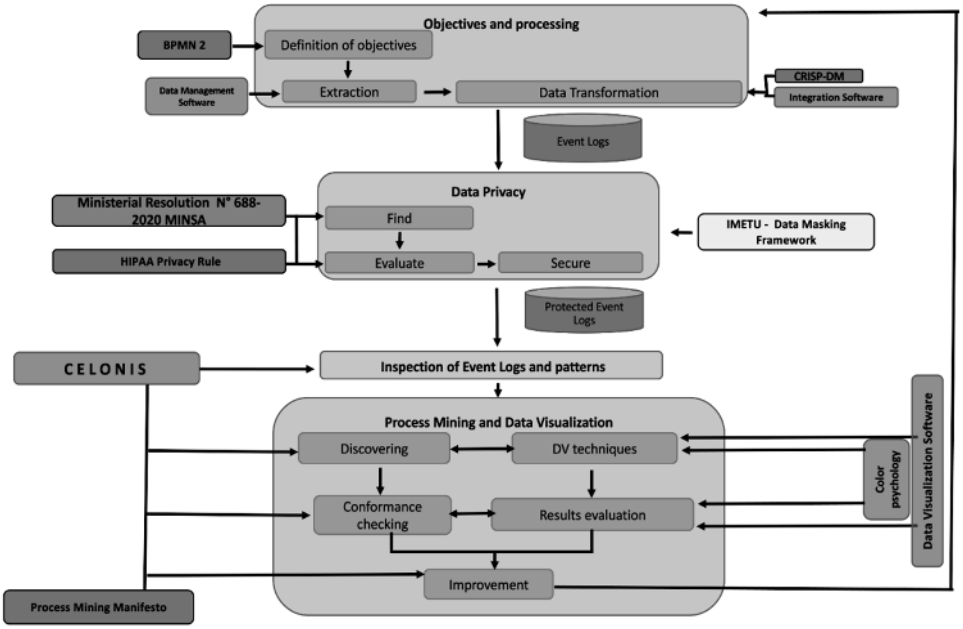


Figure 1: Reference Model for the privacy of patient personal data in processes using Process Mining

3.2. Phases of the model

3.2.1. Objectives and processing

The objectives of the project are defined, and the data are processed based on them. It has three main sub-phases: definition of objectives, extraction and transformation. In the first sub-phase, different indicators, such as time, quality and cost, must be taken into consideration. Once the objectives, both general and specific, have been defined, questions should be generated based on them. These will establish a way to evaluate objectives, specifically their progress and fulfillment. The second sub-phase involves the extraction of event logs from various sources, which will be used later for the application of Process Mining. For the last sub-phase, a cleaning, integration and quality assurance of the event logs will be performed, in order to have only one merged base, showing the ID, the activity and the timestamp.

3.2.2. Data privacy

This phase is based on [10], where all event logs already transformed go through a masking process to ensure the privacy of sensitive data, leading to patient identification. This phase also comprises three sub-phases: find, evaluate, and protect. The first sub-phase is based on the identification of the sensitive data within the event logs. After the analysis, there will be a generated list of the data that will be masked to maintain the privacy of the patients. The second sub-phase aims to identify the optimal masking algorithm for the event logs. Each attribute deserves a specific form of masking. The last sub-phase is the one where the chosen technique is executed for each of the attributes. Thus, the event logs are ready for application within Process Mining

3.2.3. Inspection of event logs and patterns

In this phase, the first impression is obtained from the event logs and different statistics that are collected to create a summary of the pattern that they follow. The number of cases, events, and their duration, resources, patterns, and event frequencies are inspected to have a prompt visualization of the process and to understand it completely.

3.2.4. Process Mining and Data Visualization

This phase is based on the application of the different process mining techniques with the protected event logs, in order to get information about the process and its compliance, as well as to adapt the data so that they can be correctly understood by non-expert users. This phase comprises five sub-phases: discovery, verification, data visualization techniques, evaluation of results, and improvement. In the first sub-phase, the real flow of the process will be found as recorded by the event logs within the tool used. In the second sub-phase, inconsistencies related to the compliance of the initially designed process and the event logs obtained will be detected. In the third sub-phase, the techniques that will apply to the different attributes of the logs in the data visualization will be defined, seeking the best representation of these and thus generate relevant information for the evaluation of the process. In the fourth sub-phase, the results presented through the different visualization techniques are evaluated in order to propose subsequent improvements or corrections that will help to optimize the current process. In the last sub-phase, after the analysis of results and measurement of indicators that allow answering the questions raised in the first phase, improvement opportunities will be obtained to start a new cycle of the model.

4. Case Study: Experimentation

4.1. Organization

Following the best model validation practices outlined in [11], which show that successful model validation requires that all its steps are fulfilled, the model validation process was performed by processing, securing, and analyzing a dataset from the appointment process of a public health institution in Lima, Peru.

4.2. Validation Process

4.2.1. Definition of objectives and process

First, as part of the model's objectives and processing phase, the objectives related to the project were defined through the formulation of questions, and with variables and indicators to answer them, check Table 1.

Table 1
Objectives and Indicators

Objective	Question	Variable	Indicator
Know the performance of ESSALUD's appointment process	How is the process going?	Process integrity	Number of cases in the process
	What are the most common flows?	Process flow compliance	Percentage of occurrence of the most frequent flow
	What are the most limited flows?	Process flow compliance	Percentage of occurrence of each alternative flow
	To what degree is the optimal process flow met?	Process flow compliance	Percentage of occurrence of the optimal flow
	What are the bottlenecks in the process?	Process flow compliance	Average waiting time between activities

Know the level of protection of sensitive data in the appointment process	How many event logs with sensitive data are protected?	Probability of event logs theft	Number of protected records

Finally, a BPMN diagram of the process was made for later comparison with the model discovered during the process mining application, see Figure 2.

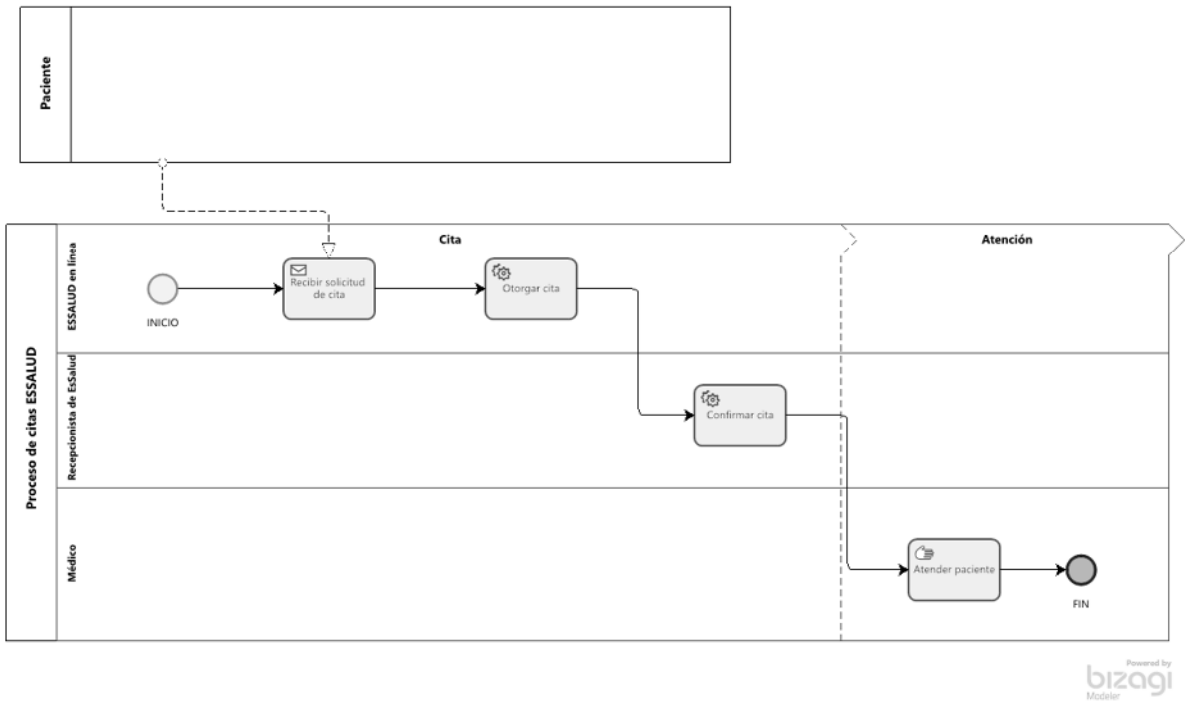


Figure 2: Diagram of the appointment process

4.2.2. Extract data

Continuing with the objectives and processing phase, and with the support of the health institution's staff, three Event Data bases were obtained: Request, Granting, and Appointments in Excel. All of them had an identifier named "ACTO_MEDICO", which will later allow the consolidation of the databases.

4.2.3. Process data

First, the three extracted databases were subjected to a cleaning process to eliminate null, incomplete, and inconsistent data that could negatively affect the reliability and accuracy of the results; for example, reserved appointments where the patient did not attend, thus leaving a gap in the field of Attention, or appointment confirmations made at a later time than their attention, due to errors human error caused by workers. Then, the three databases were integrated into a single database through the

"ACTO_MEDICO" field mentioned above; however, this integrated database still does not meet the minimum characteristics required by an event log. Finally, Phyton 3.9 was used to, through the Pandas library, generate the event logs with the required field (ID, activity, and timestamp). Each event log was composed of four activities: Request, Grant, Appointment, and Attention.

4.2.4. Data masking

As part of the data privacy phase, the event logs were masked in three steps. First, all event logs containing DPS (Personal Health Data) were identified, which, as stated in Ministerial Resolution N° 688-2020 MINSA, are highly confidential. Then, the masking technique was evaluated for each field based on the IMETU (Identify, Map, Execute, Test, and Utilize) masking framework. Finally, the techniques were applied to the database stored in Excel, check Table 2.

Table 2
Protected Event Logs

DPS of event log	Example	Selected masking	Masked DPS
National Identification Card (DNI)	12345678	Remove last 4 characters	1234####
Patient	Rafael Pedro Ramirez Vela	Use Excel Kutools Add-in	9E7475F70KJYdCys5Aoqckh cnuSvaXqQG8m0mTQi7HZw h/R87cQ=
Age	44	Increase value by 20	64
Sex	M	The value "*" will be taken	*
Physician's National Identification Card	87654321	Remove last 4 characters	8765####
Physician	Javier Mateo Lopez Zarate	Use Excel Kutools Add-in	8BCF7BD70KJYdCys5AoPCf W+5Ua3dYGICQMd45wV9e F9JJ7SIETuSC4TWOiD+w==

4.2.5. Process Mining Application

As part of the event log and pattern inspection phase, the masked data were loaded into the Celonis platform to get an overview of the process using the metrics it provides, such as daily cases and events, average process time, or bottlenecks. Next, the process mining phase proceeds with the discovery of the process model through the Celonis Overview tool, which allows us to see the discovered model with all its deviations. Then, the process model is loaded to be compared with the model discovered in verification. In the first data load, in Figure 3, this verification was 12% of event logs. Following the continuous improvement approach of the model, problems in the data were identified and corrective actions were taken, such as using the Excel DATEDIFF function to validate the correct sequence of dates.

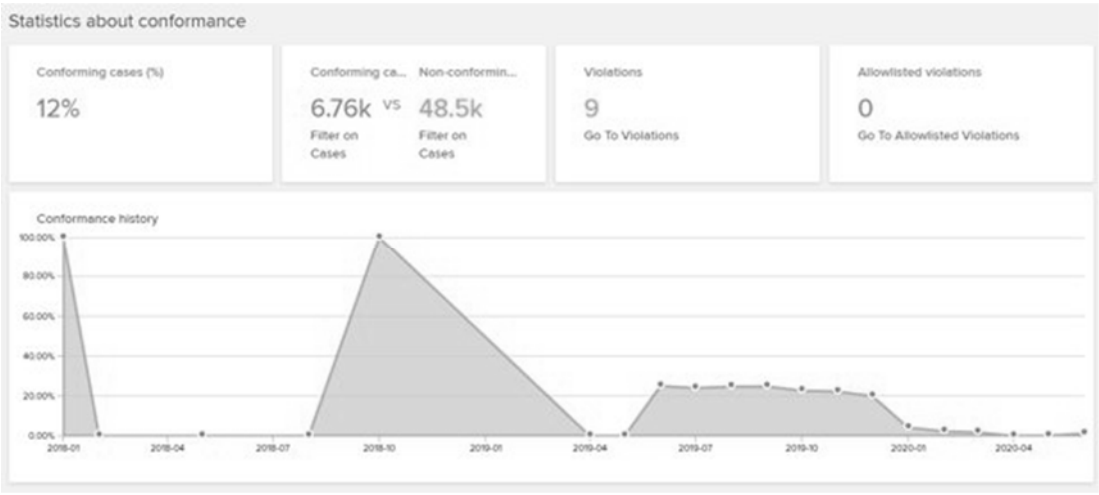


Figure 3: Safety verification of the first load

In the second load, see Figure 4, the verification was 80%, but the diagram obtained looked forced because Celonis did not organize correctly the activities that occurred on the same day due to the absence of the correct time in the timestamp, so the date and time fields were unified in Excel to allow the timestamp to take it into account.

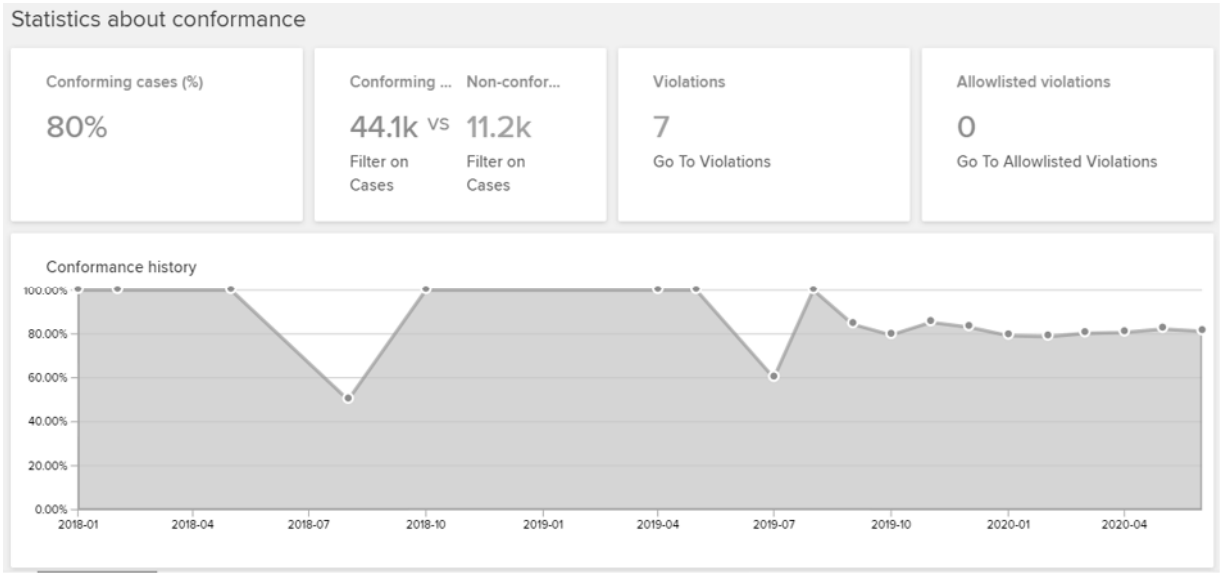


Figure 4: Safety verification of the second load

In the Figure 5, the third load, satisfactory results were obtained, so we proceeded with the next phase. The discovered model is shown below in Figure 6, followed by the safety verification.

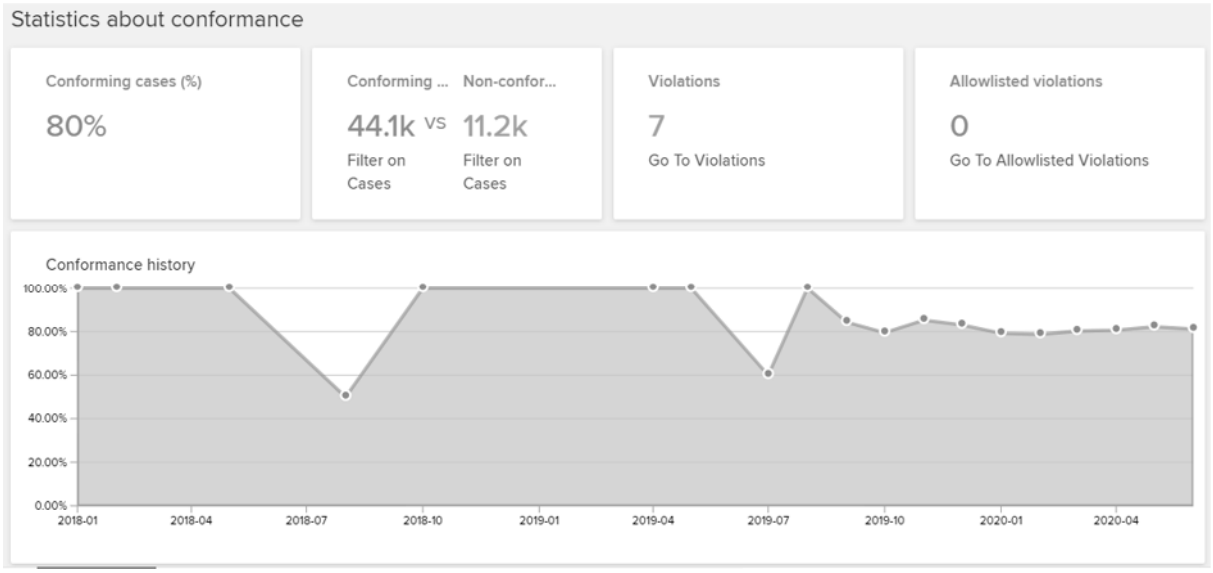


Figure 5: Safety verification of the third load

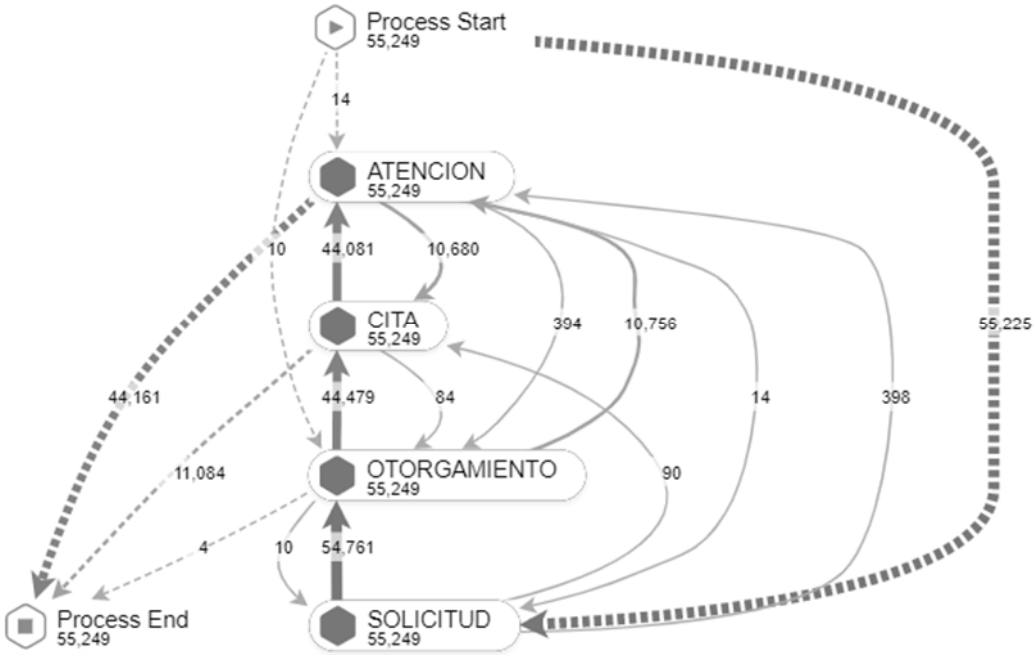


Figure 6: Model discovered in the third load

4.2.6. Data Visualization and Decision Making

Using the Celonis Studio and Celonis Business Views tools, dashboards facilitated the understanding of the analysis by non-expert users using some traditional charts such as pie charts to show the distribution of cases by medical service, or the bar chart, where the average process time by age group in days is shown. Subsequently, the data from the Celonis results are evaluated and improvement actions are determined, such as the definition of start and end times for the activities, the

creation of a variable in the confirmation activity that shows that the appointment has been attended, among others. Regarding the results, all event logs that contained DPS are protected, complying with the Ministerial Resolution N° 688-2020 MINSA, while the average time of the process was reduced by 89.4% and the percentage of compliance with the flow increased by 68%. With this, it can be affirmed that masking data to ensure its privacy does not prevent an effective process mining analysis, nor does it affect the reliability and accuracy of the analysis.

5. Conclusions and perspectives

In the paper, we proposed a reference model to ensure the privacy of confidential patient data in the health appointment process using Process Mining. The model was applied in an operational context in the search of answering questions that help to know the behavior of the process and find improvements. 55,249 event logs were reviewed for the case study, through which all confidential records were obtained masked, ensuring their privacy, the compliance of the process increased by 68% and the average execution time decreased by 89.4%. This not only ensures the privacy of confidential records in the event logs, but also has a positive impact on the process. It is recommended to evaluate the addition of a data protection and governance phase, which includes the definition of roles and authentications that reinforce the protection of sensitive information recorded in the healthcare sector.

As future work, it is recommended to improve the quality of the data recorded in the databases by periodically cleaning null or empty data and incorrect dates and inconsistent data, since these may affect the analysis and the results obtained are not as accurate, so there is the probability that the improvement of the process will be focused in the wrong direction. It has also been noted the need for the definition of start and end times for the care activity. In this way, it will be possible to justify the number of appointments to be carried out in a period or for a specific service and also to know the number of resources that can be allocated to minimize time.

6. References

[1] Hurst W., Boddy A., Merabti M., & Shone N. (2020). Patient Privacy Violation Detection in Healthcare Critical Infrastructures: An Investigation Using Density-Based Benchmarking. Future Internet, 12(6), 100. Recovered from http://dx.doi.org/10.3390/fi12060100

[2] Banco Interamericano de Desarrollo (BID), Organización de los Estados Americanos (OEA). (2020). Reporte Ciberseguridad 2020: Riesgos, avances y el camino a seguir en América Latina y el Caribe

[3] Van Der Aalst, W. et al. (2011). Process mining manifesto. En International Conference on Business Process Management (p. 169-194). Springer, Berlin, Heidelberg. Recovered from https://doi.org/10.1007/978-3-642-28108-2_19

[4] Pika, A., Wynn, M. T., Budiono, S., ter Hofstede, A. H., van der Aalst, W. M., & Reijers, H. A. (2019). Towards privacy-preserving process mining in healthcare. In International Conference on Business Process Management (pp. 483-495). Springer, Cham. Reovered from https://doi.org/10.1007/978-3-030-37453-2_39

[5] Aguirre, J. A., Torres, A. C., & Pescoran, M. E. (2019). Evaluation of operational process variables in healthcare using process mining and data visualization techniques. Health, 7, 19. Recovered from http://dx.doi.org/10.18687/LACCEI2019.1.1.286

[6] Mannhardt, F., Koschmider, A., Baracaldo, N., Weidlich, M., & Michael, J. (2019). Privacy-preserving process mining. Business & Information Systems Engineering, 61(5), 595-614. Recovered from https://doi.org/10.1007/s12599-019-00613-3

[7] Ibarra, J., Jahankhani, H., & Kendzierskyj, S. (2019). Cyber-physical attacks and the value of healthcare data: facing an era of cyber extortion and organised crime. In Blockchain and Clinical Trial (pp. 115-137). Springer, Cham. Recovered from https://doi.org/10.1007/978-3-030-11289-9_5

[8] Ibarra, J., Jahankhani, H., & Kendzierskyj, S. (2019). Cyber-physical attacks and the value of healthcare data: facing an era of cyber extortion and organised crime. In Blockchain and Clinical Trial (pp. 115-137). Springer, Cham. Recovered from https://doi.org/10.1007/978-3-030-11289-9_5
[9] Ministerio de Salud (MINSA). (2020). Resolución Ministerial 688 – 2020.
[10] Ali, O., & Ouda, A. (2016, October). A classification module in data masking framework for business intelligence platform in healthcare. In 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 1-8). IEEE. doi: 10.1109/IEMCON.2016.7746327
[11] Anderson, M. P., & Woessner, W. W. (1992). The role of the postaudit in model validation. Advances in Water Resources, 15(3), 167-173